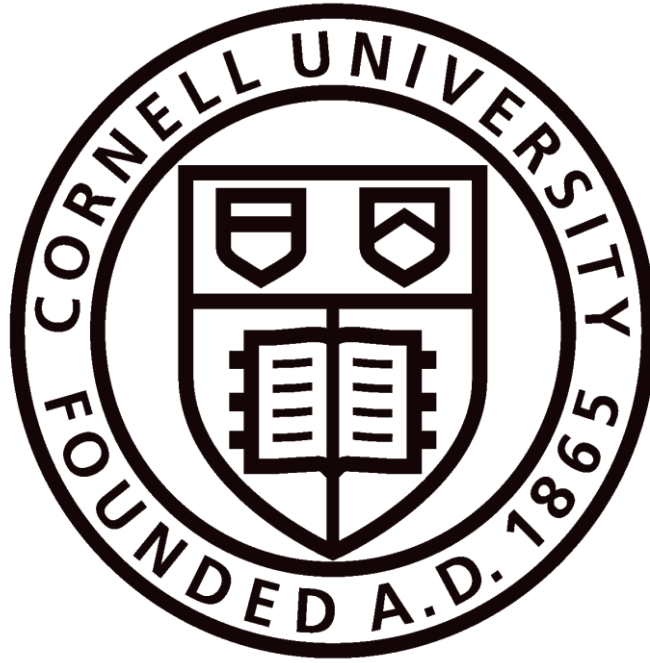**SCHRÖDINGER'S CATEGORIES:**
**THE INDETERMINACY OF**
**FOLK METAETHICS**

**A Dissertation**
*Presented to the Faculty of the Graduate School*
*of Cornell University*
*in Partial Fulfillment of the Requirements for the Degree of*
*Doctor of Philosophy*

by

Lance Spencer Bush
May 2023

SCHRÖDINGER'S CATEGORIES:
THE INDETERMINACY OF FOLK METAETHICS

Lance Spencer Bush, Ph.D.

Cornell University 2023

*Metaethics* is a field of philosophy that addresses fundamental questions about the nature of morality. One of the central disputes in metaethics is whether *moral realism* is true. Moral realism is the claim that there are *stance-independent moral facts*, moral facts that are true independent of the standards or values of individuals or groups, much like scientific facts (e.g., the shape of the earth) aren't made true by personal preference or cultural consensus. *Moral antirealism* is the claim that *there are no stance-independent moral facts*. Research on *folk metaethics* studies whether ordinary people (i.e., *nonphilosophers*) endorse realism or antirealism, or speak and think in ways that commit them to one of these views. Some researchers maintain that nearly everyone endorses either realism or antirealism, but not both. Yet most research suggests significant interpersonal and intrapersonal variation in *folk metaethics*: some people are more inclined towards realism, and others antirealism, while most people are *metaethical pluralists*: they are moral realists about some moral issues and antirealists about others. Regardless of the account in question, *all* existing research presumes that there is a *determinate* fact about whether people are realists or antirealists. I argue that existing evidence does not support this conclusion. Instead, the best account of folk metaethics may be *metaethical indeterminacy*: ordinary people are neither realists nor antirealists, and neither best explains the way people speak or think. The case for *metaethical indeterminacy* proceeds in two steps. First, I argue that *all published studies on folk metaethics rely on invalid measures*. Second, I present empirical evidence that challenges the validity of existing research on folk metaethics and supports metaethical indeterminacy. I evaluate the proportion of people who interpret questions about metaethics as intended, using open response questions, as well as multiple choice questions and Likert scale items. These studies show that most people do not interpret questions

about metaethics as researchers intend. I conclude with a study that demonstrates how forced choice paradigms can create the misleading appearance of a genuine pattern of determinate folk philosophical views, even where none plausibly exist.

# BIOGRAPHICAL SKETCH

I have been told that I tend to focus on the negative, and I think that's mostly correct. I'd like to think that this gives me an edge. But it's just who I am. It's not in my nature to provide a positive biographical sketch; the kind that presents an array of highlights and memories like polished trophies. We all know these glossy histories are at best only half-truths. Sure, I've had bright moments. My mom tucking me into bed, folding half the blanket over me and declaring me a "taco." Foggy memories plodding through the swamps of New Jersey with my brother, casting a wary gaze over my shoulder for signs of the Jersey Devil. Winning first place in the pinewood derby. Tearing open a pack of cards from the *Weatherlight* expansion of "Magic: The Gathering" to find a Thundermare staring back at me, eyes ablaze.

Childhood was a time of wonder. But it was also a time of pain, confusion, and loneliness. I'd like to think the person that I became benefited from the hardships, but I fear I'm simply all that remains after so many enervating experiences left me pitted and gouged. I dropped out of high school in tenth grade—a rocky start, but I turned things around. Community college at sixteen. Finished two bachelor's degrees, then completed an MA at Tufts. Hobnobbed with students at Harvard and Oxford. Spent a few months as a research assistant in the Morality Lab at Boston College, and a brief stint teaching philosophy before I managed to get into Cornell. I've had a lot of opportunities, to be sure, and no small amount of luck.

But like I said, I tend to focus on the negative. In August 2019 I developed a chronic medical condition that nearly scuttled any hope I had of finishing my dissertation. A few months later, my mom died after a long battle with cancer—she beat the cancer, but she died anyway. I see pictures of myself from 2018 and I don't recognize the smile there anymore, or the absence of gray hair.

This dissertation is the culmination of over ten years of thinking about metaethics, moral

psychology, and moral philosophy. My aims are largely critical: critical of every study on the psychology of metaethics, of moral psychology, and of philosophy in general. If my central arguments are correct, then an entire literature is fundamentally misguided and thoroughly mistaken. Yet I think it'd be a mistake to see my ambitions as merely destructive, or to think that nothing good can come from an emphasis on criticizing the work of others. When we clear away the mistakes and confusions of the past, wherever possible, we make way for something new, something better. But the pursuit of something better is a task for someone else. Because, after all, I tend to focus on the negative.

*To mom*

# ACKNOWLEDGEMENTS

No self-respecting person would begin their dissertation without an obligatory tipping of the hat to everyone who helped them along the way. As cynical as I am, this is one ritual I wholeheartedly endorse. Since I have many people to acknowledge, I'll begin by listing everyone I mention: Dr. David Pizarro, Dr. Laura Niemi, Dr. Shaun Nichols, Dr. Morten Christiansen, David Moss, Andres Montealegre, Tyler Millhouse, Don Loeb, Michael Gill, Sam Rosen, Brian Tebbitt, Bryce Gessell, Todd K. Shackelford, Sebastian Deri, Rebekah Davis, Thomas Pölzler, Dillon Bowen, Lieuwe Zijlstra, Geoffrey Goodwin, Jennifer Cole Wright, Ross Colebrook, Adam Gibbons, Tyler John, David Newman, Geoffrey Goodwin, and my wife Priscila Serpa.

First, I owe the deepest debt of gratitude to my advisor and the chair of my committee, Dr. David Pizarro. I am grateful for the chance to demonstrate what someone with a fundamentally different approach to psychology can bring to the field and can only hope the gamble pays off. I am also grateful for the approach David took as an advisor and a mentor. I'm an inveterate loner. Stubborn and independent, many of my greatest shortcomings stem from my unwillingness to conform with conventions even when they are there for good reason. David provided the breathing room for me to retain a sense of autonomy, Yet David also knew how and when to step in as it was necessary. There were many times I was afraid to step out of my comfort zone and address the weaknesses in my knowledge and skills. More than merely serving as a mentor, David excelled at creating an atmosphere that acknowledged the pressures of graduate study, but with an ever-present humor that made the whole process tolerable.

I also want to acknowledge the rest of their committee, Dr. Laura Niemi, Dr. Shaun Nichols, Dr. Morten Christiansen, for their insights, thoughts, and contributions. I haven't had the chance to

interact with the rest of the committee as much as I'd have liked, but I am nevertheless grateful for their insights, feedback, and criticisms.

I owe a special thanks to David Moss. David has contributed more than anyone to my research and my development as a thinker. David has been a constant source of insight, moderating my brasher impulses with a more measured approach to research and presenting countless insights and observations that have furthered my knowledge of psychological research and folk philosophy. In fact, although I've refined and implemented the coding scheme employed in chapter 4, it was David Moss who initially proposed this approach. David has also contributed a great deal to the studies reported here, assisting in the formulation of the hypotheses and the measures used, as well as contributing to the many criticisms I level against existing research on folk metaethics. David is in every way a research collaborator and deserves much of the credit for the final product.

I also want to thank my other colleagues and collaborators. Andres Montealegre has been a friend, research colleague, and supporter of the direction of my research for many years now. We have worked on numerous projects, and I have learned more from Andres about how to use R, quality research design, open science practices, the importance of a rigorous approach to survey design, preregistration, and much more than I ever could have imagined.

Thanks, as well, to Tyler Millhouse. Tyler was my roommate when I was a philosophy graduate student at Tufts. Our many discussions about philosophy both deepened my understanding of metaethics and broadened the scope of my knowledge with other areas of philosophy. Tyler also tempered many of my worst philosophical habits and served as a role model for the ability to approach philosophical disagreements with calm, composure, and candor. Tyler has provided feedback on several projects related to folk metaethics and has worked with me on many of my earliest projects, and he continues on occasion to work with me on current projects.

Don Loeb has also earned my gratitude. Don's work has had a profound influence on me, inspiring and motivating me to continue my pursuits with the confidence that a graduate student without the academic credentials of professional philosophers wasn't completely misguided in thinking there was something wrong with the methods and conclusions of contemporary metaethics. Don has also been an open book about his life, sharing many stories and anecdotes. His avuncular demeanor and his ability to befriend virtually everyone he meets is a reminder that philosophers need not be insular and aloof. I have yet to learn how to be as kind and compassionate towards others, and I often fear that it's beyond me. Thanks for being so awesome, Don!

I also want to thank Michael Gill. Dr. Gill's paper on indeterminacy and variability in metaethics could be seen as the key paper motivating my research. This is the only paper I know of that explicitly argued for metaethical indeterminacy. In many ways, this project is a very long footnote to Loeb's incoherentism and Gill's Indeterminacy-variability thesis.

I want to thank Sam Rosen for suggesting the title for the paper. I love bad jokes and terrible wordplay, and *Schrödinger's Categories* scratched that itch perfectly.

Thanks also to Brian Tebbitt, who provided valuable feedback on the content and direction of the dissertation and helped remind me to focus rather than go on a bunch of tangents. Thanks also for the many helpful discussions which helped shape the way I think by refining my views about language and the relationship between language, philosophy, and psychology.

There are far too many other people to thank as well. Thanks to the editors and reviewers at *New Ideas in Psychology* and *Diametros*. A special thanks to Bryce Gessell for helpful discussions and feedback. Thanks to Todd K. Shackelford for inspiring me to pursue psychology in addition to philosophy, for providing so many opportunities to publish articles and to interact with incredible researchers, and most of all for helping me to develop a richer understanding of evolution. If his classes hadn't been so engaging, I may have never moved into my current line of research. Thanks to

Sebastian for extensive help with R and feedback on how to write a dissertation. Thanks to everyone who attended the folk moral relativism workshop and to Thomas Pölzler for organizing it. Thank you to Rebekah Davis for providing critical feedback on the dissertation and for encouragement with my research. Thanks to many others for their helpful discussions and feedback, including Dillon Bowen, Lieuwe Zijlstra, Geoffrey Goodwin, Jennifer Cole Wright, Ross Colebrook, Andrew Maul, Hagop Sarkissian, Adam Gibbons, Tyler John, David Newman. I'd also like to thank Geoff Goodwin for generously sharing data from his research, despite my critical analysis of portions of it. Thank you to everyone else at Cornell who provided support or feedback or simply attended one of my talks. Thanks also to the many people I don't remember or never met who saw one of my posters or presentations at conferences. Thanks also to my parents, who supported me the whole way. Thanks to our wonderful dog Tchutchu, who is sadly no longer with us. She was my sole company for most of the writing process and throughout the solitude of the pandemic. I'd like to think she did the most to motivate me.

Finally, I'd like to thank my wife, Priscila. She has had to put up with all of my flaws and shortcomings the most, yet she's done more than everyone else to keep my head in the game and my fingers typing. Priscila undoubtedly deserves the most credit for seeing this work to completion. Priscila helped with formatting my figures and tables and assisted with numerous aspects of the text, from feedback on the overall organizational structure, deciding which studies to include, providing an outsider's perspective on studies and how I frame my arguments, to virtually every element of writing, formatting, and thinking about the topics I discuss. Yet all the academic feedback and support in the world would be irrelevant if I didn't have the motivation and focus to finish. A lot of terrible things happened in the past few years that held me back and made finishing this project an immense struggle. Priscila has been with me every step of the way, providing encouragement and support, and pushing

me to get organized, to get focused, and to take care of my mental and physical health. I couldn't have

done this without her.

# TABLE OF CONTENTS

**CHAPTER 1:**

**Introduction**

Alex and Sam are two typical Americans. They are planning to have a barbeque tomorrow, but according to a local weather forecast, it might rain. If it does rain, they will have to cancel their plans. This has led to an argument about the weather:

> Alex: *"Given the indeterministic nature of the universe, it is ultimately unknowable whether it will rain tomorrow."*

> Sam: *"Nonsense. It will definitely both rain and not rain tomorrow. The universe will diverge down two separate branches. It will rain in one universe, and not in the other."*

Alex and Sam seem to be endorsing the Copenhagen and Many Worlds interpretations of quantum mechanics, respectively. This disagreement probably seems absurd. It is incredibly unlikely that interpretations of quantum mechanics would arise in everyday disputes about the weather, and even if they did, neither perspective would be helpful in resolving the practical question of whether to cancel a barbeque. The scenario is also absurd because *most people don't have a position on how to interpret quantum mechanics*. Such considerations are irrelevant to everyday decisions, and nothing about the way ordinary people think or speak requires them to take a side in disputes that occupy theoretical physicists.

I believe that a similar absurdity plagues empirical research on *folk philosophy*. *Folk philosophy* refers to the philosophical stances and commitments of ordinary people. *Philosophical stances* are the philosophical beliefs ordinary people hold. *Philosophical commitments* are the philosophical positions implicit in the way ordinary people speak, think, and act, independent of any particular psychological states.[1] *Ordinary people* are people who lack significant formal philosophical training and have not

---

[1] I will often refer to these as simply *stances* and *commitments*.

engaged in significant philosophical reflection. The goal of empirical research on folk philosophy is to *describe the philosophical stances and commitments of ordinary people*.

Researchers have devised a cunning array of paradigms for cataloging the stances and commitments of ordinary people, spanning every major branch of philosophy from **metaphysics** (Dink & Rips, 2017; Korman & Carmichael, 2017; Nichols & Bruno, 2010; Paul, 2010; Rose , Schaffer, & Tobia, 2018) to **epistemology** (Nagel, San Juan, & Mar, 2013; Starmans & Friedman, 2012; Swain, Alexander, & Weinberg, 2008; Weinberg, Nichols, & Stich, 2001) to **ethics** (Alicke, & Gordon, & Rose, 2013; Greene, 2008; May & Holton, 2012; Phillips, Nyholm, & Laio, 2014) and **aesthetics** (Cova et al., 2015; Cova & Pain, 2012; Rabb et al., 2020).[2]

Researchers have frequently interpreted these studies as evidence that ordinary people have stances or commitments that correspond to the categories, accounts, and distinctions recognized by philosophers, e.g. *deontology* and *consequentialism* (e.g. Greene, 2008; Kahane et al., 2018; Johansson-Stenman, 2012; May, 2014; cf. Mihailov, 2022), **compatibilism** and **incompatibilism** (e.g., Carstensen, 2022; Nadelhoffer et al., 2020; Nahmias & Murray, 2011; Nahmias et al., 2004; 2005; 2006; 2015; Nichols, 2012; Nichols & Knobe, 2007, cf. Nadelhoffer, Murray, & Murray, 2021), **psychological** and **non-psychological** conceptions of **personal identity** (e.g. Nichols & Bruno, 2010; Shoemaker & Tobia, 2022; Strohminger & Nichols, 2014; 2015; cf. Starmans & Bloom, 2018a; 2018b), adherence to or rejection of the **'ought' implies 'can' principle** (Buckwalter & Turri, 2015;

---

[2] Some of these studies may only be concerned with measuring philosophical *intuitions* of a kind that do not correspond to what I mean by philosophical *stances* and *commitments*. "Intuition" is a term regularly employed by philosophers, but for which there is unfortunately no single definition. Researchers studying folk *intuitions* may or may not be studying stances or commitments, e.g., they might be studying *dispositions* to believe certain propositions, rather than studying what people pretheoreticallly believe are committed to (Earlenbaugh & Molyneux, 2009). Since I am offering a stipulative set of terminological distinctions other authors have not used, I would be unable to confirm whether they conceive of their findings as indicators of stances and commitments without consulting each author, explaining the distinction, and asking them for their position. Failing that, I could selectively opt for studies that appear to fit my distinctions, this task would be largely guesswork on my part. Instead, I chose broadly representative research characterized by scope and impact. I believe these studies provide a more useful picture of the kind of research I am referring to than a curated list of more obscure studies that would risk presenting a skewed notion of what folk philosophical research is about, even if some of these studies are not subject to my criticisms.

Chituc et al., 2016; Cohen, 2018; Henne et al., 2016; 2019; Kissinger-Knox, Aragon, & Mizrahi, 2018; Mizrahi, 2015; Semler & Henne, 2019; cf. Thompson, 2022) **causal-historical** and **descriptivist** theories of reference (e.g. Machery et al., 2004, Mallon et al., 2009; van Dongen et al., 2021), and so on.

Yet few researchers have seriously considered the possibility that ordinary thought and language is *indeterminate* with respect to at least some of these philosophical distinctions.[3] In other words, there may be philosophical distinctions that are absent from the way ordinary people speak and think, and that, for these philosophical issues, *ordinary people don't have any particular philosophical stances or commitments at all*. If so, then there may be no way to resolve competing analyses of the philosophical content of folk philosophy, since these analyses could equally accommodate (or fail to accommodate) the data (Gill, 2009). If this is the case, the implications would be catastrophic for much ongoing research, since this research would be attempting to describe features of ordinary thought and language that don't exist.

I am not claiming that ordinary people have *no* philosophical stances or commitments. Even the most skeptical account of folk philosophy would acknowledge that ordinary people hold some minimally construed stances and commitments. To the extent that people endorse first-order moral judgments (e.g., "torture is morally wrong") they could be said to have a philosophical stance. And insofar as ordinary people speak as though some beliefs are more justified than others, that there is an external world, or that we have a *prima facie* duty to keep our promises, ordinary people may be said to have epistemic, metaphysical, and moral commitments, respectively. I am not challenging the existence of determinate stances and commitments of this kind. Yet these are mundane and uncontroversial beliefs that are rarely the focus of research on folk philosophy.

---

[3] Gill (2009) and Pravato (2020) are notable exceptions. Gill explicitly argues for the possibility of indeterminacy in folk metaethics (for replies see Johansson & Olson, 2015; Sinnott-Armstrong, 2009). Pravato (2020) argues that normative language (e.g., "good," "ought") is indeterminate, though it is not meaningless.

Research on folk philosophy instead tends to focus on how ordinary people think about ongoing disputes central to academic philosophy, often with the goal of illuminating or resolving these disputes (Alexander, Mallon, & Weinberg, 2010; Knobe, 2007; Sytsma & Livengood, 2012). As a result, most research on folk philosophy offers insight into ordinary thought and language that is nonobvious or at least subject to reasonable doubt. For instance, researchers studying folk philosophy could plausibly wonder whether ordinary people are committed to the notion that *'ought' implies 'can'*. Yet researchers studying folk philosophy would have little interest in studying whether people think it's morally wrong to torture babies for fun. This is not to say that researchers wouldn't ask this sort of question, but simply that responses to it would not be interpreted as discoveries about folk philosophy, and they wouldn't play an important role in resolving philosophical disputes. Such measures would instead serve some conventional psychological purpose, such as measuring psychopathy.

At the other extreme, it seems unlikely that ordinary people hold philosophical stances about more esoteric philosophical issues, or speak and think in ways that exclusively fit one or another side of obscure philosophical disputes. For example, Gill (2009) finds it implausible that ordinary people speak or think in ways that commit them to mathematical Platonism or anti-Platonism when they use mathematical language, since he believes "[t]he way people use numbers in everyday math simply does not contain answers to the questions that animate philosophy of mathematics" (p. 218). Likewise, it is unlikely that the way people speak or think commits them to a particular stance about the legitimacy of the analytic-synthetic distinction, the optimal decision-theoretic solution to Newcomb's problem, or other debates that concern academic philosophers, and it is even less plausible that ordinary people have explicit stances on these issues.[4]

Thus, there is a continuum between folk stances and commitments that uncontroversially exist and those that are likely absent from folk philosophy altogether. Somewhere between these extremes

---

[4] For other examples or popular philosophical disputes, see Bourget and Chalmers (2014; ms).

lies a nebulous middle ground of philosophical accounts, concepts, and distinctions that possess a less-than-certain status as features of folk philosophy. This middle ground is the uncertain territory where I intend to build my case for folk indeterminacy. In making my case, I am not claiming that there are *no* determinate features of folk philosophy between the extremes. There are many features of the way ordinary people think and speak that have been or could be discovered, and at least some of these findings may correspond to traditional philosophical distinctions (e.g., perhaps most ordinary people really are *compatibilists* or *incompatibilists* about "free will"). Yet each hypothesis must be evaluated on an individual basis. We should not presume that a given distinction that is important to philosophers is a part of folk philosophy. In the end, empirical evidence will be the final arbiter of the content of folk philosophy.

While my concerns apply to folk philosophy in general, I will focus exclusively on *folk metaethics*. Metaethics is a branch of philosophy that deals with abstract questions about the nature of morality, such as whether there are moral facts, what makes moral facts true, and how we might acquire knowledge of moral facts (Sayre-McCord, 2014). *Folk metaethics* is simply the subset of *folk philosophy* dedicated to studying the metaethical stances and commitments of ordinary people. While there are several questions that fall under the purview of metaethics, most research focuses on whether ordinary people endorse or are committed to some form of *realism* or *antirealism* about moral facts (Pölzler &

Wright, 2020a; 2020b). Roughly speaking, this distinction concerns whether there are *stance-independent*[5] *facts about what is morally right or wrong* (Shafer-Landau, 2003, p. 15).[6]

> **Realism***: There are stance-independent facts about what is morally right or wrong*

> **Antirealism***: There are no stance-independent facts about what is morally right or wrong*[7]

Most studies have found that the majority of ordinary people are *metaethical pluralists* who endorse realism about some moral issues and antirealism about others (Beebe, 2014; Beebe et al., 2015; Beebe & Sackris, 2016; Goodwin & Darley, 2008; 2012; Pölzler & Wright, 2019; 2020a; 2020b; Wright, 2015; 2018; Wright, Grandjean, & McWhite, 2013, Zijlstra, 2019; cf. Beebe, 2020). For instance, a person may judge that there is a stance-independent moral fact about whether murder is morally wrong, but that there is no stance-independent moral fact about whether abortion is morally wrong. They might instead think that claims about abortion can only be true or false relative to the moral standards of different individuals or cultures, or that such claims merely expression nonpropositional content, e.g., a negative emotion or an imperative to not get an abortion.

---

[5] I follow Shafer-Landau's lead by using the term stance-independent rather than the more common term mind-independent. This is because, as Shafer-Landau notes, the latter is more disposed to prompt confusion since there is one respect in which realism entails that moral facts are mind-independent: they are not made true by the beliefs or values of people. Yet there is another respect in which some moral facts may crucially depend on people's attitudes or values. People sometimes mistakenly think that mind-independence means that whether an action is right or wrong depends on the psychological impact that action would have on an individual. For instance, facts about whether it would be okay to hit someone would depend on whether that person would suffer. And since suffering is a mental state, we might think of claims like "hitting someone is wrong" as mind-dependent, since the reason hitting someone may be wrong is because it causes suffering. Yet this is simply a different sense in which a moral fact could depend on mental states than is intended by the notion of mind-independence or stance-independence.

[6] Realism is sometimes defined more minimally, and instead consists of two claims: that moral sentences are truth-apt and at least some of them are true (Sayre-McCord, 2015). There seems to have been a shift in recent decades towards reserving the term realism for more robust forms of realism that include stance-independence or other considerations.

[7] Technical terms like "stance-independent" are likely to prove unhelpful. Roughly, the distinction concerns whether there are moral facts that are not made true by our goals, standards, or values, or whether there are no such facts. Comparison with more familiar concepts might help. Many of us, on reflection, may agree that scientific facts don't depend on our goals, standards, or values. Believing, or really wanting the earth to be flat couldn't make it true that it was flat. Yet many of us would, if we reflected on the matter, deny that there are any stance-independent facts about which food or music is best. We might instead insist that it's a matter of personal preference, and that facts about which food tastes good or bad can only be understood relative to different people's stance towards the food (e.g., whether it tastes good or bad to them), if there are any facts at all. If so, we might say that we're gastronomic antirealists: we deny that there are facts about which foods taste good or bad that are stance-independently true. See Loeb (2003) for an amusing discussion of gastronomic realism and its relation to moral realism.

Some researchers have accepted these findings at face value and argued that *folk metaethical pluralism* is the best interpretation of the data (Davis, 2021; Feltz & Cokely, 2013; Hopster, 2019; Pölzler, 2017; Wright, Grandjean, & McWhite, 2013). Others have interpreted findings in folk metaethics to support the claim that most people are realists (Goodwin & Darley, 2008; 2012) or antirealists (Beebe, 2020; Sarkissian et al., 2011; Pölzler & Wright, 2020b) or at least that we can rule out specific folk metaethical views such as folk noncognitivism (Pölzler & Wright, 2020a). All of these interpretations share the presumption that folk metaethics is *determinate* (Gill, 2009).[8] In other words, they presume that we can decisively demonstrate that ordinary people adopt a realist or antirealist stance towards moral claims, or speak in ways that best fit some form of realism or antirealism.

I contend that all of this research is fundamentally flawed, and that these interpretations are all mistaken. I don't mean that the studies are poorly designed, and that, with a few tweaks, we'll be able to properly assess what the folk think about realism and antirealism. Nor do I mean that a proper interpretation of these studies would allow us to determine which stances or commitments ordinary people have regarding realism and antirealism. I mean that such efforts cannot succeed in principle because, with few exceptions, *ordinary people have no determinate stances or commitments about the truth status of moral claims.* With respect to realism and antirealism, most ordinary people exist in a state of philosophical superposition that collapses only by engaging in philosophy. I will refer to this as the *metaethical indeterminacy thesis*, which is the claim that *ordinary people have no determinate metaethical stances or commitments with respect to moral realism and antirealism.*

The primary purpose of this dissertation is to explain why I think this is the case, and to convince you that *metaethical indeterminacy* is a plausible account of folk metaethics. Since all existing research on folk metaethics has been interpreted as evidence that ordinary people have determinate

---

[8] One exception to this is Pölzler and Wright (2019; 2020a). They don't presume determinacy so much as explicitly argue that their findings cast doubt on indeterminacy. I am indebted to them for clearly laying out their reservations and doubts, and for offering such high quality and insightful studies for our disagreements to center around.

metaethical stances and commitments, this leaves me with three central tasks: (1) to demonstrate that the measures used in existing research on folk metaethics are invalid, (2) to demonstrate this can be readily explained by metaethical indeterminacy and (3) to provide empirical evidence that supports metaethical indeterminacy.

My third objective is intentionally modest. I do not intend to provide anything even approaching decisive evidence of metaethical indeterminacy. This would be an incredibly difficult task, since there are few straightforward ways to convincingly demonstrate that ordinary people have no determinate metaethical stances or commitments. Researchers could always insist that they do, but that existing methods have simply failed to reveal them. Such a task, if it is to succeed, will call for multiple, converging lines of evidence and a broader theoretical foundation than is feasible for the scope of this project. My efforts here should be seen, instead, as laying the groundwork for such a project, providing both a theoretical rationale for the plausibility of metaethical indeterminacy and preliminary empirical support.

It should also quickly become apparent *why* demonstrating metaethical indeterminacy will prove difficult. Although the folk metaethics literature is new and of modest size, no researchers have interpreted their findings as evidence of metaethical indeterminacy. Quite the contrary, all existing research supports a determinate account of folk metaethics, with the only question being what the distribution of different folk metaethical positions is (i.e., whether most people are some form of realist, antirealist, or a combination of the two). Overturning an entire literature is no easy feat. Nevertheless, I *do* believe I can decisively succeed at this task. The methodological critiques that I level against existing studies on folk metaethics reveal far more than the mere difficulty of measuring folk metaethics, they raise serious doubts about whether there is any viable way to measure folk metaethics *at all*. Even if the case for metaethical indeterminacy cannot be decisively established, but only gestured at, I nevertheless hope to show that existing studies fail to provide good evidence that people have

determinate metaethical stances and commitments. In other words, before demonstrating that I am right, I first set out to demonstrate that everyone else is wrong.

If I am correct that ordinary people have no determinate metaethical stances or commitments, the implications would be considerable. Whole research lines in metaethics will prove little more than misguided attempts to measure metaethical stances that most people simply don't have, or to measure commitments that are almost entirely absent from folk philosophy, and, where present, do not vindicate traditional metaethical accounts or existing interpretations of the data. This troubling possibility has been obscured by flawed research design, poor measurement, and a host of methodological problems (Beebe, 2015; Bush & Moss, 2020; Pölzler, 2018b; 2018c). Taken together, these issues give the superficial appearance of nascent fields of research that simply need to refine their tools before progress can be made.

In the chapters that follow, I will address these methodological shortcomings, and argue that researchers have not defended the most plausible interpretation of available data. Yet in doing so, I do not wish to give the impression that the solution is to devise better measures. Rather, I hope to show that the observed patterns of results, as well as the underspecified questions, ubiquitous confounds, pervasive ambiguity in scale items, and other methodological shortcomings are not merely the result of flawed research design, but an unavoidable byproduct of the mismatch between philosophical theories and ordinary thought and language (Bush & Moss, 2020). As I will show, most participants do not understand questions about metaethics in the ways researchers intend, and I will argue that a plausible explanation for this is not that researchers have yet to ask questions in a way participants interpret as intended, but that the types of responses researchers are trying to elicit simply aren't features of folk thought and language to begin with.

I don't believe the problem of folk indeterminacy is a problem exclusive to metaethics. There may be broad ramifications of the kind of analyses I marshal to support this conclusion. If an entire

field of research could present the superficial appearance of legitimate psychological phenomena where none exist, or at least exist in a form radically different from what researchers propose, could the same be true of other lines of research as well? Much to my alarm, I fear an answer in the affirmative.

Given these potentially broader implications, it would be reasonable to ask why I focus on metaethics in particular. Why choose an obscure field with a comparatively small body of literature? The main reason is simply that I'm interested in the topic. But focusing on folk metaethics is also strategic. I am more familiar with research on folk metaethics than any other area of study, and metaethics is my primary area of interest in philosophy. Most importantly, it is an area where the case for indeterminacy is especially strong. Finally, I also suspect that I can build a more persuasive case by focusing on a single topic than by spreading myself too thin by addressing many different areas of research.

If I am successful, this will open the door to future efforts to build similar cases for indeterminacy in other areas of folk philosophy. I am currently developing a similar case for indeterminacy about folk notions of free will and will turn my attention to other folk philosophical distinctions in the future.[9] But for now, folk metaethics is a rich, active, and evolving field of research that provides fertile ground for developing a template for similar arguments in favor of indeterminacy in other areas of folk philosophy. My critique of research on folk metaethics is thus intended also as a proof of concept for the broader possibility of folk indeterminacy with respect to other philosophical distinctions, and that other areas (or potential areas) of research on folk philosophy may be dead ends.

---

[9] I could have presented a similar case for indeterminacy with respect to other areas of folk philosophical research, such as *folk epistemology* (Beebe, 2012; Kim & Yuan, 2015; Machery et al., 2004; Nichols, Stich, & Weinberg, 2003; Weinberg, Nichols, & Stich, 2001) or *personal identity* (De Freitas et al., 2017; Molouki & Bartels, 2017; Shoemaker & Tobia, 2019; Starmans & Bloom, 2018a; 2018b; Strohminger & Nichols, 2014; 2015), among other possibilities. Hopefully researchers who are more familiar with these bodies of literature will investigate the possibility of indeterminacy in these areas of research as well. Even if there are determinate folk views with respect to these philosophical literatures, my approach may aid in refining the measurement tools used to assess other topics.

As bleak as this may seem, my hope is that coming to terms with the intractability of some research programs will redirect efforts towards new lines of research. After all, beliefs and attitudes related to many aspects of morality—including agency, responsibility, and normativity—are deeply embedded in our languages and cultures, and are thus plausibly part of our evolved psychology (Cosmides & Tooby, 2016; Devaine, Hollard, & Daunizeau, 2014; Mallon and Machery, 2010, Zahn-Waxler & Robinson, 1995). There must be patterns in the way people speak and think that could shed light on these and other topics that traditionally fall within the ambit of moral philosophy. The same is true of folk epistemology and metaphysics. Yet our understanding of how ordinary people think about these issues has been hindered by an overreliance on the traditional concepts, distinctions, and framing of questions that dominates analytic philosophy and Western philosophical traditions generally. We should draw insights from academic philosophy whenever possible, but I hope to show that attempts to force ordinary thought into the artificial strictures that animate historical philosophical disputes can distort and obscure the factors that *do* characterize folk philosophy.

On a personal note, these criticisms are not motivated by hostility towards research on folk philosophy or folk metaethics in particular. After years of frustrating attempts to study folk philosophy, I have grudgingly accepted that a great deal of projects are dead ends. Rather, I am motivated by the desire for philosophers and psychologists working at the intersection of our fields to conduct the highest quality research possible. I suspect that others share my feeling of never being completely at home in a philosophy or psychology department. But if we intend to craft an interdisciplinary home for ourselves, I want to ensure it is one we are proud to live in. Much as criticism of the rigor and methods of social psychology can be motivated by a love for the discipline and optimism about its future, I see my efforts not as a call to abandon the study of folk philosophy, but as an effort to rectify significant methodological shortcomings in existing research and, more importantly, to call on philosophers to show more appreciation for the potentially rich and surprising

ways in which folk philosophical thought might differ from academic philosophy. I have no doubt that I will err many times here and elsewhere, and that if my work receives any attention at all, much of it will be to correct such mistakes. I also have no illusions that the case presented here will convince everyone or stem the tide of misguided research on metaethics or folk philosophy in general, but at the very least, I hope my criticisms will contribute to the improvement of future research.

The chapter structure is as follows. In chapter two, I critique *the disagreement paradigm*, the most common method used to measure folk realism and antirealism. In chapter three, I critique the most prominent alternative measures to the disagreement paradigm and a few attempts to experimentally manipulate folk realism and antirealism. In both chapters, my goal is to show that existing methods suffer such substantial methodological shortcomings that they are unlikely to provide valid measures of folk metaethics. In chapter four, I present a series of studies that center on the collection and evaluation of open response data. My goal in this chapter is to demonstrate that ordinary people struggle to interpret questions about metaethics as intended. Chapter five is split into two sections. In the first section, I supplement the studies conducted in the previous chapter with a series of multiple choice studies and a study using Likert scale items that also assess interpretation rates for metaethics stimuli. Once again, I show that ordinary people struggle to interpret metaethics stimuli as intended. Finally, I present a study which asks participants to assess a highly sophisticated theoretical issue: different interpretations of quantum mechanics. I argue that it is highly implausible that ordinary people would have sufficient prior knowledge of quantum mechanics to have a meaningful philosophical stance, and that it is even less plausible that ordinary language is implicitly committed to distinct philosophical positions on quantum mechanics. Nevertheless, nonrandom patterns of data emerge when using a forced choice paradigm. This illustrates that even when we possess strong theoretical grounds for presuming folk *indeterminacy*, studies can create the superficial appearance of

folk *determinacy*. Finally, in the conclusion, I offer a brief commentary on the implications of my arguments and findings, and an optimistic prognosis for the future of research into folk philosophy.

I am also aware of the interdisciplinary nature of this project. I employ jargon in both philosophy and psychology, and introduce new terminology. I have put together an extended explanation of some of the major terms, themes, concepts, and arguments introduced here and throughout the dissertation. I also include caveats and qualifications to the central argument which were too lengthy to include in the main text. These are available in **Supplement 1**. For a brief summary of the relevant terms used in this chapter and subsequent chapters, please consult the glossary in **Appendix A**.

# CHAPTER 2:
## Critique of the Disagreement Paradigm

### 2.0 Introduction

Most research on whether ordinary people are moral realists or antirealists employs the *disagreement paradigm* (Bush & Moss, 2020). Researchers have also employed a variety of other measures, including scales (e.g., Collier-Spruel et al., 2019; Zijlstra, 2019), implicit measures (Wagner, Pölzler, & Wright, 2021; Zijlstra, 2021), fMRI (Theriault et al., 2017; 2020) and training paradigms (Pölzler & Wright, 2020a; 2020b; Wright, 2018). This chapter provides a summary of the many methodological reasons to doubt the validity of the disagreement paradigm.[10] I explore problems with other measures in the following chapter. Both chapters provide a cursory overview of the many methodological shortcomings with existing folk metaethics paradigms. Although their methods differ, most of these studies provide *prima facie* evidence of moral pluralism, though some researchers have argued that their findings, or folk metaethical findings, support folk uniformity (e.g., Zijlstra, 2021).[11]

Unfortunately, *none of these paradigms provide valid measures of folk metaethical stances or commitments*. Each study exhibits a host of methodological shortcomings that render its findings at best inconclusive. Of course, I cannot review every study. Instead, I have organized studies by the methods used, and chose examples that are most representative of those paradigms. There are no established conventions or standardized paradigms for folk metaethical research, so my organizational scheme will not perfectly distinguish all study designs, and some studies will exhibit characteristics that overlap with multiple categories or that will not comfortably fit into any particular category. In some cases,

---

[10] Although my critiques are more comprehensive than others, I am not the first to raise these or similar objections. Pölzler (2018a; 2018b) has previously raised some of the same criticisms of research on folk metaethics, and has raised some concerns that I don't focus on here. Although I've consolidated many critiques here, Pölzler's criticisms are excellent and worth considering as well.

[11] This includes uniform folk relativism (Beebe, 2020), a nearly uniform tendency towards antirealism, but pluralism about the type of realism (i.e., most people endorse cultural relativism, others individual subjectivism, and others noncognitivism; Pölzler & Wright, 2020a; 2020b: Wright 2018, Pölzler and Wright 2019).

specific studies exhibit virtues or shortcomings that are not present in the examples I use to represent that paradigm, or employ additional measures that I did not deem sufficiently important to consider. Where relevant, I will make note of the relevant idiosyncrasies of individual studies. Yet there are so many methodological problems with this research, both in general and with respect to individual studies, that there is no practical way to adequately present them in the main text. I provide a more comprehensive critique of the disagreement paradigm in **Supplement 2** and a more comprehensive critique of alternative measures in **Supplement 3**.

Philosophers may also question the value of research on folk metaethics. After all, aren't we more concerned with whether moral realism or antirealism are true than what people think or how they speak? Aside from the fact that the answer for some of us will simply be "no," there are still important questions about the relationship between academic metaethics and folk metaethics. I address this in **Supplement 1** where I argue that a great deal of academic metaethics *is* concerned with folk metaethics and that many disputes in metaethics turn in part on empirical questions about how ordinary people speak and think. Some researchers may also wonder whether research on the Moral/Conventional Distinction (the MCD) provides evidence of folk realism. I argue that it does not in **Appendix E**.[12]

### 2.1 The disagreement paradigm

There is no single, canonical form of the disagreement paradigm, but they all share a similar structure:

> (1) Typically, participants are first asked to rate how much they agree or disagree with a moral claim, e.g., *"Robbing a bank in order to pay for an expensive vacation is morally bad."*

---

[12] I must apologize for the inclusion of so many supplements and appendices. I have done so to spare the reader an interminable slog through the methodological musings that motivated the empirical portions of this project, which appear in chapters four and five. However, it would feel like a betrayal of my principles to simply omit such discussions from my work. One of the central objectives of this project is to criticize the tendency for researchers to dive straight into experimental research without adequately situating their research on solid theoretical foundations reinforced by substantive descriptive findings. It would therefore make little sense to conceal the theoretical contributions I can offer when they can be made available to those interested in advancing the study of folk metaethics.

(2) Next, they are presented with a disagreement about the moral claim. This disagreement can be between themselves and another person, or between two other people.

(3) Finally, they must select a response that indicates their metaethical position. Usually, this involves a set of multiple choice response options that reflect different realist and antirealist positions.[13]

Precise wording may vary, but all versions of the disagreement paradigm attempt to distinguish realism from antirealism by asking whether *(i) both people can be correct* (which indicates *antirealism*), or if *(ii) at least one person must be incorrect* (which indicates *realism*). For example, Beebe and Sackris (2016) present the disagreement paradigm as follows:

> *If someone disagrees with you about whether [claim], is it possible for both of you to be correct or must one of you be mistaken?*
> ❏ *At least one of you must be mistaken*
> ❏ *It is possible for both of you to be correct*[14,15]

The judgment that *at least one person must be mistaken* is interpreted as *realism* because it seems to reflect the view that there is a single stance-independent standard of truth. Conversely, the judgment that both people could be correct is interpreted as an expression of *antirealism* because, if a person that judges that a moral claim can be *correct* according to one person's moral standards but *incorrect* according to another person's standards, they seem to believe the truth of moral claims can only be judged relative to each person's moral standards and can vary depending on what that person believes. This judgment reflects a form of *relativism*, which for our purposes is a form of *antirealism*.[16]

---

[13] Studies sometimes differ. Sarkissian et al. (2011) present a realist statement then ask participants how much they agree or disagree with it on a 7-point Likert scale.

[14] I changed the order of the response options because it seems more natural to describe realism first, then antirealism.

[15] I chose Beebe and Sackris's version of the disagreement paradigm for its clarity and simplicity.

[16] To illustrate the rationale behind the paradigm, consider how we judge disagreements about nonmoral issues. Some disagreements involve mutually incompatible truth claims. If Alex claims that "Water is H2O," and Sam claims that "Water is not H2O," we know they cannot both be correct, because these claims represent jointly exhaustive and mutually incompatible descriptions of reality. But if Alex claims that "Cheese is delicious," and Sam claims that "Cheese is not delicious" it is not obvious that one of them must be incorrect. Perhaps these statements are best understood not as competing truth claims, but as indexicalized expressions of the subjective preferences of the individual making the claim. That is, Alex could mean "I find cheese delicious," while Sam could mean "I don't find cheese delicious." If so, Alex and Sam would each be making an assertion that is true *relative to their own preferences*. These statements may *appear* to conflict with one another, but this is simply because the indexical element in each statement is implicit. Once it is made explicit,

We could ask participants whether people who disagree about morality (in general) can both be correct or if at least one must be incorrect, but most studies present participants with a series of concrete moral claims (e.g., murder, abortion, donating to charity, etc.) and repeat the same set of questions for each moral issue. This allows researchers to assess variations in metaethical judgments towards distinct moral issues. For instance, a person may provide a realist response for the claim "murder is wrong," but an antirealist response for the claim that "abortion is wrong." Most researchers also include questions about nonmoral issues, such as claims about science or history, social conventions, and matters of taste or aesthetic judgment. This allows for cross-domain comparisons. For instance, researchers may hypothesize that most people will consistently choose the *realist* response for disagreements about science, but *antirealist* responses for social conventions and matters of taste. Researchers then assess the total proportion of *realist* and *antirealist* responses participants provided for each moral and nonmoral claim. This allows them to assess the proportion of realist and antirealist responses for each moral and nonmoral issue individually and to assess the overall proportion of realist and antirealist responses within each domain.

Most studies find that the majority of participants endorse realism for some moral issues and antirealism for others. In fact, *all studies that employ the disagreement paradigm find at least some evidence of intrapersonal variation*. While some participants do provide uniformly realist or antirealist responses, they are typically in the minority (but see Pölzler & Wright, 2020a; 2020b for notable exceptions). In other words, most studies suggest not only that pluralism is a common response, but that *most* people appear to be metaethical pluralists. These studies also reveal consistent evidence of *interpersonal variation*, since

---

any apparent conflict dissolves. We are already familiar with indexical claims, and have no trouble understanding them in other contexts or when they are made explicit. Obviously, there would be no conflict if Alex were to say, "I am Alex," and Sam were to say, "I am not Alex." Both statements would be true and obviously don't conflict with one another.

some participants exhibit a consistent inclination for realism, while others are more disposed towards antirealism.[17]

Many researchers have accepted these findings at face value, i.e., people appear to be metaethical pluralists because they *are* metaethical pluralists (e.g., Colebrook, 2020; Pölzler & Wright, 2018; 2020a; 2020b; Wright, Grandjean, & McWhite, 2013; Wright, 2018). However, others have characterized the overall evidence as favoring a uniform account, or at least a substantially uniform account even if they acknowledge a bit of pluralism (e.g., Beebe, 2020; Goodwin & Darley, 2008; Sarkissian et al., 2011; Zijlstra 2021), while still others have expressed skepticism about the plausibility of pluralism on theoretical grounds (Johansson & Olson, 2015; Sinnott-Armstrong, 2009). Finally, one could favor metaethical indeterminacy by arguing that existing studies are invalid and contend that they show neither pluralism nor uniformism because ordinary people don't have determinate metaethical stances or commitments. This leaves us with three competing hypotheses about folk metaethics: *uniformism*, *pluralism*, and *indeterminism*.

> *(i)* **Uniformism** holds that there is a single determinate account of folk metaethics with respect to a particular metaethical distinction.

> *(ii)* **Pluralism** holds that there is ineliminable but determinate variability with respect to one or more traditionally competing accounts of folk metaethics.

> *(iii)* **Indeterminism** holds that there is no determinate position that characterizes some or all folk metaethics with respect to one or more metaethical distinctions.[18]

---

[17] Since every study finds evidence of metaethical variability, this might appear to be overwhelming evidence of *pluralism*. If *uniformism* were correct, we would expect little interpersonal or intrapersonal variation. This doesn't mean we should expect *perfect* uniformity. People can be subject to performance errors, or misunderstand questions, or lack competence with regard to the relevant terms or concepts, or even deliberately select responses that do not reflect their beliefs out of spite or amusement. The uniformist is not, therefore, committed to the position that *all* respondents should *always* choose the same response in practice. They need only capture a consistent and strong majority. Yet even this low bar cannot be met. Instead, we find most major metaethical positions well-represented among participants. If these findings accurately reflect the metaethical stances or commitments that participants genuinely hold, this would be straightforward evidence of metaethical pluralism, at least among the populations these samples represent (i.e., mostly WEIRD populations; see **Supplement 2**).

[18] Each of the hypotheses may also apply to one dispute but not another, and are thus not incompatible in that respect, either. It may be that folk metaethics is determinately cognitivist, and that prototypical moral claims are thus best understood as propositional claims (i.e. *uniformism* towards cognitivism), while there could also be no determinate answer

At present, all published interpretations of folk metaethical data support either uniformism or

pluralism. Though Gill (2008; 2009) was the first to defend metaethical indeterminacy, this proposal

was coupled with pluralism and preceded the bulk of empirical research on folk metaethics, though it

nevertheless served as an impetus for this project. Thus, to my knowledge, no one has argued that

indeterminism (with a qualified splash of pluralism) may offer a plausible explanation of current

empirical findings and the position most likely to be vindicated by future research.[19,20] In the remainder

of this chapter, I review the state of the empirical research on folk metaethics and argue that existing

data is too methodologically flawed to support uniformism or pluralism.[21]

## 2.2 Overview of critique of the disagreement paradigm

There are many reasons to doubt the disagreement paradigm's ability to reliably distinguish realists

from antirealists. Documenting all these shortcomings may seem like overkill. But I do so to make a

point: measuring folk metaethics is *difficult*. By reviewing the many pitfalls that researchers face when

studying folk metaethics, I intend not merely to illustrate that the disagreement paradigm should be

discarded, but to gesture towards the more pessimistic conclusion that surveys are not up to the task

---

as to whether these assertions are relative or not (i.e., *indeterminism* towards relativism vs. nonrelativism), or there could be instances of relativism and nonrelativism that cannot be explained away as conceptually confused or nonstandard (i.e., *pluralism* about relativism vs. nonrelativism, Gill, 2008, p. 218).

[19] However, Gill (2009) does mention early empirical evidence that supports metaethical pluralism.

[20] According to Gill (2009), it is:

"...quite likely that meta-ethical indeterminacy characterizes much of ordinary discourse. But it may not characterize all of it. The best descriptive analysis of some other uses of moral terms might involve robust meta-ethical commitments. Those commitments, however, might not all be uniformly consistent with one side of the traditional meta-ethical debate over the other." (p. 218)

I agree. But where Gill and I differ is that I emphasize indeterminacy to a much greater extent.

[21] Pluralism and indeterminism are not mutually incompatible but may instead characterize subsets of folk thought and language. I discuss this in **Supplement 1.** Note that the compatibility of pluralism and indeterminacy could take multiple forms. It could be that some people's moral thought and language is consistently determinate but other people's moral thought and language is consistently indeterminate, or it could be that moral thought and language tend to be determinate in particular contexts but not others, or towards particular moral issues but not others, or among the members of some communities but not others, or among adults but not children, etc. There are too many possibilities for me to entertain all of them.

of evaluating folk metaethics.[22] Pölzler (2018b; 2018c) has done much of the work in documenting these shortcomings, and several of the concerns I raise summarize or are similar to points he has made. However, I raise a variety of novel concerns and add my perspective to criticisms originally raised by Pölzler.

All criticisms of the disagreement paradigm are summarized in **Table 2.1**. Including each subcategory of a particular criticism, there are at least 28 methodological criticisms. It would be impossible to discuss each in detail. I will discuss a handful in detail and provide a summary of the rest. If any of these criticisms seem obscure, confusing, implausible, or underdescribed, see **Supplement 2** for a more complete discussion.

**Table 2.1**

*Summary of methodological problems with the disagree paradigm*

| | Problem | Description |
|---|---|---|
| 1 | *Inadequate response options* | Missing options: Typical response options are not mutually exhaustive and often fail to include standard forms of realism and antirealism |
| 1.1 | *Missing noncognitivism* | Studies typically exclude a response option for noncognitivism |
| 1.2 | *Missing error theory* | Studies typically exclude a response option for error theory |
| 1.3 | *Missing subcategories of relativism* | Studies typically fail to distinguish different forms of relativism (e.g., subjectivism, cultural relativism, appraiser relativism, agent relativism) |
| 1.4 | *Missing other subcategories* | Studies typically fail to distinguish different forms of realism, including naturalism, non-naturalism, constructivism, relation-designating accounts, incoherentism, hybrid theories, indeterminacy, pluralism, and quietism |
| 2 | *Poor specificity* | Typical response options have poor specificity: they cannot distinguish different metaethical positions from one another |
| 3 | *Conflations* | Stimuli is often interpreted in unintended ways due to conflating realism and antirealism with unintended concepts and distinctions |
| 3.1 | *Conflating metaethics with normative ethics* | Participants often conflate metaethics stimuli with *normative* moral concerns. This undermines the validity of the disagreement paradigm since many |

---

[22] Pölzler (2018) was the first to document many of the methodological shortcomings discussed in this section. I am grateful for his clear and careful documentation of many of these shortcomings.

| | | responses express the participant's normative moral views, *not* their views on realism or antirealism |
|---|---|---|
| 3.2 | *Conflating metaethics with epistemic concerns* | Participants often conflate metaethics stimuli with *epistemic* concerns. This undermines the validity of the disagreement paradigm since many responses express the epistemic judgments, *not* their views on realism or antirealism |
| 3.3 | *Conflating realism with universalism* | Participants often conflate realism with *universalism*. This undermines the validity of the disagreement paradigm since many responses express judgments about universalism, *not* their views on realism or antirealism |
| 3.4 | *Conflating realism with absolutism* | Participants often conflate realism with *absolutism*. This undermines the validity of the disagreement paradigm since many responses express judgments about absolutism, *not* their views on realism or antirealism |
| 3.5 | *Conflating relativism with contextualism* | Participants often conflate relativism with *contextualism*. This undermines the validity of the disagreement paradigm since many responses express judgments about contextualism, *not* their views on realism or antirealism |
| 3.6 | *Conflating relativism with descriptive claims* | Participants often conflate relativism with *descriptive claims*. This undermines the validity of the disagreement paradigm since many responses express judgments about descriptive claims, *not* their views on realism or antirealism |
| 4 | *Evaluative standard ambiguity* | Studies often fail to provide context that would resolve ambiguities about which moral standards people who disagree are indexing. Providing context may fail to resolve the problem since it may not be salient and may not provide a valid measure for all forms of realism (e.g., appraiser relativism) |
| 5 | *Abstract norm ambiguity* | Moral disagreements can be readily attributed to different ways of conforming to the same abstract moral principles. Participants often judge two people could both be correct because both actions conform to the same abstract moral principle, *not* because the participant endorses relativism. |
| 6 | *Misattributing source of disagreement* | People may misattribute the source of disagreement, e.g., to differences in nonmoral beliefs or misunderstanding the question, rather than a difference in fundamental moral values. |
| 7 | *Domain classification inconsistency* | Studies rely on *a priori* categorization of issues as "moral." However, research demonstrates that ordinary people (a) exhibit high levels of variability in which issues they consider moral *within* samples, (b) high levels of variability *between* populations and (c) systematically differ from researchers in what they consider moral issues. This can introduce considerable noise and interpretative difficulties and may reduce measurement invariance |
| 8 | *Presumption of correspondence theory of truth* | As Pölzler and Wright observe, studies on realism and antirealism presume participants are committed to a *correspondence theory*. However, there is little empirical evidence suggesting most participants endorse or are committed to correspondence theories of truth in relation to moral claims. If they do not, it is not clear if their responses are valid. |
| 9 | *Signaling and reputational concerns* | Participants may be motivated to select response options in order to signal desirable character traits rather than because those responses accurately reflect their views on realism and antirealism. |

| 10 | *Lack of realism* | Stimuli suffer a variety of problems that undermine their capacity to induce consistent, intended psychological states |
|---|---|---|
| 10.1 | *Lack of experimental realism* | Stimuli are often implausible or unrealistic, which can reduce engagement |
| 10.2 | *Lack of mundane realism* | Stimuli are often underspecified. This may force participants to "fill in the blanks" with their own inferences, decreasing consistency in interpretations |
| 10.3 | *Lack of psychological realism* | Stimuli sometimes include humorous or unrealistic stimuli (e.g., aliens that want to turn everything into pentagons), which can prompt unintended psychological states |
| 11 | *Lack of external validity* | Studies exhibit poor external validity due to overreliance on WEIRD populations, unrepresentative response options, and lack of ecological validity |
| 11.1 | *Overreliance on WEIRD populations* | Most studies are conducted on WEIRD populations and college students, limiting generalizability |
| 11.2 | *Stimulus as fixed effect problems* | Studies employing concrete moral issues are subject to the stimulus-as-fixed-effect fallacy |
| 12 | *Forced choice obscures indeterminacy* | Studies force participants to choose from available response options. This can create the misleading appearance of determinacy |
| 13 | *Inaccurate, biased, or misleading stimuli* | A problem originally described by Pölzler (2018b; 2018c) stimuli may mislead or bias participants, prompting unintended interpretations or motivating people to choose responses that inappropriately favor particular response options, though stimuli may also simply be inaccurate |
| 14 | *Questionable a priori categorization* | Hypotheses are derived primarily from concepts and distinctions drawn from contemporary analytic metaethics, a small, elite academic field almost exclusively composed of people from WEIRD populations, whose ways of thinking may not reflect how ordinary people think |
| 15 | *Intrinsic complexity* | The metaethical concepts studies are designed to assess are: (i) complex and difficult to grasp (ii) unfamiliar to ordinary people (iii) abstract and distant from everyday contexts of moral judgment (iv) there are typically unintended interpretations of stimuli that are more plausible than the intended (metaethical) interpretation The complexity of metaethical concepts makes efforts to mitigate the conflations and ambiguities outlined in this list difficult, if not infeasible, using conventional social scientific methods. |
| 16 | *Cumulative invalidity* | In isolation, the preceding problems may not decisively threaten the validity of the disagreement paradigm. But the cumulative effect of so many issues does. Adequately addressing all such issues may be too difficult without training participants, which could result in spontaneous theorizing or cause them to differ so much from the population they were sampled from that studies lose external validity |

**2.3 Inadequate response options**

One problem with standard versions of the disagreement paradigm is lack of response options that reflect the full range of possible metaethical stances and commitments. There are two reasons why the response options used in standard versions of the disagreement paradigm are inadequate: (1) response options are *not jointly exhaustive*, and thus fail to provide response options that reflect the full range of possible metaethical positions[23] and (2) response options are typically *low resolution*: they *may* tell us whether the participant is a realist or antirealist, but even still, they don't tell us *what kind*.

### *2.3.1 Response options are not exhaustive*

Although there is an overarching division between metaethical *realism* and *antirealism*, both positions may be further subdivided into a number of subcategories. Realists may be subdivided into naturalists and non-naturalists, while antirealists are generally divided into relativists, noncognitivists, and error theorists, though antirealism also includes constructivist and response-designating accounts, as well as less common positions, such *incoherentism* (Loeb, 2008) and my own view, *metaethical quietism*.[24] Thus, both realism and antirealism are umbrella terms that sit at the top of a hierarchy of lower-level categories and subcategories.

In principle, this does not pose any challenge to the validity of the disagreement paradigm. The goal of standard versions of the disagreement paradigm is to simply determine whether the

---

[23] Pölzler (2018a; 2018b) discusses this shortcoming as well.

[24] Metaethical quietism is the view that the dispute between realism and antirealism is a pseudoproblem resulting from linguistic and conceptual confusions. In my own case, this is paired with the position that nontrivial accounts of realism are unintelligible. I cannot elaborate on these views here. Roughly, I maintain that notions such as irreducible normativity, or standard accounts of "reasons," as they figure in contemporary analytic metaethics, are unintelligible, and thus by extension any metaethical account that appeals to such concepts is likewise unintelligible. Other accounts are merely *trivial* in a technical sense, which in most instances results from the account in question bottoming out in inert descriptive claims that lack the kind of authoritativeness or clout to compel action (e.g., typical naturalist accounts of realism). Although I reject the realism/antirealism dispute as a pseudoproblem, I still reject realism, making me an antirealist. However, since I reject other antirealist positions for buying into many of the conceptual errors that realism is subject to (e.g., confusions about "reasons"), I reject all conventional forms of antirealism as well. Given that my position rests on metaphilosophical foundations that reject the very framework in which metaethical positions are ordinarily classified, my position is difficult to place in conventional taxonomies.

participant believes (or is committed to) the claim that there are stance-independent moral facts (*realism*) or not (*antirealism*).[25] Yet in practice this is *not* what standard versions of the disagreement paradigm measure. Instead, participants are only able to judge whether two people who disagree about a moral issue (a) *are both correct* or if (b) *at least one must be incorrect*. This does not represent a jointly exhaustive pair of options. Participants could, in principle, think that *neither* side of a moral dispute could be correct. Yet they have no way to express this. Such a view is consistent with *noncognitivism*. Noncognitivism is the view that moral claims don't express propositions, and are thus incapable of being true or false. For instance, the moral claim "murder is morally wrong" could be understood as an imperative ("don't murder") or an emotional expression ("murder?! Boo!").

The failure to include adequate response options can also systematically *miscategorize* participants. Error theorists would have to choose option (b), and be classified as realists, even though error theory is an antirealist realist position. Error theorists believe that all claims about what is morally right or wrong are false, because they are implicitly committed to a false presupposition. For instance, if "stealing is morally permissible" presumes that stealing is *stance-independently* permissible, but there are no stance-independent moral facts, then this and all other moral claims like it are false, including the contrary moral claim that "stealing is morally impermissible." Since error theorists would agree that "at least one person must be incorrect," they would be included among realists. Changing the response option doesn't help either, e.g., changing it to "at most one could be correct" or indicating that just one of the two positions is correct isn't available to the antirealist, because they don't think anyone is correct.[26]

---

[25] There is nothing wrong in principle with a simple, high-level measure that isn't designed to assess subcategories. If you want to study whether people like pizza, you don't *have to* exhaustively canvass everyone's favorite pizza toppings or risk having useless data. Just the same, it would be valuable in principle to know whether people are realists or antirealists.

[26] Technically, an error theorist could think it is logically possible that one of the positions is correct, but that it isn't *metaphysically* possible. I suspect researchers would have a difficult time designing measures that could detect this distinction.

Both the noncognitivist and the error theorist would think that *neither* side of a moral dispute could be correct, either because nobody could be correct *or* incorrect (*noncognitivism*) or because everyone is incorrect (*error theory*). Participants who hold such views have no way to adequately express their positions, leaving them without the ability to provide an accurate response to standard versions of the disagreement paradigm. Newer versions of the disagreement paradigm have corrected for this, by including noncognitivism or both noncognitivism and error theory (Beebe, 2015; Davis, 2021; Pölzler & Wright, 2020a; 2020b). However, this may solve one problem only to introduce another. Noncognitivism and error theory are subtle, sophisticated philosophical positions. By including additional response options, researchers increase the cognitive load on participants and risk introducing response options that are difficult to interpret as intended (Bush & Moss, 2020). Whatever these costs may be, the exclusion of legitimate response options is worse, since previous studies show that people do choose such options when they're provided. Forcing participants to choose from a narrow set of responses that they wouldn't otherwise choose risks undermining the validity of one's measures by mixing participants who don't endorse the response options they selected with those who do, and requiring participants to choose response options that don't reflect what they think. Such added complexity would only scratch the surface. I haven't even addressed the typical exclusion of different forms of realism (e.g., naturalism and non-naturalism) or antirealism (e.g., constructivism, relation-designating accounts, and so on), with the exception of Davis (2021), who does address different forms of realism, no other studies address any of these response options, and if they did, the complexity and number of response options would quickly become unmanageable.

### 2.3.2 Poor specificity

Standard versions of the disagreement paradigm also suffer from poor specificity. Even if the disagreement paradigm could tell us whether people are realists or antirealists, typical versions can't tell us *what kind* of realist or antirealist people are. Without additional follow-up questions, we cannot

further determine whether a realist endorses naturalism or non-naturalism, or whether an antirealist endorses relativism, noncognitivism, error theory, or some other position. The low resolution of the disagreement paradigm does not render it invalid. It simply means that it could at best only provide limited information.

One solution to this limitation is to ask a variety of auxiliary questions that allow one to provide a more fine-grained assessment of each participant's metaethical position. This is precisely what Davis (2021) did. Participants were asked whether they believed in God before

follow-up questions, which influenced the subsequent set of questions they were asked. Once they were presented with the disagreement paradigm, participants were able to choose realism, relativism, error theory, or noncognitivism. Those who chose realism were then presented with a set of follow-up questions to determine whether they endorsed a supernatural, natural, or non-natural conception of realism. This is a step in the right direction, but it introduces a variety of potential concerns. Briefly, (1) the task becomes longer and more cognitively demanding, (2) introducing additional stimuli results in new ways to increase unintended interpretations and interpretation variation, (3) follow-up and auxiliary questions may still employ forced choices that limit response options and generate the artificial appearances of determinacy, (4) introducing additional questions can introduce novel methodological problems, (5) such approaches rely on potentially controversial judgments by researchers on how to classify different metaethical positions.

The disagreement paradigm also does not typically include different response options for different forms of relativism (Pölzler, 2018a; 2018b). In particular, there is no distinction between *subjectivism*, the view that moral claims are true or false relative to the standards of *individuals*, and *cultural relativism*, the claim that moral claims are true or false relative to the standards of different *groups*. Once Pölzler & Wright (2020b) included distinct options for each, they found a substantial number of participants who endorsed one or the other. This is yet another limitation that can be circumvented

by including additional response options, but once again at the cost of increased complexity and risk of unintended interpretation.

A less discussed limitation is the absence of any attempt to distinguish *appraiser* and *agent* relativism (Quintelier, De Smet, & Fessler, 2014). Relativism allows the truth of a moral claim to be relativized to the standards of the agent performing an action (agent relativism), the standards of those judging the agent judging the action (appraiser relativism), or both. Only one study tested for both, and found evidence of *both*, leading the authors to conclude that "there is inter-individual as well as intra-individual variation in whether individuals relativize moral standards to agents or appraisers" (p. 227).[27]

In short, the disagreement paradigm does not typically allow us to measure people's metaethical stances or commitments with any level of specificity, and any attempt to do so faces a tradeoff: the higher the resolution, the harder the task. With more response options, follow-up questions, and auxiliary questions used to pin down the precise metaethical stances or commitments of the participant, the more onerous the task and the greater the risk that something *else* goes wrong. This would not be a problem if the risk of something else going wrong were low, but as I show in the remaining sections and chapters, the risk is anything but low.

## 2.4 Conflations with unintended concepts

This section presents the most substantive and serious set of methodological shortcomings with the disagreement paradigm: *conflations. Conflations* occur whenever participants interpret stimuli in a particular unintended way, such that their response is influenced by (or fully reflects) their stance or commitment towards *something other* than what researchers were intending to measure.

---

[27] Notably, they acknowledge the increased complexity of drawing so subtle a distinction, adding that "Notwithstanding the fact that we excluded all participants who did not fill out all comprehension questions correctly, given the complexity of our scenarios and questions, future investigations would benefit from simpler materials" (p. 227).

There are two types of conflation: *formal conflations* and *informal conflations*. Formal conflations occur whenever stimuli *itself* conflates the intended interpretation with one or more unintended interpretations. Such errors can cause participants to systematically interpret stimuli in unintended ways due to researcher error, rather than participant misunderstanding. *Informal conflations* occur whenever participants conflate stimuli that have been presented accurately (i.e., without conflations) but participants nevertheless conflate the intended interpretation with some other interpretation(s), resulting in a response that does not reflect the measure of interest. Formal conflations could, in principle, be eliminated, though it may be infeasible to eliminate informal conflations. As such, the latter present the greater potential threat to the validity of the disagreement paradigm. The findings discussed in **Chapter 4** and **Chapter 5** suggest that unintended interpretations are the primary threat to folk metaethics research, and these unintended interpretations are, in turn, largely due to participants conflating stimuli intended to elicit judgments about realism and antirealism with other distinctions. Here, I document the most common conflations.

### 2.4.1 Conflating metaethics with normative ethics

Metaethical distinctions concern second-order questions about the nature of morality. Metaethics is distinct from *normative ethics*, which is concerned with first-order questions about what is in fact morally right or wrong, permissible or impermissible, etc. When a person judges that e.g., "murder is wrong," this is a *first-order* (normative) position. When they judge that there is a stance-independent fact about whether murder is wrong, this is a *second-order* (metaethical) position. In other words, metaethical positions are "philosophical views *about* such first-order moral judgment" (Pölzler, 2018b, p. 657; emphasis original; see also Huemer, 2005, pp. 1-2).

Participants responding to the disagreement paradigm often appear to conflate metaethical considerations with normative considerations. For instance, suppose Alex expresses the view that abortion is morally permissible. Then Alex is told a previous participant judged that abortion is *not*

morally permissible. Alex is then asked whether both she and the previous participant can be correct, or if at least one of them must be incorrect. Alex might reason as follows: "Well, abortion isn't morally wrong, so the other person is mistaken." Alex may then select the response option "at least one of us must be incorrect." This would be interpreted as a *realist* response, yet Alex's thought process had nothing to do with realism. Alex was simply expressing a first-order moral judgment: abortion isn't morally wrong, so someone who thinks that it is wrong is mistaken. When asked to explain their answers, many participants expressed first-order judgments of this kind, or, alternatively, simply stated that the person who disagreed with them was mistaken or a bad person (see **Chapter 4**, Study 1). Unfortunately, such responses cannot tell us whether the participant is a realist or not, since we cannot be sure whether the response option they selected reflects their metaethical position or simply reiterates their normative stance.

Unfortunately, participants are also far more familiar with expressing first-order moral judgments, and may have trouble "turning off" their tendency to think in terms of what's right or wrong, rather than in terms of abstract considerations about what it means for something to be morally right or wrong. This possibility prompted Pölzler (2018b) to suggest that "Avoiding first-order moral intuitions in studies on folk moral realism altogether may be methodologically infeasible" (p. 658). If so, normative moral standards may represent an ineliminable impediment to measuring metaethical stances and commitments. At best, we may be able to mitigate the distorting influence of normative judgments, but doing so may come at the cost of further complicating studies and may limit the kinds of moral issues we can assess (e.g., we may not be able to assess more emotionally charged moral issues).

### 2.4.2 Conflating metaethics with epistemic concerns

Metaethical realism and antirealism are typically construed positions on the truth status of moral claims with *metaphysical* but not necessarily *epistemic* implications.[28] Metaphysical considerations are distinct, but related, to *epistemological* questions about how we can acquire moral *knowledge*, whether (and how) our moral beliefs can be *justified*, and whether we can be *certain* of our moral views (Pölzler, 2018). There is no easy way to disentangle metaphysical and epistemological considerations, since metaphysical stances often have epistemological implications, and vice versa. For example, if error theory or noncognitivism are true, then it is impossible to have moral knowledge because there are no moral facts. On occasion, philosophers also incorporate epistemic stances in their characterizations of moral realism, e.g., some claim that moral realism requires that we have (or can have) knowledge of at least some moral facts (Miller, 2009; Sinnott-Armstrong, 2009).[29]

Despite their close relationship, questions about folk metaethical realism and antirealism are not intended to directly assess epistemic considerations about the means or possibility of moral knowledge or justification. Rather, they are concerned exclusively with whether there are moral facts, and if so, whether those facts are stance-independent (which are metaphysical or conceptual questions, *not* epistemic questions). Skepticism about moral knowledge is entirely consistent with the belief that

---

[28] I say "typically" because Parfit and Scanlon apparently maintain non-metaphysical notions of realism (see Veluwenkamp, 2017). This is likely a very uncommon view, and it's unclear how plausibly it can be maintained. Nevertheless, Parfit is quite explicit on this point. As Veluwenkamp notes:

Parfit maintains that in the normative domain these truths have "no positive ontological implications" and are not "about metaphysical reality" [...] And for Scanlon, normative truths "need no natural or special metaphysical reality in order to have the significance that we commonly grant them" [...]. (p. 751, see Parfit, 2011, vol. 2, p. 479, p. 747, and Scanlon, 2014, p. 52)

I have no idea what they are talking about. Of course moral realism has metaphysical implications. At the risk of sounding impertinent towards eminent scholars, I think these remarks are implausible and desperate attempts to insulate realism from objections through little more than stipulative fiat.

[29] This is a reasonable criterion to include, since it allows the realist to exclude undesirable forms of skeptical moral realism: that there are moral facts, but we can't know any of them. People who believe that there are stance-independent moral facts are typically animated not just by the belief that they exist, but by the confidence (or at least hope) that we already know, or could eventually come to know, at least some of those moral facts.

there are stance-independent moral facts. Conversely, the belief that we can have moral knowledge may not directly entail whether that knowledge is of stance-independent facts or knowledge or relativized or response-dependent moral standards. In other words, skepticism about moral knowledge is consistent with realism, while a belief that moral knowledge is possible is consistent with antirealism. As a result, questions that do not neatly distinguish epistemological and metaphysical considerations from one another risk being unable to identify whether a response reflects an epistemic stance, a metaphysical stance, or both, and thus cannot serve as valid measures of realism/antirealism.

There are many ways in which participants could conflate metaethical and epistemic considerations. Here, I will focus on just one: *modal operator scope ambiguity*.[30] Consider the following claim:

(i) Brazil and Germany could both win the World Cup this year.

This sentence could be making one of two distinct claims:

(a) *Either* Brazil could win the World Cup this year, *or* Germany could win the World Cup this year, *but they cannot both win.*

(b) Brazil and Germany could both win the World Cup this year *at the same time.*

To formalize this ambiguity, take ♦ *(x)* to mean that x is possible, take W to refer to the predicate "win the World Cup", and take b to refer to Brazil and g to refer to Germany. This leaves us with two formalizations of the possible meaning of (i):

(a) ♦ *(Wb)* & ♦ *(Wg)* (exclusive)
(b) ♦ *(Wb & Wg)* (inclusive)[31]

(i) is ambiguous between the first, *exclusive* reading, and the second, *inclusive* reading. In practice, context and pragmatics will tend to resolve this ambiguity. Given our background knowledge of the World Cup and the rules of tournaments in general, most of us know that only one team can win at a time,

---

[30] This problem was originally proposed by my colleague, Tyler Millhouse.
[31] Tyler Millhouse formalized the inclusive and exclusive readings of the disagreement paradigm (Millhouse & Bush, 2016).

and this will lead us to conclude that when someone asserts (i), the exclusive interpretation is more plausible than the inclusive one. In other cases, context will tend to favor the opposite interpretation. For instance, consider the statement:

(ii) Alex and Sam could both have cereal for breakfast this morning.

Despite its similarity to (i), there is little reason to assume that two people would be unable to eat the same thing for breakfast, so most people would lean towards an inclusive reading of this statement.

It is unclear how people interpret moral disagreements that exhibit the same ambiguity. However, versions of the disagreement paradigm closely mirror the wording used in these examples. For instance, Beebe (2015) uses the following response options:

(1) It is possible for both of you to be correct.
(2) At least one of you must be mistaken.

The first option, "it is possible for both of you to be correct," invites the same ambiguity between an inclusive and exclusive reading. This may lead participants to be uncertain how to resolve this ambiguity. They may interpret in the intended exclusive way, but they may also interpret in the inclusive way. One way the inclusive interpretation makes sense is if the participant is *uncertain* about whether the moral claim is true or false, and recognizes that future information could reveal that either they or the person they disagree with could ultimately turn out to be correct. If so, judging that they could both be correct would not represent a relativist stance because they will have interpreted the question as an (incomplete) *epistemic* question about moral knowledge, and not a metaethical question about whether conflicting moral claims are mutually exclusive.

This may be the best explanation for some otherwise puzzling results. Beebe et al. (2015) and Beebe and Sackris (2016) asked participants to judge disagreements about empirical claims, such as the age of the earth or the size of planets. Most people probably believe that claims like these have a single established answer that they could verify for themselves. But Beebe and colleagues also included

claims that may have a single, objectively correct answer in principle, but lack any practical means of verification; or, as they put it, are "not only unknown but practically unknowable" (Beebe & Sackris, 2016).

The most intriguing example involved conflicting claims about whether Julius Caesar ate soup on his 21st birthday.[32, 33] If participants interpreted this question in the intended *inclusive* reading, they would have to consider whether two logically incompatible historical claims could both be true *at the same time*. It is possible that some participants might endorse a radical form of relativism, and believe that truth claims even in the domain of historical truth are relative. Nichols (2004) found some evidence of this in the explanations participants gave for their responses. However, these individuals represented only a handful of participants. Beebe & Sackris (2016) found that 45% of US participants judged that both people could be correct in 45% of the disagreements about Julius Caesar, while a similar pattern held for participants in China (53%), Poland (48%), and Ecuador (53%, Beebe et al., 2015). One plausible explanation for these findings is that participants interpreted what they were being asked as an epistemological question, and gave an exclusive reading of the question about whether both could be correct.

Scope ambiguity represents just one of the ways participants could conflate epistemic considerations with metaethical considerations. Even if scope ambiguity were minimized or eliminated, ordinary people may still be prone to conflating epistemic and metaphysical considerations more generally. Disentangling epistemic and metaphysical considerations is sufficiently difficult that training in philosophy may be necessary to develop the competence to recognize the distinction and tease it apart, and even philosophers, on occasion, slip in doing so. Like normative conflations, it may

---

[32] Beebe et al. (2015) adapted the materials used by Beebe and Sackris (2016) to populations in Poland, China, and Ecuador, and modified the name of the historical figure to one that was more culturally relevant, e.g., Confucius.
[33] Beebe & Sackris (2016) asked about Caesar drinking wine rather than eating soup as in Beebe et al. (2015), but this difference is irrelevant to any points that I make here.

be difficult or impossible to fully eliminate the unintended influence of epistemic considerations on ordinary people's response to metaethical stimuli, and in practice, it may require so much training that it would be infeasible to provide participants with the requisite training; it may even be impossible, since the amount of training required would result in participants no longer being "ordinary people." We wouldn't expect, for instance, ordinary people to possess the requisite expertise to draw difficult, counterintuitive, and unfamiliar distinctions in mathematics or science, so it's puzzling that researchers would assume that they can, or simply ignore whether people could readily do so for highly abstract philosophical distinctions.

### 2.4.3 Conflating realism with universalism

Moral realism is often confused with *moral universalism*. Moral universalism is the view that a given moral principle or standard applies to all moral agents, regardless of their location in time or space.[34] For instance, if it is a universal moral fact that it is wrong to own slaves, then it is not only wrong to own slaves in the United States, it is also wrong to own them in any nation on earth, or anywhere else in the universe, so it would also be wrong even for aliens to enslave one another.[35] Moral universalism is sometimes contrasted with moral relativism, in that the former holds that moral standards apply to all people, while the relativist may hold that moral standards can vary depending on an individual's subjective values, or the standards of their culture or group. For instance, a relativist may claim that it

---

[34] A *moral agent* is any entity that is appropriately subject to moral appraisal. A typical adult human is a moral agent, while babies, nonhuman animals, and inanimate objects are not. This restriction is intended to limit the scope of universality to appropriate targets. A position may still count as *universal* even if it fails to hold lightning morally accountable for striking people.

[35] Assuming they possess the relevant characteristics to be appropriate subjects of moral consideration, e.g., relevant forms of agency such that they can be held to the same moral standards as humans. For the record, moral philosophers do not reference aliens as often as they should. For instance, most forms of group relativism seem to implicitly refer only to differences between human cultures. Yet in principle one could advocate species-relativism, and defend the view that moral standards can be correct or incorrect relative to the standards of an entire species, rather than to particular cultures within that species. This position is rarely explored, presumably because there are no known alien civilizations to compare ourselves to, but if there were, this might very well be a popular position. It is interesting to note, then, that the conceptual space of metaethical positions that people happen to defend seems to some degree circumscribed by contingent features of our circumstances. If multiple advanced species had evolved on earth (e.g., advanced elephantine or cephalopod civilizations) and existed today, species-relativism might be a common position.

is morally wrong for Catholics to have abortions, but it is not morally wrong for non-Catholics to have abortions.[36]

Universalism represents one end of a continuum that represents the *scope* of moral facts. At one extreme, moral facts apply only on an individual basis: the individual subjectivist may believe that each of us ought to do that which is consistent with our personal moral standards. On such a view, the scope of a given person's moral standard, e.g., "do not murder," applies only to themselves. Someone else may endorse the same moral rule, but to the extent that moral rule applies to them, it does so in virtue of it being *their* standard, not someone else's. At the other extreme, the moral standard "do not murder" applies to everyone, everywhere, in all times and places. Other positions may fall somewhere between these two extremes e.g., cultural relativists may believe a moral rule applies to all members of a particular culture. It is also possible to believe some moral rules apply universally but others do not.

Regardless of where one falls on the continuum between moral facts being more or less universal in scope, universalism/localism is orthogonal to the distinction between realism and antirealism. Some researchers have drawn explicit attention to this and have sought to carefully avoid conflating universalism with realism (e.g., Goodwin and Darley, 2008).[37] Unfortunately, researchers have on occasion conflated realism and universalism (Yilmaz & Bahçekapili, 2015a; 2018; 2020). Even

---

[36] I provide this example rather than an example based on different cultures as a revolt against the common tendency to speak of relativism only in terms of cultural standards, rather than other potential group-based standards. Relativists need not hang their hat on one, and only one way in which one's standardized can be relativized.

[37] Goodwin and Darley are very clear not to mix the two up:

"[T]he question of whether ethical standards should apply to all cultures is a question about the *scope* of ethical standards, and is independent of the question of whether such standards and beliefs are objectively or subjectively true. Our interest centers on this second question, which concerns the *source* of such beliefs or standards - whether they derive their truth (or warrant) independent of human minds (i.e., objectively) or whether instead, their truth is entirely mind-dependent or subjective." (p. 1341)

so, participants often conflate universalism with relativism even when researchers don't formally conflate them (see **Chapter 4**).

### *2.4.4 Conflating realism with absolutism*

People also conflate realism with *absolutism*. Absolutism is the view that certain moral rules or principles are *exceptionless* (Hawley, 2008). In other words, there are certain types of actions that an agent may not perform under *any* circumstances. For example, an absolute prohibition against killing other people would entail that it is wrong to intentionally kill people, even in self-defense or to prevent that person from killing a greater number of people.

Absolutism is a property of first-order moral norms, is orthogonal to the dispute between realism and antirealism, and *it does not entail realism*. A person who endorses absolute moral rules could have a stance or commitment towards cultural relativism or individual subjectivism. If so, then their endorsement of an absolute moral rule would best be understood as e.g., "there are no exceptions to this rule *according to my culture's standard*s," or "there are no exceptions to this rule *according to my personal moral code*." Likewise, realism does not entail absolutism. A realist may believe that there are stance-independent moral facts, but believe that those facts represent a body of rules and principles that are flexible and result in moral principles and rules that permit exceptions in particular circumstances. For instance, a realist may believe that it is morally wrong to kill people, *except* in self-defense or to stop someone from causing extraordinary harm to others. As I show in **Chapter 4**, the conflation between realism and absolutism is one of the most common causes of unintended interpretations.

### 2.4.5 Conflating relativism with contextualism

Another common source of unintended interpretations is the conflation between contextualism and relativism (Chappell, 2007).[38] Contextualism is the claim that whether a given action or type of action is permissible under some circumstances but not others. For instance, one might hold that it is morally permissible to kill someone in self-defense, but it is not morally permissible to kill someone for personal gain. Or it may be morally permissible to lie to an enraged murderer about the whereabouts of their intended victim, but morally wrong to lie to avoid responsibility for your actions.

Contextualism is *not* relativism. Contextualism is a feature of *normative* moral principles and standards. It concerns the sensitivity of a given category of action, e.g., "killing," "lying," "stealing," to situational factors that influence the moral status of the action. Both realists and antirealists can favor moral principles that are sensitive to contextual considerations. Consider a contextualist account of the conditions in which killing is justified:

*It is morally wrong to intentionally kill others except in self-defense, justified wars, or to prevent great harm.*
A moral realist could believe this statement is stance-independently true. But a moral subjectivist could also think it is true, just not stance-independently true. In other words, *contextualism is completely orthogonal to moral realism and antirealism.*

Nevertheless, people routinely conflate relativism with contextualism. This confusion likely stems from the fact that both relativism and contextualism allow the moral status of an action to vary. Relativism has to do with moral claims being true or false on the basis of which standards they are indexed to. As such, relativism allows *the exact same action* to be right relative to some moral frameworks, and wrong relative to others. Contextualism doesn't allow for this. Contextualism has to do with moral actions being right or wrong on the basis of specific details relevant to a given situation, and that can

---

[38] I use the term "contextualism," here, but there is no standard usage in the literature, and it is also used to refer to other concepts or distinctions which I do not address here (e.g., Brogaard, 2008; Evers, 2014).

vary from one situation to another. Thus, it allows for an *action type* to be right in some cases and wrong in others, but the rightness or wrongness always turns on *differences* between these circumstances, e.g., lying could be morally wrong in situations where it is used for selfish gain, but morally right in situations where it would save lives.

Both relativism and contextualism allow the truth of some general moral action, like "stealing," to be morally good sometimes and morally bad other times, but for completely different reasons. Thus, in a certain sense, relativism *does* hold that a moral claim is true or false "depending on the context," but it refers to a very distinct type of context: the context of utterance. *Who* makes the claim and *what* moral framework is indexed by their claim determines the truth conditions for the statement in question. Yet this sense of "context" has nothing to do with the contextual considerations relevant to contextualism. Contextualism treats a given moral action as morally right or wrong on the basis of distinctive features of a given moral situation *other than* standards of the people engaging in moral actions or judging the moral actions of others. This is a subtle distinction that is difficult to convey. It is no surprise that it is one of the most common conflations people express (see **Chapter 4**).

### 2.4.6 Conflating relativism with descriptive claims

Some people believe abortion is morally permissible. Others believe abortion is morally wrong. The members of some societies are more likely on average to consider a particular action wrong than members of another society. Such observations involve a *descriptive* claim to the effect that different individuals and societies have different moral standards. And these claims are completely independent of whether the respective standards are *correct* or *incorrect*, and they are likewise independent of whatever metaethical position one might endorse. Yet, for reasons I cannot fathom, people seem to conflate the banal observation that different people have different moral standards with relativism. Not just participants, but researchers, too. As I discuss in Bush and Moss (2020), several of the items Yilmaz and Bahçekapili use conflate descriptive claims with metaethics, e.g., "*What is moral varies on the basis of*

*context and society*."[39] Yet even when researchers avoid such conflations, they are rampant among participants across a variety of studies. For instance, participants were asked what one of the items on Zijlstra's (2019) *Folk Moral Objectivism* (*FMO*) scale means:

*"What is morally right and wrong is relative to the moral beliefs of an individual, culture, or society"*

They routinely responded with statements like these:

*Different groups of people have different morals.*

*Different cultures and societies have different morals to beliefs as to what is right and wrong.*

*It is true that different individuals and cultures have different morality belief systems*

Responses like these are very common, and illustrate that even when people are presented with a claim that *explicitly* describes morality as "relative," they frequently interpret such claims as *descriptive*.

## 2.5 A summary of remaining problems

There are many more methodological shortcomings, but they cannot be adequately addressed in detail. I will quickly summarize the remainder, before concluding with the implications not simply of each methodological worry taken in isolation, but the *cumulative* effect of so many.

### 2.5.1 Evaluative standard ambiguity

If two people who disagree about a moral issue are both referring to the same moral standards (such as the standards of a shared culture), relativists would agree that at least one of them must be incorrect. Studies that don't specify which moral standards people's claims are referring to risk encouraging some antirealists to choose the "realist" response option. Attempts at clarifying would complicate studies and may be ineffective.

---

[39] Note that they also conflate relativism with contextualism in the very same item.

### 2.5.2 Abstract norm ambiguity

There are different ways to comply with the same abstract moral rule. For instance, the proper way to respect the dead is burial in some cultures, and cremation in others. Participants may judge that people from these cultures may disagree, and both be correct, not because moral truth is relative, but because they believe both cultures have identified appropriate ways to comply with the same moral rule. This may result in participants choosing the antirealist option even though this position has nothing to do with realism or antirealism (but rather with the application conditions of normative standards).

### 2.5.3 Misattributing source of disagreement

Participants may attribute the source of the disagreement to something other than a fundamental difference in moral values, e.g., to a disagreement over the nonmoral facts or to misunderstanding the scenario. If so, their responses cannot serve as valid measures of realism or antirealism because the validity of the disagreement paradigm requires participants to interpret the cause of the disagreement as a difference in moral values.

### 2.5.4 Domain classification inconsistency

Most studies present participants with items that researchers have classified themselves as "moral." Yet ordinary people do not agree with researchers or with one another about which issues are "moral" or not (Levine et al., 2021; Machery, 2018; Wright, Grandjean, & McWhite, 2013). Researchers who ignore this problem are unable to assess the extent to which this undermines the validity of their results or at least renders them much noisier. Researchers who account for this issue by asking participants to classify issues as "moral" or not (a) face yet another challenge, in that participants won't agree with researchers or one another about what it means for an issue to be "moral," (b) face validity issues with this task as well (e.g., interpretative difficulties, performance errors, etc.), and (c) if such measures

result in excluding responses participants don't classify as moral, cross-participant comparisons will be compromised, perhaps severely.

### 2.5.5 Presumption of correspondence theory of truth[40]

The disagreement paradigm presumes participants endorse a correspondence theory of truth. There is no evidence that they do so with respect to moral claims (Pölzler & Wright, 2020a). The only pair of studies to specifically address this question suggest ordinary people *don't* exhibit a clear and uniform commitment to a correspondence conception of truth (Barnard & Ulatowski, 2013; Reuter & Brun, 2021). Instead, both studies find that ordinary people frequently reject features of correspondence theories and exhibit a variety of different positions, and may have multiple or (in Reuter and Brun's case) ambiguous meanings in ordinary language. For instance, Reuter and Brun claim that their findings "reveal that the use of 'true' shows substantial variance within the empirical domain, indicating that 'true' is ambiguous between a correspondence and a coherence reading" (p. 1).

### 2.5.6 Signaling & reputational concerns

Given the centrality of our moral thoughts, values, and behavior, ordinary people may be motivated to select response options on the basis of what they perceive would signal socially desirable characteristics. This may be exacerbated by a tendency to ascribe normative or ideological implications endorsing metaethical positions, or at least the kind of phrasing used to describe metaethical positions.

---

[40] This is a problem that was initially proposed by Pölzler (2018a; 2018b) and discussed in Pölzler & Wright (2020a; 2020b). It had not occurred to me to consider the dependence of folk metaethics on other philosophical presumptions that are often taken for granted. It leads me to wonder whether there are others that may threaten the validity of folk metaethics. Yet it also hints at a far greater worry: that many, perhaps *all* folk philosophical thought is so deeply nested in the rest of each individual's philosophical beliefs and commitments that the only way to assess any particular person's philosophical views would be holistically, taking all of their views (or at least the relevantly connected ones) into account. That is, any particular position a person takes on any given issue may only make sense by taking into consideration the rest of their beliefs and commitments, in which case studying individual philosophical views in isolation from the rest of each person's views may make little sense.

This could motivate some participants to select response options for reasons unrelated to a sincere expression of a metaethical position.

### 2.5.7 Lack of realism

Lack of *realism* can also threaten the external validity of the disagreement paradigm (Calder, Phillips, & Tybout, 1982; Findley, Kikuta, & Denly, 2021). There are at least three ways that such stimuli can lack external validity and thus exhibit poor generalizability:

> **(i) Lack of experimental realism:** Participants may be less attentive or motivated to engage stimuli they find unserious or unrealistic (e.g., people may take a scenario involving space pirates less seriously than one involving ISIS)

> **(ii) Lack of mundane realism:** *Unfamiliar* scenarios with *impoverished stimuli* may prompt responses that don't reflect more prototypical, familiar, and contextually rich instances of the relevant phenomena (e.g., people's moral evaluation of scenarios involving space pirates may tell us very little about moral judgment and behavior in real situations)

> **(iii) Lack of psychological realism:** Stimuli that are amusing, strange, unrealistic, or difficult to imagine may prompt different psychological processes than stimuli that lack these features (e.g., if people find material funny, elevated mood may have an undesirable influence on how they respond)[41]

### 2.5.8 Lack of external validity

Almost all research has been conducted on WEIRD populations, with much of it conducted specifically among college students. What research has been conducted cross-culturally is both limited and limited both in terms of how many studies have been conducted and how representative the relevant samples are. This limits the generalizability of most folk metaethics research (Henrich, Heine, & Norenzayan, 2010). Furthermore, most studies represent the moral domain via a motley array of concrete moral issues that are inappropriately treated as representative of the moral domain as a whole,

---

[41] These distinctions are my interpretation and integration of the categories proposed by Pölzler(2018b) and Bauman et al. (2014, p. 537). See Bauman et al. (2014) for a deeper discussion of these problems as they relate to popular stimuli used in moral psychology. The threats proposed here are more of a problem for that research than they are for research on folk metaethics.

exposing all research relying on concrete moral items to the stimulus-as-fixed-effect fallacy, further limiting the generalizability of existing studies (Baguley, 2012; Judd, Westfall, & Kenny, 2012).

### 2.5.9 Forced choice obscures indeterminacy

Almost all studies employ forced choice paradigms that require participants to endorse some form of realism or antirealism. This obscures the possibility of metaethical indeterminacy.

### 2.5.10 Inaccurate, biased, or misleading stimuli

Stimuli, including response options and instructions, can be inaccurate or misleading, or bias participants towards particular response options. While this is a mundane problem common to all research, metaethical concepts are subtle and hard to describe, and researchers have routinely misoperationalized realism, antirealism, and subcategories of each, suggesting it is especially difficult to operationalize metaethical concepts.

### 2.5.11 Questionable a priori theorizing

Studies rely on concepts and distinctions derived from contemporary analytic metaethics. Many of the relevant distinctions emerged in the 20th century and have been primarily described and refined among a highly educated academic elite, almost exclusively in the Anglophone world. Their distinctive way of thinking about the nature of morality may fail to represent how ordinary people think. Yet studies rely exclusively on hypotheses derived from this discipline, *not* a richer foundation in psychological research.

## 2.6 Conclusion

The previous sections provide concrete and specific reasons why responses to the disagreement paradigm may fail to accurately reflect whether participants are realists or antirealists. Setting these issues aside, we may also take a bird's eye view towards the problem of empirically assessing folk

realism and antirealism. Philosophers have historically presumed that the way ordinary people speak and think is underwritten by an implicit commitment to one or another of competing metaethical accounts. It's not clear to me why they've been so convinced that there is a determinate, shared folk metaethics that the presumption that there is has gone virtually unchallenged in academic philosophy. Yet even if the brick and mortar of everyday claims like "Alex is an asshole" and "Sam shouldn't lie" best fits some realist or antirealist semantic theory, it's not obvious we should expect to readily determine what that theory is by asking people questions about metaethics. And it's even less clear that we should reasonably expect people to have explicit metaethical stances. If contemporary analytic metaethics has demonstrated anything, it is that metaethics is *complicated*, *subtle*, and *difficult*.

In our experience, David and I have found that students struggle to understand metaethical concepts even under optimal educational settings where they have ample opportunity to read about the relevant distinctions and receive instructions from specialists (Bush & Moss, 2020). David and I have also pointed out that philosophers routinely accuse one another of failing to accurately characterize particular metaethical distinctions, noting that if "trained professionals struggle to grasp these concepts, it is unclear why we should suppose that untrained laypersons should readily do so" (p. 7). In addition, we've outlined several criteria that, along with the complexity of metaethical concepts, cast serious doubt on the plausibility that ordinary people could be reasonably expected to interpret questions about metaethics as intended:

>    *(i) the relevant metaethical theories are complex and difficult to grasp*
>    *(ii) most people are unfamiliar with these distinctions prior to encountering them in studies*
>    *(iii) metaethical theories are generally abstract and distant from real world practical questions lay populations would be more familiar with and expect to be asked about*
>    *(iv) there are typically plausible non-metaethical interpretations of the questions posed to respondents* (p. 6)

Given these considerations, it may be *less* likely that people's responses reflect people's stances and commitments towards realism and antirealism than other unintended intended distinctions.

The many methodological problems highlighted in this chapter corroborate this claim. *Some* of the issues can be mitigated or avoided. Considered in isolation, it may seem the disagreement paradigm could survive any particular methodological limitation. Yet each problem would typically require extending the instructions, providing more response options, or otherwise expanding the length and complexity of the disagreement paradigm, which could result in a different set of methodological problems: the resulting "corrections" would result in a bloated monstrosity of a study, so laden with clarifications and instructions and complicated response options that even professional philosophers would balk, while ordinary people would be left with the task of taking what amounted to a brief introductory course in philosophy.[42] I discuss efforts to devise *training paradigms* that do just this in the next chapter.

---

[42] Indeed, this is precisely what has happened with efforts to design *training paradigms*: Pölzler (2021) reports that it took on average *fifty minutes* to complete the studies appearing in Pölzler and Wright's (2020a; 2020b), and I don't think their efforts were adequate (see **Chapter 3** and **Supplement 3**). In my own experience, it would take far more work than a few minutes for participants to develop adequate understanding of what's at stake in the dispute between realists and antirealists.

## CHAPTER 3:
## Critique of Alternative Paradigms

**3.0 Introduction**

In addition to the disagreement paradigm, researchers have employed several other methods to study folk metaethics. The primary alternative is to devise *scales* in which participants are asked to express their level of agreement or disagreement with a set of sentences. These sentences are intended to express various metaethical positions (e.g., realism, relativism, and so on) with an aggregate score in any given subscale reflecting one's overall endorsement of that particular metaethical position. *Training paradigms* are another prominent alternative. This approach involves instructing participants in the relevant metaethical distinctions, and may include exercises intended to reinforce and confirm that participants understand the relevant distinctions.

There are a variety of other paradigms as well, often appearing in only one or a handful of studies. I address all of these paradigms in **Supplement 3**. The extremely short version of that commentary is that, without exception, *all of these paradigms are invalid*. Here, I will only focus on scales and training paradigms. Each serves to highlight the distinct methodological shortcomings and alternatives to the disagreement paradigms. These can be summarized as follows: (1) scales consistently suffer poor face validity, and, even when they don't, participants still struggle to interpret items that appear on these scales as intended. Training paradigms seek to minimize unintended interpretations, however (2) it's not clear they succeed at this task, and if they did, this would simply trade the problem of ensuring stimuli are interpreted with another, equally serious problem: once participants have been trained in the relevant metaethical distinctions, we can no determine whether their responses reflect metaethical stances or commitments held *prior to* training, or whether the training *induced* participants to adopt determinate metaethical stances, a process I call *spontaneous theorizing*.

**3.1 Scales**

The most prominent alternatives to the disagreement paradigm are metaethics *scales*. Metaethics scales present participants with a series of short statements that are intended to represent one or more metaethical positions. Participants are then asked to express a stance towards each such statement. This typically involves the use of Likert scales, with participants expressing the degree to which they agree or disagree with each statement. For instance, Collier-Spruel et al. (2019) present participants with a series of statements that reflect belief in relativism. Here are a pair of examples:

(1) *The viewpoint of one's culture determines whether their actions are morally right*
(2) *An action is only morally wrong if a person believes it is morally wrong*

Participants are asked to express how strongly they agree or disagree with each of these statements on a scale of 1 to 5 (1 = strongly disagree, 5 = strongly agree). Responses to these items can be totaled to provide an aggregate score of degree of belief in relativism, with higher scores representing greater belief in relativism, and lower scores reflecting disbelief in relativism. Researchers can also assess average level of agreement for clusters of items that represent a particular metaethical position, which could provide insight into whether a particular population leans more towards or against relativism (or whatever metaethical positions are included in the scale).[43]

At present, there are six genuine metaethics scales, and two "quasi-scales" (clusters of *ad hoc* items used in place of a full scale): the relativism subscale of the Ethics Position Questionnaire (*EPQ*, Forsyth, 1980), the Objectivism-Subjectivism scale (*TOS*, Trainer, 1983), a set of 3 items used by Yilmaz & Bahçekapili (2015b), the New Meta-Ethics Questionnaire (*NMQ*, Yilmaz & Bahçekapili, 2015a), a three-item moral objectivism scale (MO3, Sarkissian & Phelan, 2019), the Moral Relativism

---

[43] Not all scales assess use Likert scales. Trainer (1983) presented participants with a series of opposing pairs of metaethical statements that reflected realism and antirealism (Trainer used the terms "objectivism" and "subjectivism") and asked participants to choose the one they favor.

Scale (*MRS*, Collier-Spruel et al., 2019), and the Folk Moral Objectivism scale (*FMO*, Zijlstra, 2019).[44]

Details of each are summarized in **Table 3.1** (complete scales are available in **Appendix C**.). To keep the tedium of an item-by-item analysis to a minimum, I provide a critique of Sarkissian and Phelan's (2019) 3-item quasi-scale below. However, substantive problems often emerge only by analyzing specific items. These critiques appear in **Supplement 3**, where I also provide a general overview of the advantages and disadvantages of scales.

### 3.2 Problems of validity in metaethics scales: An example

For the sake of brevity, I will focus on one of the shortest scales, Sarkissian and Phelan's (2019) 3-item moral objectivism scale (MO3). This allows me to present a focused discussion of the distinctive issues with each of these items rather than spreading myself thin by addressing all of the items on a larger scale. There are a few other reasons for selecting this particular scale: these items suffer many of the same problems as items that are not discussed in the main text. For extended discussion of problems with other scales, see **Supplement 3**.[45] In addition, these items were *explicitly* designed in response to methodological concerns about the poor validity of earlier items, namely those that appear on Forsyth's (1980) *Ethics Position Questionnaire (EPQ)*. The MO3 appears to be an *ad hoc* set of measures devised in the midst of a project, without any explicit efforts made to validate the items beyond

---

[44] Notably, Forsyth's EPQ (1980) and Trainer's TOS (1983) predate the disagreement paradigm by several decades. Yet neither managed to spawn a literature specifically dedicated to the study of folk metaethics in a way that is well-integrated with the philosophical literature in the way contemporary research on the psychology of metaethics has managed. Like the fall of the Roman empire, any official start date to the present era of folk metaethics research will be largely a matter of convention, but if I had to choose a date, it would be with the publication of Goodwin and Darley's 2008 publication, "The psychology of meta-ethics: Exploring objectivism." While several articles preceded it, including the aforementioned Forsyth (1980) and Trainer (1983), as well as Nichols and Folds-Bennett (2003), Nichols (2004), and Wainryb et al. (2004), none of these seemed to attract the attention that Goodwin and Darley's article did. The latter seems to have attracted enough interest for the topic to reach critical mass, and establish itself a sustained area of research.

[45] This should be qualified by acknowledging that some scales went through far more rigorous validation efforts, and resulted in items that suffered fewer problems than those discussed here. Nevertheless, these items still suffer poor validity (See **Supplement 3**), and participants still exhibited low rates of clear intended interpretation (see **Chapter 4**). Thus, while other items may be better, they're still not valid. For comparison, a very accurate description of Gödel's incompleteness theorems may be less prone to particular confusions than a very bad one, but it's still plausible that most people would remain confused and incapable of understanding the theorems.

reporting Cronbach's alpha, which does not provide adequate (or even very good) evidence of validity (Sijtsma, 2009). In particular, in describing the virtues of the MO3, they state that it was "not based on previous measures," adding that they "[...] were concerned that existing measures, such as the 5-item scale adapted from Forsyth (1980), did not capture metaethical views in a precise way, potentially limiting what we can infer from their use in this context" (p. 4).

**Table 3.1**

*Summary of metaethics scales*

| Scale | Full name | Reference | Items | Construct |
|-------|-----------|-----------|-------|-----------|
| EPQ | Ethics Position Questionnaire | Forsyth (1980) | 10 | relativism |
| TOS | Trainer-Objectivism subjectivism scale | Trainer (1983) | 17 | objectivism, subjectivism |
| YB3 | 3-item face valid metaethics scale | Yilmaz & Bahçekapili (2015b) | 3 | objectivism, subjectivism |
| NMQ | New Meta-ethics Questionnaire | Yilmaz & Bahçekapili (2015a) | 8 | objectivism, subjectivism |
| MO3 | 3-Item Moral Objectivism Scale | Sarkissian & Phelan (2019) | 3 | objectivism, relativism |
| MRS | Moral Relativism Scale | Collier-Spruel et al. (2019) | 10 | relativism |
| FMO | Folk Moral Objectivism scale | Zijlstra (2019) | 20 | no truth, relativism, universalism, absolutism, divine command theory |
| JRT5 | Objectivity of Morality scale | Johnson, Rodrigues, & Tuckett (2020) | 5 | objectivism, subjectivism, normativity |

Unfortunately, the MO3 does not capture metaethical views in a precise way, either. Although their scale purports to measure "objectivism," they do not offer an explicit description of what they take objectivism to mean. However, it is reasonable to infer that they have in mind roughly the same

conception as the researchers they cite, such as Goodwin and Darley (2008). Thus, the construct they have in mind seems to correspond to moral realism in that objectivism refers to the claim that there are stance-independent moral facts. I provide a brief summary of the problems with each of the three items here in **Table 3.2**.

**Table 3.2**

*Summary of methodological problems with the MO3*

| **Methodological problems by item** |
|---|
| *There exists a single moral code that is applicable to everyone, regardless of any individual person's beliefs or cultural identity.*<br><br>(1) Conflates universalism with stance-independence<br><br>(2) Exhibits abstract norm ambiguity (conflating multiple ways of conforming to the same abstract moral rule with relativism)<br><br>(3) "Applicable" may be interpreted in practical or normative terms<br><br>(4) Use of the term moral "code" encourages unintended interpretations<br><br>(5) Could be interpreted as a descriptive claim<br><br>(6) Could prompt conflation with absolutism vs. contextualism |
| *If two people really disagree about a particular moral problem then at most one of them can be correct, since moral problems cannot have multiple correct answers.*<br><br>(1) Exhibits problems associated with disagreement paradigm<br><br>(2) Exhibits abstract norm ambiguity (conflating multiple ways of conforming to the same abstract moral rule with relativism). The wording of this item is especially vulnerable to this problem<br><br>(3) Prohibits expressing metaethical pluralism by presuming intrapersonally uniform metaethical stance<br><br>(4) Unclear what "really" means. Does not clearly convey fundamental moral disagreements, and is thus subject to attributing source of disagreement to nonmoral differences<br><br>(5) Use of the term "problem" rather than "claim" may amplify abstract norm ambiguity<br><br>(6) Could prompt conflation with absolutism vs. contextualism<br><br>(7) Worded as a negation with use of "cannot," which may increase cognitive load |
| *It is possible to compare different cultures by a single, universal standard of moral rightness.* |

(1) Conflates universalism with stance-independence

(2) Completely lacks face validity: whether it is possible to compare different cultures by a single standard does not entail that that standard is true, much less stance-independently true.

(3) Unclear: it's not clear what it means to say that it's "possible" to "compare" different cultures by a "single, universal standard."

(4) May prompt conflation between metaethics and normative ethics

(5) May conflate realism intersubjectivity or constructivism

(6) May conflate realism with absolutism, close-mindedness, rigidity, and other undesirable traits

---

### 3.2.1 MO3 item #1

| **MO3 #1** *realism* | There exists a single moral code that is applicable to everyone, regardless of any individual person's beliefs or cultural identity. |
| --- | --- |

This item conflates realism with universalism. Universalism holds that there is a single standard of moral truth, but this concerns the *scope* of moral claims (i.e., who they apply to), not what makes them true. Unfortunately, *who* a moral standard applies to has nothing to do with whether that standard is stance-independent or not. A moral antirealist could endorse a normative moral rule, e.g., "it is morally wrong to steal," and endorse the claim that this rule applies to *everyone*. For instance, ideal observer theory holds that there is a single correct moral standard, but this standard depends on what an ideally rational and fully informed agent would endorse (Firth, 1952). Thus, there is one correct moral standard, but that standard depends on the stance of a hypothetical agent. Conversely, a moral realist could believe there are stance-independent moral facts, but at least some of those facts aren't universal. They might think that e.g., different (stance-independent) moral standards apply to different people. More generally, people could believe that there can be multiple correct standards, but that does not necessarily make the stance-dependent, since it is possible to relativize moral claims to stance-independent criteria (Joyce, 2015). In short, the distinction between whether there is one or multiple correct moral standards is orthogonal to the distinction between stance-independence and stance-

independence. Asking people whether there are *universal* moral rules cannot serve as a legitimate measure of realism.[46]

Nevertheless, one might reason that universalism is an excellent proxy for realism. We may suspect this is the case by starting with the observation that we tend to be realists about other domains, such as science and math. Then, we might suppose, since we tend to think there is one stance-independent standard of truth in these domains, we may imagine that the believing there is only a single body of facts with respect to that domain is a good indication that the domain is not merely universally true but true in a stance-independent way. In short, if universalism tends to coincide with realism in other domains, it may plausibly do so with respect to realism as well. This is not an unreasonable assumption. Yet it requires us to make a theoretical assumption about the relation between universalism and realism that may be unwarranted, and at best leaves us with an indirect measure of realism. It is also possible people interpret the claim as descriptive. That is, it could be that people believe the item reflects the claim that, as a matter of fact, there are standards all societies accept, e.g., descriptive universals, such as rules against rape or murder.

Another issue is that it is unclear what "applicable" means. While philosophers may recognize that this is intended to convey that everyone's actions are right or wrong in accordance with the same standards, and that praise, blame, and whatever else may be associated with morality is justified and in all of these cases, ordinary people may not know what it means, or may interpret "applicable" in a more forceful, literal, and (most importantly) *normative* way.

---

[46] This conflation is unfortunate given that Goodwin and Darley's (2008) earlier research on folk metaethics explicitly drew attention to the distinction and made it very clear that they sought to avoid the conflation:

"[…] the question of whether ethical standards should apply to all cultures is a question about the *scope* of ethical standards, and is independent of the question of whether such standards and beliefs are objectively or subjectively true […] our interest centers on this second question, which concerns the *source* of such beliefs or standards – whether they derive their truth (or warrant) independently of human minds (i.e., objectively), or whether, instead, their truth is entirely mind-dependent or subjective." (p. 1341)

This is especially unfortunate since Sarkissian and Phelan discuss this exact paper in their article (see p. 2).

Yet another concern with this item is the use of the term "code." Participants may interpret this as a statement about whether moral standards should be formalized, written down, or otherwise enshrined in our institutions or laws. If so, they will have interpreted "code" in a literal but unintended way. In addition, the notion of a moral "code" may evoke notions of an especially absolutist or dogmatic moral system, or of stark and inflexible rules that are insensitive to situational factors. Insofar as participants tend to hold a favorable or unfavorable attitude towards a rigid and absolutist moral system, this may prompt participants to express greater or lesser agreement with this item for reasons unrelated to metaethics. For instance, participants may interpret disagreement with this item to reflect particularism or sensitivity to context, even though such positions are consistent with agreement, since people could believe there are universal, but flexible and context-sensitive moral principles.

A related, but subtly distinct conflation could occur if participants interpret this statement to imply that there are universal means of complying with a particular moral principle or rule. This is the problem of *abstract norm ambiguity* discussed in chapter 2. For instance, there may be a general moral rule that we should "show respect to others." However, the precise way in which we ought to do this may vary across cultures e.g., the proper way to show respect may be to bow, kneel, or shake hands. Participants may believe this statement would reflect the view that there is only one proper way of conforming to an abstract moral rule, and reject it on those grounds. Yet again, this is not what agreement with such a statement would entail.

Moral rules may also apply to some people or groups in virtue of their particular status, position, or relation to others. For instance, parents have special duties and responsibilities to their children, doctors have special duties to their patients, leaders have responsibilities over their subordinates, and so on. In other words, people may have *positional duties*, or duties that they possess in virtue of their status, position, or role within a particular social context Babushkina (2019). Such positional duties may be *universal* in a conditional sense, i.e., "if you are the parent of a child, then you

have an obligation to care for that child," or "if you are a police officer, you have an obligation to apprehend someone committing a crime," and so on. Such rules may be universal, but not necessarily apply to everyone, simply because not everyone occupies the same status, position, or role within a society. Yet consider the wording of the item: "There exists a single moral code that is applicable to everyone [...]" This *could* be interpreted as a view that we have only nonpositional duties, in that the same set of rules apply to all people everywhere, regardless of their particular role or relation to others.

This possibility may not be aided, and may even be exacerbated, by the latter half of the item, "[...] regardless of any individual person's beliefs or cultural identity." *Philosophers* may understand this to mean that a person's moral beliefs, or the moral standards of their culture, are irrelevant with respect to the moral facts, because those facts are stance-independent, and thus not made true by the standards of individuals or groups. Yet is this how nonphilosophers without training would interpret this phrase? This is an especially relevant question given that laypeople are unfamiliar with stance-independence and such considerations would probably not be salient without additional instructions or context in the questionnaire that the item is included in. In other words, this phrasing may *seem* to be a straightforward remark about how the moral facts in question are not stance-dependent, but this may only be obvious if one has the requisite training to recognize the purpose of this phrasing.

To make matters worse, this isn't even the standard phrasing philosophers themselves would typically use. What does it mean to say that the same moral code applies to everyone regardless of that person's cultural *identity*? Philosophers may wish to specify the *standards* of that person's culture, rather than their identity. More generally, it is simply unclear how individual beliefs or cultural identity are intended to relate to the applicability of the putative moral rules; it is left to the participant to understand that the intended relation is to serve as truth-makers for moral facts. This is a specific and sophisticated relation that would be difficult to convey, and it's asking a lot to expect nonphilosophers to reliably interpret it this way.

### 3.2.2 MO3 item #2

| | |
|---|---|
| **MO3 #2** **_realism_** | If two people really disagree about a particular moral problem then at most one of them can be correct, since moral problems cannot have multiple correct answers. |

This item is an abstract, scale-version of the disagreement paradigm. It is abstract in that it refers to the moral domain as a whole, rather than to concrete moral disagreements. As such, its most serious shortcoming is that it inherits many of the same methodological problems as the disagreement paradigm itself.[47]

Another problem with this item is that it is extremely vulnerable to *abstract norm ambiguity*, since a realist could believe that a "moral problem" could have multiple correct answers. What distinguishes a realist from an antirealist is not whether they believe moral problems have one solution rather than multiple solutions, but whether they believe the solution(s) to a problem are made true in virtue of stance-independent moral facts. Moral realists are thus *not* obligated to believe that "moral problems" have only one correct solution. Instead, they may believe that a *range* of actions could be permissible given the same abstract moral standard. By failing to make explicit that the disagreement is about the truth status of a distinct moral claim, this item conflates fundamental moral disagreement with disagreement over the number of ways one could conform to a given moral standard. The latter possibility is by no means obscure or esoteric. Consider the following actions:

(1) Building a bridge
(2) Defeating an opponent in chess
(3) Writing a computer program that solves a specific problem

---

[47] One small but clever potential advantage with this item is the inclusion of the phrase "really." This could help to emphasize that it is referring to *fundamental* moral disagreement, rather than a disagreement over nonmoral facts or in some other way not reflecting moral stances that genuinely conflict with one another in terms of their substantive moral content. However, it is not at all obvious that untrained participants would pick up on the use of "really" and interpret in the intended way.

Even with well-specified goals that allow for one to unambiguously categorize attempted solutions as correct or incorrect, all of these examples allow for *multiple* solutions to be correct. Nevertheless, there are stance-independent facts about the correct and incorrect ways to achieve each of these goals. It simply isn't the case that if one believes that the facts in a given domain are stance-independent that one believes that, for any problem within that domain, there is only one correct solution. To suggest this is to favor a form of extreme rigidity that many people would reject, regardless of whether they were realists or not. Such apparent rigidity may have given participants the mistaken impression that this item was asking about *absolutism*, the belief that there are exceptionless moral rules. Such views may be unappealing both because they are inconsistent with people's normative moral standards and because they signal undesirable character traits (e.g. close-mindedness).

Finally, this item seems almost ideally worded to encourage abstract norm ambiguity. Rather than referring to people disagreeing about a moral claim, or issue, it describes people disagreeing about a moral "problem." A "problem" is *precisely the kind of term people would be inclined to use when describing something that could have more than one possible solution.*

Thus, not only does this item potentially conflate metaethics with the idea that there could be more than one way to comply with a moral rule, it seems to actively encourage the unintended interpretation.

### 3.2.3 MO3 item #3

| MO3 #3 *realism* | It is possible to compare different cultures by a single, universal standard of moral rightness. |
|---|---|

Like item #1, this item conflates realism with universalism. Yet this item is much less appropriate as a measure of realism for the simple reason that it is *unclear* and, if literally interpreted, *is not about metaethics at all*. First, it is unclear what it means to "compare" cultures by a single standard of rightness.

This does not necessarily entail anything about whether different cultures have different moral standards, nor does it indicate whether those standards are stance-independently true.

In principle, one could simply choose to compare culture according to a single, uniform standard without that standard being "true." Food or movie critics could develop criteria that allows them to compare different food or movies "by a single, universal standard." Yet simply because you *can* compare different things according to a particular standard it does not make that standard *true*, much less *stance-independently true*. Even if everyone agreed to follow that standard, this would not entail stance-independence. If everyone agreed on what movies or pizza toppings were best, this also would not make it stance-independently true that they were the best, it could just mean that everyone has the same stance.

For this item to be valid, participants would have to interpret *comparison* in such a way that it refers to holding all cultures to the same standard in a vantage-point neutral way. That is, it is possible to compare all societies by a single universal standard *because there is in fact only one universal standard*. However, it could be interpreted to reflect the claim that a person or group could compare different cultures by *their* standards, which could be universal in scope. In other words, we could imagine that Alex endorses a universal standard of moral rightness, that is, a standard which applies to everyone. And Alex could then compare different cultures according to this standard to evaluate the degree to which they comply with it. Yet simply because Alex can compare different cultures according to this standard, it does not follow that this standard is stance-independently correct.

Second, whether it is possible to compare different cultures according to some standard is conceptually distinct from whether it is in fact the case that there is a stance-independently *correct* standard. After all, even an antirealist could agree that it is *possible* to compare cultures in this way, either because it is possible realism is true, or it is possible to compare cultures in this way whether or not realism is true. This item is open to a variety of such interpretations because it is ambiguous in

what sense it is "possible" to do what this statement asks. The intended interpretation is that it is possible to compare different cultures by a single universal standard because as a matter of fact *there is* such a standard.

Finally, references to a "single" standard could prompt associations with inflexibility, rigidity, and absolutism. That is, participants may associate or conflate universalism with the adherence to a moral standard that is insensitive or blind to local cultural context. This could also be reflected in a dogmatic adherence to broad and abstract moral principles, or with a generalist approach to normative morality, rather than a more particularist approach. In short, there may be a persistent threat of *normative entanglement*: the conflation between metaethical and normative considerations. In practice, it may be difficult to conceptually pull metaethical and normative apart in a way where one can present a sentence that has unambiguous metaethical content without also prompting associations with particular normative moral stances. And if those moral stances imply undesirable traits or attitudes, people may eschew response options for unintended reasons.[48]

### 3.2.4 Concluding remarks on the MO3

All three items on the MO3 lack face validity and suffer numerous shortcomings that raise serious questions about whether they could serve as valid measures of moral realism. Items #1 and #3, in particular, are not face valid at all. Item #2 suffers numerous shortcomings in addition to simply being another incarnation of the methodologically inadequate disagreement paradigm. Given these concerns, it's unlikely responses to these items provide reliable information about folk realism and antirealism. While some other scales fare better, they all suffer a variety of similar (and in some cases, distinct) shortcomings. They conflate unrelated distinctions with realism and antirealism, are prone to

---

[48] I address this possibility in greater detail in **Supplement 3**, where I discuss the role that signaling and reputation may be at play whenever items intended to convey realism are interpreted as expressions of absolutism, or are thought to convey this in addition to whatever else they may express.

be interpreted in a variety of unintended ways, present participants with a forced choice between items that don't reflect legitimate dichotomies, and so on.

### 3.3 Training paradigms

The primary methodological hurdle for both the disagreement paradigm and alternative paradigms is the difficulty of ensuring participants interpret questions about realism and antirealism as intended, and are given appropriate response options to express their positions. Unfortunately, conventional methods are not up to the task. Presenting people with standard survey items is insufficient, because conventional survey methods cannot adequately specify what they are asking in a way that ordinary people can reliably interpret as intended. This is because they do not provide adequate instruction and disambiguation to successfully minimize causes of unintended interpretations.

Adequately ensuring people interpret questions as intended would require extensive training and instructions to ensure participants could competently circumvent the numerous alternative interpretations that result from ambiguities, an inclination to interpret stimuli in more familiar ways (e.g., as questions about normative ethics or epistemology), and other potential factors, such as reputational concerns (e.g., endorsing realism may give the impression of being close-minded and intolerant).

One solution to this problem is to simply provide the requisite training and instruction. This is precisely what Pölzler and Wright (2020a; 2020b) have done in a series of innovative new paradigms, which I refer to as *training paradigms*. Pölzler and Wright have been at the forefront of identifying, documenting, and developing workarounds for the methodological shortcomings present in previous research on folk metaethics, and this shows in the quality and thoughtfulness that has gone into their studies. They have streamlined stimuli, improved instructions, minimized ambiguities, employed comprehension checks of the relevant metaethical distinctions, and provided converging evidence by comparing results across a diverse array of distinct paradigms. They have also

demonstrated a sophisticated understanding of the metaethical positions available to participants, and have included a wider variety of response options that reflect the array of metaethical positions previous studies failed to include. Yet the primary feature of this recent wave of studies has been the introduction of stimuli to minimize unintended interpretations by teaching participants about metaethics and inducing them to reflect on the nature of morality.

Since Pölzler & Wright's later studies correct for methodological shortcomings with Wright's (2018) earlier research, I will focus on Pölzler & Wright's training paradigms (I address Wright's 2018 article in **Supplement 3**.) Pölzler & Wright also explicitly intend their training paradigms to correct for the interpretive difficulties of earlier folk metaethics research. After describing the shortcomings of previous research, they state that "Our design starts with an explanation that purports to prevent participants from misreading its main tasks as asking for first-order moral intuitions [...] or other matters that are unrelated to the issue of moral objectivity [...]" (p. 59).[49]

Their explanation consists of a description of the distinction between metaethics and normative ethics, and makes it clear that they are going to ask people about metaethics. Participants are then asked to "bracket" their normative moral standards and to "ignore these intuitions or put them to the side" (p. 59). This is not a reasonable demand to make, and it's not clear people are psychologically capable of complying with it. Participants are effectively asked to suppress their moral attitudes and values, e.g., their repugnance at murder, torture, genocide, and cruelty. I can't do this, and I doubt that most ordinary people can do so either. This brings to mind instances in which jury members are instructed to disregard information that comes up in a courtroom, even if it's incriminating or reflects negatively on the character of the defendant. If the prosecution slips and mentions inadmissible evidence that the defendant's DNA was found on the murder weapon, and the

---

[49] Notably, they explicitly acknowledge the possibility that participants "do not have any determinate intuitions about the existence of objective moral truths at all" (p. 58). This is a rare acknowledgment of the possibility of indeterminacy.

judge asks the jury to "disregard" this information, do we seriously expect the jury to able to ignore this? I doubt it. And for the same reason, I doubt people can "bracket" their normative moral standards.

This is followed by presenting participants with "the least controversial and least biasing examples of normative and metaethical sentences we could think of" (p. 59). Finally, they "test and improve participants' understanding of the normative/metaethical distinction" using a pair of comprehension checks:

(1) *Distinction comprehension*: Participants must convey their understanding of the instructions by correctly choosing the response option, "Normative sentences express moral judgments and meta-ethical sentences make claims about the nature of morality itself" (p. 59)

(2) *Exercise*: Participants must successfully classify a set of sentences as either normative or metaethical statements

These training exercises preceded five distinct paradigms. Each of these paradigms includes additional details and instructions distinct to that particular paradigm, which I will only discuss when relevant.

Finally, participants in all of the paradigms that they use receive detailed response options that attempt to clearly describe the relevant philosophical position. These response options also present participants with a richer variety of answers than previous studies. For instance, Pölzler and Wright (2020b) distinguish between "secular realism" and "theistic realism."[50] Taking all of these modifications into consideration, Pölzler and Wright's training paradigms (2018a; 2018b) represent the most comprehensive instructions to date. This includes:

(1) *Disambiguating instructions*: Explaining what metaethics is and distinguishing it from other topics participants may conflate with metaethics (e.g., normative ethics)

---

[50] They describe secular realism as follows:

"When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they are independent from what anybody thinks about them. For example, an action that is morally wrong is wrong no matter what anyone thinks. So it would still be wrong even if you yourself, or the majority of the members of your culture, thought that it is not morally wrong." (Pölzler and Wright, 2020b, p. 60)

(2) *Target instructions*: Explicitly telling people the study is about metaethics

(3) *Response option instructions*: Explicitly describing different metaethical positions participants can endorse (e.g., realism, relativism, noncognitivism, and so on)

(4) *Disambiguating example*s: Providing examples that illustrate the distinction between metaethics and non-metaethical topics. Other studies (e.g., Wright, 2018) may also include examples that illustrate metaethical distinctions (e.g., between propositional and nonpropositional sentences)

(5) *Comprehension checks*: questions designed to ensure participants understand the relevant concepts and distinctions presented in (1)-(4). Participants who fail comprehension checks may be excluded from analysis.

(6) *Training exercises*: Participants may engage in various training exercises to develop competence with the relevant concepts and distinctions. Participants who fail training exercises may be excluded from analysis.

(7) *Detailed response options*: Participants are presented with detailed response options.

(8) *Additional comprehension checks*: Participants may be asked additional follow-up questions to assess what particular questions were about

(9) *Open response questions*: Participants were asked to explain their answers. Participants whose responses failed to demonstrate adequate relevance may be excluded from analysis.[51]

Given this many criteria, Pölzler and Wright pulled out all the stops, making a concerted effort to ensure participants interpret questions as intended. Training paradigms do not have to exhibit all seven of these characteristics, or any of these traits in particular. So long as the researchers conducting a study seek to cultivate competence in metaethics, a study qualifies as a training paradigm. Studies may involve minimal training (e.g., simply providing detailed instructions, but nothing else), or go through much more extensive efforts (e.g., requiring participants to pass a college course on metaethics). There is no bright line that categorically distinguishes training paradigms from everything else. Like many concepts, it's a matter of degree.

---

[51] Open response questions or other comprehension checks that appear after the main measures of a study are not, strictly speaking, a necessary feature of a training paradigm. However, since they can be used to evaluate the effectiveness of training, I have included them. On their own, efforts to evaluate interpretation rates or assess whether participants understand metaethical concepts does not induce a *change* in their competence with those concepts. The essential characteristic of any training paradigm is that it employs some method for improving or modifying the way participants understand metaethical concepts, typically for the purpose of enhancing the probability that they will interpret questions about metaethics as intended.

Finally, note that training paradigms are consistent with a variety of measures that follow the training portions of the paradigm (and that precede comprehension checks or other measures that follow the main measures). In this case, Pölzler and Wright employed modified versions of the disagreement paradigm, along with several novel paradigms, including a "theory task," a "metaphor task," and a "truth-aptness" task (I discuss these in **Supplement 3**).

The question is: were Pölzler and Wright's efforts successful? As you've probably already guessed, I believe the answer is a decisive, if regrettable, *no.* There are three main reasons why training paradigms fall short. The first two are serious, but survivable. However, the third reason is fatal to all training paradigms:

(1) *Inadequacy*: There are good reasons to believe that existing training paradigms fail to ensure adequate reflection, competence, and emphasis on semantic considerations
(2) *Irrelevance*: Competence in the relevant distinctions may have little or no significant causal influence on measures of the participant's metaethical stances/commitments
(3) *Spontaneous theorizing & philosophical induction*: Successful training yields a sample whose responses may have been formed by the training, rather than reflecting their metaethical stances/commitments prior to participant in the study.

I discuss problems (1) and (2) in **Supplement 3**. Although both present such serious challenges to existing paradigms that each independently undermine the validity of existing studies, space constraints limit me to discussing (3), *spontaneous theorizing* and *philosophical induction*, because they present *insurmountable* obstacles to the validity of training paradigms.

### 3.3.1 Spontaneous theorizing

Suppose for the sake of argument that training paradigms successfully enable participants to interpret questions about metaethics as intended.[52] The goal of research on folk metaethics is to discover the

---

[52] If the need for participants to *reflect* on their reactions is also considered a necessary condition, then we can assume this task succeeds as well. This *is* a requirement that Pölzler and Wright agree to for their purposes (see Pölzler & Wright, 2020a) since they accept Kauppinen's account of the necessary conditions for philosophically relevant measures (see Kauppinen, 2007). It may not be necessary for some research goals, however. For instance, Pölzler (2018a) points out that the failure of Goodwin and Darley's (2008) studies to meet Kauppinen's criteria "should not be understood as a criticism of Goodwin and Darley. After all, they did not aim at providing semantically relevant evidence in the first place. Their

metaethical stances/commitments of *ordinary people*, i.e., people *without* philosophically significant training. How could we possibly achieve this goal by recruiting ordinary people and then training them in philosophy? Granted, we might suppose that they haven't received *significant* training. But there is an important difference between those who receive (a) no training, (b) some training, and (c) a lot of training.

Suppose I am correct about folk metaethical indeterminacy: that in the absence of formal philosophical training, people have no determinate metaethical stances or commitments. And suppose that, like the various interpretations of quantum mechanics, the way that ordinary people speak and think about morality does not contain any implicit commitment to realism or antirealism. If so, any attempt to measure how ordinary people think that *changes* their psychology by *causing* them to have determinate stances or commitments would not be a measure of how ordinary people speak and think outside the experimental context; it would be a measure of how they speak and think once they are placed in an experimental context that, by design, *induces* them to have determinate stances and commitments. And in the absence of evidence to the contrary, we cannot know whether the measures obtained via training paradigms represent what the participant thought prior to participating in the study, or whether their pattern of responses were caused by the study itself. Thus, we could not justifiably use the results of such studies as evidence of how "ordinary people" think about metaethics unless we could be sure that the stimuli employed in our studies wasn't causing them to develop or at least superficially appear to express determinate metaethical stances and commitments.

Unfortunately, there is no feasible way to do this. Here's why: on the one hand, we could employ conventional, non-training measures, then assess how participants interpreted questions after the fact to test whether they interpreted questions as intended. However, this is exactly what I've done

interest was rather with illuminating the prevalence and causes of ordinary people's intuitions about the moral objectivism/subjectivism distinction" (p. 4925).

(see **Chapter 4** and **Chapter 5**). The results of these studies reveal that people consistently fail to interpret questions as intended due to ambiguities, conflations, and other factors that render ordinary people incapable of reliably interpreting conventional questions about metaethics (i.e., questions that lack disambiguating instructions, training, and so on) as intended.[53,54]

On the other hand, we could provide precisely those clarifications and disambiguations that form one of the foundations of training paradigms. Yet the very act of providing such clarifications is precisely what could *cause* participants, who are supposed to be "ordinary people," to think about the issues at hand in ways that they hadn't prior to participating in the study by causing them to reflect on metaethical concepts, develop competence in those concepts and distinctions that appear in measures, and so on. If this occurs, we would be unable to distinguish responses that reflect how that person thought *prior* to participating in the study from responses that reflect how the participant thinks *as a result of participating in the study*. I refer to this phenomenon as *spontaneous theorizing*. Spontaneous theorizing occurs *whenever a participant who held no determinate stance or commitment prior to participating in a study is induced to develop or express a stance or commitment due to the experience of participating in the study itself*.

For example, as Gill (2009) observes, most ordinary people have probably not considered whether mathematical Platonism is true. That is, they have not given much thought to whether our mathematical discourse refers to abstract objects that exist independent of how we think, speak, or act (Linnebo, 2018). And it is doubtful that everyday mathematical thought and language commits people to a particular account. Yet if we gave people a survey that required them to express agreement with either Platonism or anti-Platonism, we'd *necessarily* get *some* pattern of responses that would

---

[53] Even if participants seem to have interpreted questions as intended, such responses could in principle be confabulatory or have failed to reflect the actual reasons why participants answered as they did. However, this would be true for a wide variety of research and represents a more general methodological concern. Of course, that doesn't mean it isn't a legitimate concern. In this case it's moot since people don't typically interpret standard metaethics paradigms as intended anyway.
[54] It's worth noting that even before these studies were conducted, Pölzler and Wright had already outlined enough methodological worries that they felt the need to devise training paradigms.

suggest *everyone* endorsed either Platonism or anti-Platonism. Even if participants held no determinate stance or commitment prior to the study, the very fact that they are now prompted to do so may cause them to reflect on and form such a view. That is, even if ordinary people have no determinate stance or commitment with respect to a particular philosophical distinction, the very act of asking participants may *cause them to develop a determinate stance*. Any study that *causes* participants to hold a particular view cannot be used as evidence that they already held that view, and it would be extremely difficult to empirically demonstrate whether the study itself caused the view or whether they held the view prior to participating in the study. This leaves any efforts to evaluate folk metaethics in a catch-22: *either* we don't provide training, in which case our measures aren't valid because participants don't interpret questions about metaethics as intended, or we do provide training, in which case we can no longer be sure that responses aren't the result of *spontaneous theorizing*.

### 3.3.2 Philosophical induction

Spontaneous theorizing is not even the most serious, or only, problem with training paradigms. There is a second, even more damning catch-22: if your attempt to train participants fails then you won't have a valid measure of their metaethical stances or commitments. However, if you succeed, then your participants are no longer ordinary people, and thus no longer members of the sample population they're supposed to represent. This is because adequately training participants to understand the relevant philosophical concepts and distinctions necessarily involve inducting them into the categories and distinctions *as they are understood in academic philosophy*, and thus requires them to think *within the framework of contemporary analytic philosophy*.[55] In other words, training paradigms succeed only insofar as

---

[55] and more specifically in accordance with mainstream metaphilosophical presuppositions implicit in the structure and framing of the metaethical concepts and distinctions presented in instructions and exercises. I develop on this concern in **Supplement 3.**

they succeed at *philosophical induction*, i.e., *inducing participants to become philosophers*, if poorly trained novice philosophers.

Of course, people who have just spent a few minutes of training are very different from academic philosophers who've spent years studying philosophical topics. But this is not a problem for my account: is this training good enough, or not? If it's not good enough, then we're back to square one. If it is good enough, then what does that entail other than that these people have crossed the threshold to be inducted into the hallowed halls of "philosophers"? Even if we wish to reserve this moniker for people who have passed an even greater threshold, that is still not a problem. We might think of participants who successfully pass the tasks presented in training paradigms not as philosophers *or* ordinary people, but something in between: *quasi-philosophers*. Training paradigms don't fail only if participants are philosophers, all that's necessary is that they are no longer ordinary people. If participants in these studies are something in between, then we still face a serious problem: the responses of quasi-philosophers may systematically differ from ordinary people (and philosophers, for that matter). And since we have no extraneous way to measure the responses of ordinary people as a point of comparison, we have no way to know how significant this deviation is. Measuring the responses of quasi-philosophers is thus never going to serve as an appropriate method for drawing inferences about ordinary people. What we'd be left with is a set of measures for a sample of people who were so changed by the study they participated in that our findings would lack any external validity. If the only way to measure someone's views is to teach them philosophy, and an ordinary person is, by definition, someone who hasn't been taught philosophy, then ordinary people are a population that is inaccessible to folk metaethics research. After all, you can't study the psychology of "people who have never seen *The Matrix*" by having people who hadn't seen *The Matrix* before they came into the lab come in, watch it, then answer questions about it.

This pair of catch-22s demonstrate that training paradigms were doomed from the start. The training paradigm requires researchers to causally interact with the participants in a way that (a) may cause relevant psychological changes in the sample and (b) *requires* changing the sample in such a way that the sample no longer represents any meaningful population outside the context of the study (and thus strips the paradigm of external validity). Pölzler and Wright correctly recognize that previous research on folk methods suffers fatal methodological shortcomings. Unfortunately, their cure is worse than the disease.

### 3.4 Conclusion

I have raised a variety of methodological worries about the current state of research on folk metaethics. These are not trivial concerns that can be cast aside and ignored. The current direction researchers in folk metaethics have taken suggests that they are inclined to agree. At least some researchers have introduced a variety of methods to circumvent the shortcomings of earlier work. Unfortunately, these efforts don't succeed, and indeed, *can't* succeed, for the reasons I've just outlined.

Yet so far my critiques have remained purely theoretical. I have yet to present any empirical work that would support these criticisms or support indeterminacy. That is the goal of the remaining two chapters. In chapters 4 and 5, I embark on an effort to demonstrate that most people do not clearly interpret questions about metaethics as intended. These rates are high enough to cast serious doubt on the validity of existing folk metaethics research. In addition, I present the results of a study that reveals how easy it is to conduct a study that could give the misleading impression that I've identified genuine patterns of folk philosophical thought, even when no such patterns plausibly exist. This highlights the possibility that research on folk metaethics is a byproduct of the forced choice design employed by most research.

# CHAPTER 4:
## Open Response Paradigms

### 4.0 Introduction

In the preceding chapters, I argued that existing research on the psychology of metaethics is too riddled with methodological shortcomings to provide compelling evidence that ordinary people hold determinate metaethical stances or commitments. While I am confident these criticisms are sufficient to justify this conclusion, I have yet to provide any empirical evidence that would support folk indeterminacy. The remaining chapters seek to rectify this.

In this chapter, I assess the results of open response data. This data consists of written responses that reveal how participants interpreted questions about metaethics. The goal of these studies is to determine the proportion of participants who interpreted questions as researchers intended, i.e., as questions about moral realism or antirealism. If most participants do interpret these questions as intended, this would provide significant evidence that these studies are valid, and that people's responses to these questions genuinely reflect their metaethical stances or commitments. Conversely, if most people don't interpret questions about moral realism and antirealism as researchers intend, that would suggest that these studies are not valid.

Even if low rates of intended interpretation do not decisively undermine the validity of existing paradigms, widespread failure to interpret questions as intended calls for an explanation. While there are many plausible reasons why so many participants would appear to interpret questions about metaethics in unintended ways, indeterminacy offers a straightforward explanation for low intended interpretation rates: people do not understand the questions because they have no metaethical stances or commitments with respect to realism and antirealism. Thus, this chapter serves both to reinforce the case made in previous chapters that existing metaethics paradigms are invalid, and to offer empirical evidence of metaethical indeterminacy.

**4.1 Rationale for using open response data**

Given the methodological limitations of qualitative research (Guerin, Leugi, & Thain, 2018; Willig, 2008), coupled with its poor reputation among more quantitatively oriented social scientists (Kvale, 1994; Povee & Roberts 2014; Queirós, Faria, & Almeida, 2017; Rahman, 2016) it is reasonable to question an extensive reliance on open response data. After all, my conclusions are largely based on coding items based on my own judgment. This may introduce an unacceptable degree of bias and subjectivity into my results (e.g., confirmation bias, Nickerson, 1998). In addition, I am not blind to my own hypotheses, and awareness of my vulnerability to cognitive biases may do little to mitigate their impact (Aczel et al., 2015; Compen et al., 2022, Fischoff, 1982; Welsch, Begg, & Bratvold, 2007; cf. Savitsky & Gilovich, 2003). Qualitative data can also be vague. Researchers may be hesitant to draw firm conclusions based on the rough impressions they or anyone else has of a set of written responses. In short, qualitative data may seem too error-prone and crude to be worth serious consideration.

These concerns are not without merit. Concerns about my findings are especially plausible since my results are likely to come as a shock. When I initially set out to assess open response data, I expected many participants to interpret questions about metaethics in unintended ways. But I never expected the rates to be so low. Across all studies, *most* participants did not interpret questions about metaethics as intended. In some cases, the percentage of participants who demonstrated a clear understanding of what they were asked is in the single digits. I am confident these findings challenge anyone convinced that ordinary people have determinate metaethical standards, yet skeptical readers may believe my assessment is too uncharitable, and that I have grossly underestimated the degree to which ordinary people correctly interpreted questions about metaethics.

I readily acknowledge this concern. Perhaps my method of coding is flawed, or there is some other reason why these findings tell us little about what, if any, metaethical stances and commitments characterize folk morality. Yet I believe the data will speak for itself. I have made all of the data readily

accessible, and invite readers to examine it. While I cannot guarantee that everyone will agree with how I coded each response (in fact, I am sure they won't), I am confident that few people familiar with the relevant metaethical distinctions will come away with the impression that there is little cause for concern.

Of course, confidence in my conclusion hardly justifies relying so extensively on qualitative data. Why, then, have I devoted an entire chapter to these results? First, the data I present in this chapter is not intended to provide an independent and decisive case for or against any particular hypothesis. Rather, my findings primarily function to check the validity of paradigms that are themselves subject to quantitative analysis. Insofar as my findings have additional implications, they should be taken only as suggestive evidence to be considered in conjunction with theoretical considerations and other sources of evidence.[56] Whatever its limitations, qualitative data can be used alongside quantitative data to *triangulate* on and potentially reach the same conclusions (Morses, 1991; Olson, Haralambos, & Holborn, 2004). When this occurs, quantitative and qualitative data can work in tandem to provide mutually corroborating evidence for the same conclusions. As such, I do not propose to supplant the rigor of quantitative data, but to complement it.

Second, my central argument is that most people consistently fail to interpret questions about metaethics as intended. Analysis of open response data can provide evidence for or against this hypothesis, whatever its limitations may be. And if I am correct, this is not a trivial concern. That participants understand what they are being asked is a *necessary* condition for validity. After all, if people do not understand what they are being asked, their responses will not reflect the psychological

---

[56] Thus, these findings serve an auxiliary role alongside other sources of evidence. I do not wish to understate the relevance of these findings, however. In principle, findings across the social sciences should converge on the same conclusions if those conclusions are genuinely robust. We should expect findings in fields as disparate in their methodology as history, anthropology, psychology, and neuroscience to converge or at least be mutually intelligible and consistent with one another. If anthropological data overwhelmingly suggested that the vast majority of people were moral realists, but the methods used by social psychologists and experimental philosophers indicated that most people were moral antirealists, this would call for an explanation.

phenomena researchers intend to measure. If open response data could provide even modest evidence that a paradigm or an entire body of empirical research were predicated on invalid measurement tools, such findings could have considerable value in directing subsequent efforts to further evaluate the validity of these measures.

Third, I rely on a large and diverse (though not demographically diverse) body of data, and conduct some quantitative analysis on the data itself, insofar as I estimate proportions of intended and unintended interpretations. This circumvents two common shortcomings with qualitative data: that findings are often based on comparatively small datasets (such as a handful of interviews) and the lack of any form of quantitative analysis. Lastly, I present a series of studies in **Chapter 5** which do take a quantitative approach to assessing how participants interpret metaethics stimuli.

With these caveats in mind, I concede that open response data may prove unhelpful in many cases. However, its use is warranted when there is theoretical justification for suspecting extremely low rates of intended interpretation for quantitative research. In such circumstances, the most efficient path to assessing validity isn't *necessarily* to conduct more quantitative studies. We could imagine a scenario where, for instance, the reason for a reliably recurring pattern in a given dataset is due to an ambiguity in how people interpret a question that researchers didn't notice. Simply asking people how they interpret the question could potentially reveal the presence of that ambiguity more quickly, easily, and decisively than running additional quantitative studies without attempting to directly pinpoint the problem. Indeed, inattention to interpretation may be invisible to many exclusively quantitative approaches, since such approaches can reliably yield data with psychometric properties that don't raise suspicion. Thus, I don't employ qualitative methods as a *replacement* for quantitative methods, but as an *auxiliary* means of assessing the validity of particular quantitative methods, methods which remain the primary way to address empirical questions.

In most cases, there may be little reason to worry that a significant proportion of participants would wildly misconstrue the nature of a task. Competent native speakers of a language typically have little trouble understanding questions or following instructions that are simple, straightforward, and deal with familiar phenomena. When major concerns arise, researchers are (hopefully) able to spot them, and when they overlook potential problems, these problems are (at least under ideal conditions) caught during the peer review process or after publication. In addition, researchers can (and frequently do) run their designs by colleagues, present them in lab meetings, or conduct pretests that can catch any serious issues with the interpretability of their stimuli. And where we recognize the potential for confusion or misunderstanding, we often provide clarificatory instructions or include comprehension checks that assess whether participants interpreted features of a study as intended, which can improve validity (Meisters, Hoffman, & Musch, 2020). In short, the way psychologists conduct research already builds checks against unintended interpretations into the process, and where these fail, the self-correcting nature of science will eventually reveal and allow us to correct for interpretative deficiencies.[57]

These platitudes represent an overly optimistic picture of the social sciences built more on its presumptive ideals than how it actually functions. It is possible that in many cases the aforementioned methods aren't employed or are inadequate even when they are widely used, e.g., the efficacy of peer review remains a matter of serious contention (Goldbeck-Wood, 1999; Justice et al., 1998; Kelly, Sadeghieh, & Adeli, 2014; Mavrogenis, Quaile, & Scarlat, 2020; Smith, 2006; 2010; Tennant & Ross-Hellauer, 2020). There is also an irony in treating quantitative methods as the *sine qua non* of research when the validity of quantitative approaches often relies on the informal procedures (e.g., lab

---

[57] Even so, few of these concerns are self-consciously designed with the goal of ensuring proper comprehension. While standard procedures may mitigate concerns about interpretation to some limited extent, they aren't typically *designed* to do so. Quite the contrary. There is a remarkably cavalier attitude among social scientists, at least within social psychology and, astonishingly, experimental philosophy, with ensuring adequate comprehension on the part of participants. Simply put, researchers rarely bother to check whether participants interpret questions as intended.

meetings) and the subjective, *qualitative* impressions of researchers, who frequently employ intuitions, hunches, and subjective experience in lieu of formal methods for establishing the validity of their measures.

In many ways, quantitative methods serve as the public face for methods that, when we pull back the curtain, often depend on non-quantitative and informal procedures that draw on the subjective judgments of researchers. Indeed, in many (or perhaps even most) cases researchers don't use quantitative approaches to validate their measures, or at least don't report doing so. Instead, they frequently employ measures developed "on the fly," without addressing *any* formal steps relevant to validation.[58] Even though formal validation procedures are available for e.g., scales, a recent assessment of the scales reported in articles published in the *Journal of Personality and Social Psychology (JPSP)* in 2014 found that:

> Roughly, half of these scales (46%) included no reference to previous validation, appearing to have been developed on the fly. α was the only psychometric information reported for half of these scales which had no previously published validity evidence, and 19% had no accompanying psychometric information. (Flake, Pek, & Hehman, p. 5)

These findings suggest that even researchers publishing in premiere journals frequently ignore the validity of their measures. This is not a minor concern. As Flake and Fried (2020) observe, lack of transparency "prevents the evaluation of all aspects of a study's validity: its internal, external, statistical-conclusion, and construct validity" (p. 457). They add that "recent research on commonly used measures in social and personality psychology showed that measures with less published validity evidence were less likely to show strong evidence for construct validity when evaluated in new data" (p. 457). Lack of validity evidence thus plausibly serves to conceal poor validity and enhance opportunities for researchers to (inadvertently or otherwise) engage in *questionable measurement practices*

---

[58] And I haven't even begun to assess the role subjectivity plays in researcher degrees of freedom, which have allowed p-hacking to persist for so long (Wicherts et al., 2016).

(QMPs), i.e., "decisions researchers make that raise doubts about the validity of the measures used in a study, and ultimately the validity of the final conclusion" (p. 458). An uncritical reliance on exclusively quantitative methods, coupled with a culture that places little emphasis on validity, does not put us in a strong position to reject qualitative and mixed methods approaches as soft or inadequate, especially when the very purpose of those methods is to evaluate the validity of the measures used in quantitative research, measures the are routinely devised using informal and non-quantitative procedures.

If so, researchers may have overlooked serious methodological shortcomings, such as high levels of interpretative variation and unintended interpretations using particular paradigms, or with respect to particular types of questions. Psychology is in the midst of a replication crisis (Nelson, Simmons, & Simohnson, 2018; Open Science Collaboration, 2015; Wiggins & Christopherson, 2019), and researchers continue to grapple with numerous methodological shortcomings related to design and data analysis; it is not outlandish to propose that there may be yet another one: a widespread failure to ensure participants interpret stimuli as intended, an instance of a broader but oft-ignored problem crisis of measurement and validity (Flake & Fried, 2020; Lilienfeld & Strother, 2020; Schimmack, 2021).

Such concerns may be irrelevant to a great deal of research that presents participants with comparatively simple, straightforward, and familiar considerations. Unfortunately for researchers who study the psychology of metaethics, there is nothing simple, straightforward, or familiar about moral realism and antirealism. The previous chapters provide more than ample reason to conclude people struggle to think in metaethical terms. Indeed, even researchers themselves struggle to clearly and consistently describe the metaethical landscape, or to devise paradigms that adequately operationalize metaethical concepts in their stimuli.

As argued in Bush and Moss (2020), this is because metaethics is unlike many other subjects social scientists present in surveys, and its differences render it uniquely vulnerable to low rates of intended interpretation. There are four main factors that contribute to an increased likelihood that participants will interpret questions about metaethics in unintended ways:

*(i) the relevant metaethical theories are complex and difficult to grasp*

*(ii) most people are unfamiliar with these distinctions prior to encountering them in studies*

*(iii) metaethical theories are generally abstract and distant from real world practical questions lay populations would be more familiar with and expect to be asked about*

*(iv) there are typically plausible non-metaethical interpretations of the questions posed to respondents, and participants may be more likely to interpret surveys as posing these more prosaic questions rather than abstruse metaethical questions* (p. 7)

Given these characteristics, when we compare research on folk metaethics to conventional psychological research, it might come as more of a surprise if most participants *did* interpret questions about metaethics as intended. Although I do not consider any of these factors especially controversial, critics may object that moral realism is not difficult to grasp and that people are familiar with it. David and I discuss why we do not believe this is likely in Bush and Moss (2020).

## 4.2 General methods and procedures

### 4.2.1 General methods

The studies presented in this chapter all share a common structure. In some cases, participants were first asked a question about their metaethical stance. They were then asked to explain *why* they answered this question the way they did or *how* they interpreted one or more of the questions they answered. In other cases, they were presented with a statement, and were then asked to explain what that statement means in their own words or to explain how they interpreted the statement. These questions were either presented in a survey alongside other questions, or as standalone questions (i.e.,

the entire study simply consisted of asking such questions). In all cases, participants were presented with a textbox and wrote out their responses, rather than using Likert scales or selecting from among a set of multiple choice options. All responses were collected in English, and were uncorrected—any typos, misspellings, poor grammar, or uninterpretable remarks were left as they were.

All studies consist of short and straightforward questions, and the typical responses reflect this simplicity, since most consist of a few words or at most a sentence or two. Note that only the *structure* of these questions was simple; since the content itself concerns metaethics, it remains a daunting and complicated task to adequately interpret what these questions are asking, or at least, that is what I hope to show in analyzing the results. In some cases, participants opted to write more than this, but this is atypical. Thus, most of the data consists of short written responses, in which participants explained their answer to a question or offered an interpretation of a question or statement. Here is a sample question and response:

> Question: *In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is objective?*

> Response: *You are saying it is an objective fact murder is wrong. An objective fact is one which is not based on human judgement or belief.*[59]

A few datasets were collected by other researchers and were included in previously published articles. Wherever I analyze such data, I will make an explicit reference to whose data I am assessing, and which published articles are associated with that dataset. However, most of the datasets were collected by myself, either alone or in conjunction with collaborators. Unless otherwise specified, any data presented is my own and all analyses were my own.

To my knowledge, no previous studies have systematically analyzed open response data with the goal of assessing the proportion of participants who interpreted questions about metaethics as

---

[59] This is an excellent and clear response.

intended. While some researchers have included open response data in their studies (e.g. Goodwin & Darley, 2008; Sousa et al., 2021; Wainryb et al., 2014; Wright, Grandjean, & McWhite, 2013), and do use it to check interpretation rates, such data has rarely featured prominently in research results.[60,61]

### 4.2.2 General procedures

All studies were analyzed in two ways. First, *interpretation rates* for each study were coded using the same method. This is the primary method of analysis and the central focus of this chapter. My goal was to determine the proportion of participants that interpreted questions (and other stimuli) used in research on folk metaethics as researchers intended. David Moss has a background in philosophy and social science. We have collaborated on projects in metaethics and folk metaethics and he has published independent research on methodological problems in folk metaethics research in the past (Moss, 2017). Given our extensive interactions over the course of my research, I consider David an extraordinarily competent coder who is familiar both with the relevant metaethical distinctions and is highly sensitive to the methodological concerns motivating this project, e.g., ambiguity, conflations, and the role pragmatics can play in influencing how people interpret stimuli.

Second, I developed a set of themes to provide richer qualitative assessment of the content of people's responses. My approach does not rigidly adhere to any formal methodology, and as such could be considered a generic qualitative approach (Kahlke, 2014). However, my approach most closely reflects *thematic analysis* that is similar in many ways to the *reflexive thematic analysis* approach pioneered by Braun and Clarke (2006; 2014; 2019; 2020; Clarke & Braun, 2013). This involved an

---

[60] One notable exception is Wainryb et al.'s (2004) extensive analysis of open response data designed to assess how children interpreted questions about metaethics. Unfortunately, this data was not available for me to assess directly, so I had to rely on the analysis presented in the article. As I discuss in **Supplement 2**, virtually all of the participants in these studies appeared to interpret questions about metaethics in unintended ways. This is consistent with my findings, but since I cannot assess the data myself, this conclusion should be interpreted with caution.

[61] Even when researchers do appeal to open response data, they rarely do so in a systematic and thorough way. For instance, Sousa et al. (2021) appeal to open response data to support their explanation of their findings. However, they only highlight one or two examples to support a point, rather than appealing to systematic analysis of the data they collected.

iterative process of familiarizing myself with the general patterns and themes that emerged in each set of data, reviewing, refining, and naming those themes, then reviewing the data again in light of these themes to code each item in accordance with one of the named themes.

This type of analysis plays only a secondary role in the findings reported here. While thematic analysis of the particular ways participants interpreted questions about metaethics in unintended ways is both fascinating and worthy of further study, my primary concern is with the overall rate of intended and unintended interpretation. As such, I have relegated discussion of this data to **Supplement 4**. Here, my focus will focus on interpretation rates.

### 4.2.3 Coding and analysis for interpretation rates

All written responses were coded using a two-dimensional quaternary system. First, all items were coded based on whether they appeared to reflect the intended metaethical distinction (1) or not (0). Second, they were coded based on whether the response was clear (1) or unclear (0). This created four possible ways an item could be coded:

1 | 1 = Clearly intended interpretation
0 | 1 = Clearly unintended interpretation
1 | 0 = Unclear (leans towards intended interpretation)
0 | 0 = Unclear (leans towards unintended interpretation)[62,63]

Judgments about how to code items in accordance with this system developed over the course of coding, and rely as much (or more) on experience and familiarity with the data that would be difficult to articulate as they do on general and definable characteristics that could be readily summarized. Nevertheless, I can state roughly how judgments were made. Items coded as clearly intended interpretations (1|1) needed to plausibly express a view, however inchoate or clumsily phrased, that

---

[62] Participants who did not respond were coded as 0|0.
[63] Early versions of this coding scheme were initially proposed by David Moss.

seems consistent with the intended interpretation of the stimuli. There was no expectation of a sophisticated or detailed response. For instance, the validity of the disagreement paradigm depends on participants interpreting disagreements to result from genuine disagreement in moral values (rather than e.g., disagreements over nonmoral considerations). To highlight just how simple a clearly intended interpretation could be, the response "Different standards of morality" was coded as a clearly intended interpretation. I did my best to extend this same low bar for a clearly intended interpretation across all datasets.

In contrast, another participant proposed that the "Other individual misread the question." This was coded as a clearly *unintended* interpretation (0|1). Clearly unintended interpretations represent any instance in which the participant expressed a coherent interpretation of the stimuli, but this interpretation was inconsistent with the intended interpretation. This could be due to conflating some other consideration with a metaethical one (e.g., interpreting the question to be asking about their first-order normative judgments), not accepting the information presented in the study as stipulated (e.g., if they are told another person judged that murder or stealing were acceptable, they propose that the other person misunderstood the question), understanding specific terminology in a way inconsistent with researcher intent (e.g. understanding the term "objective" to mean *unbiased*), or invoking metaethical concepts, but expressing an understanding of them that is inconsistent with the relevant metaethical distinction (e.g. stating that morality "is objective," then adding that objectivism means that moral standards "apply to everyone" or are "clear and well-defined."). An unintended interpretation is *not* a "misinterpretation." Misinterpretation implies fault or error on the part of participants. Yet as much or more of the problem lies in the inherent ambiguity and underspecificity in the way questions were phrased, and in some cases, due to researcher error (e.g., researchers employed stimuli that formally conflated realism/antirealism with other concepts or distinctions).

103

The distinction between unclear intended and unintended interpretations can be a bit murkier. Roughly, unclear intended interpretations could be consistent with researcher intent, but are sufficiently unclear, internally inconsistent, or ambiguous to undermine confidence about how the participant interpreted stimuli. Unclear unintended interpretations include several types of response that are easy to classify: instances where the participant did not respond or did not offer a substantive response (e.g., writing "n/a" or "."), incoherent or uninterpretable remarks, responses that clearly ignored the task (e.g., directing a remark towards researchers, such as "this study sucks!"), or were otherwise irrelevant or lacking in substance (e.g., "MORALITY"). Finally, some instances or unclear unintended interpretations leaned in the direction of suggesting an unintended interpretation, but were too confusing, vague, or underdeveloped to justify confident classification. These general remarks about how responses were coded are far from adequate to capture the full range of considerations behind coding decisions, and discussion of individual items is often the best way to highlight how and why particular decisions were made. I provide a more extensive discussion of the coding procedure, along with examples, in **Supplement 4**. With the exception of Study 1A, all of the coding reported here and in the supplements was conducted exclusively by myself.

### 4.3 General predictions

I had no initial predictions about the precise proportion of participants I expected to fall within each category. Since Study 1A was the first dataset that I analyzed, I didn't know what to expect, so there was no particular prediction made in advance for those findings. However, the incredibly low interpretation rate initially obtained for that study (see Bush & Moss, 2020 and **Supplement 4**) led me to anticipate that fewer than half of participants would interpret questions as intended, regardless of the measure used. Since no previous research adequately estimated interpretation rates for the folk metaethics research, I had no previous data to rely on in offering precise estimations. And given the

wide variation in interpretation rates across studies, it would be difficult to make precise predictions about any particular measure.

However, the main reason I didn't set out with precise predictions is that it would be inappropriate to do so. Coders (such as myself) were not blind to expectations about interpretation rates. To code data with precise prediction in mind would risk biasing coding to conform to expectations. Thus, while I offer a nominal prediction that fewer than half of participants will interpret metaethics stimuli in a clearly intended way, due to the non-blinded nature of the coding it would be inappropriate to treat these anticipated interpretation rates as true "predictions." Since I am coding the data myself, there would never be a way to eliminate the risk that I am biased in such a way that I code to conform to a given "prediction." As such, statistical tests should be interpreted as *post hoc* and effectively exploratory, not reflections of an unbiased assessment of the data, and definitely not as predictions in the absence of observing the data, since this wasn't possible.

This is obviously less than ideal. To anticipate one criticism with this data: it's unclear whether it's feasible in practice to obtain blind coders with sufficient competence to adequately code the data. Insofar as this limits the inferential power of my analyses, this shortcoming is simply unavoidable. While one might suppose that blind coders would be better than nothing, I am not convinced this is the case. *Systematic* bias may result in interpretation rates that I'm not confident would accurately reflect the true proportion of clear interpretations. Nevertheless, future efforts can and should be made to train blind coders. The only other ways to shore up concerns about bias would require corroborating evidence, perhaps from a combination of adversarial coders conducting similar research and the use of conventional quantitative methods where genuine predictions about interpretation rates can be tested without concern. I take the latter approach in **Chapter 5**.

## 4.4 Summary of studies

Studies 1A-1C evaluate whether people presented with the disagreement paradigm interpret the source of disagreement as intended (i.e., as a difference in fundamental moral beliefs.

Studies 2A-2D present participants with a question and a response from a previous participant. Interpretation rates focus on how participants interpret the previous participant's *response* rather than the question the previous participant was asked. This allows us to artificially construct a clear articulation of an intended response to the initial question, with the participant's task merely serving to interpret a statement reflecting a metaethical position.

Studies 3A-3C explore interpretation rates in the context of a concrete moral decision related to charitable giving. A noncognitivism condition was added to broaden the range of metaethical views reflected in the questions.

Studies 4A and 4B evaluate interpretation rates for questions that explicitly ask about the participant's metaethical views using standard metaethical terminology, e.g., "objective" and "relative."

Studies 5, 6, and 7 evaluate interpretation rates for the questions that appear on scales used to measure folk metaethical views. Study 5 assesses interpretation rates for the Moral Relativism Scale (MRS, Collier-Spruel et al., 2019). Study 6 assesses interpretation rates for scale items employed by Yilmaz and Bahçekapılı (2015a; 2015b; 2018). Finally, Study 7 addresses the items used in the FMO (Zijlstra, 2019). All three studies employ two distinct measures: asking participants *why* they answered the question the way they did, and asking them to *explain* what they think the question means. This provides a converging pair of measures of interpretation rate for each item.

## 4.5 Studies 1A & 1B: Source of disagreement in Goodwin & Darley (2008)

**Methods**

Studies 1A and 1B are a reanalysis of open response data collected by Goodwin and Darley (G&D, 2008). G&D gathered open response data for two of the three experiments they conducted,

Experiment 1 and Experiment 2. The goal of these studies was to evaluate whether ordinary people are moral realists or antirealists.[64] Although both studies used the disagreement paradigm, there are additional details to this study that I will describe to put my reanalysis of the open response data in context.[65]

All participants rated their level of agreement with a set of 26 moral and nonmoral statements on a six-point Likert scale (1 = *strongly disagree,* 6 = *strongly agree*). Nonmoral statements were drawn from three additional categories: social convention, aesthetics, and "factual" statements (e.g., statements about science or geography). For each statement, they were then asked whether it was a:

- *True statement.*
- *False statement.*
- *An opinion or attitude.[66]*

This was followed by having the participant perform an unrelated task. The goal of this phase was to give researchers time to select five responses (2 moral, and 1 social convention, aesthetic, and factual question). Once these items were selected, the study resumed, and participants were told that participants from a previous study using the same stimuli disagreed with each of these statements. They were then asked to consider this disagreement and select one of the following responses:

- *The other person is surely mistaken.*
- *It is possible that neither you nor the other person is mistaken*
- *It could be that you are mistaken, and the other person is correct*
- *Other*

---

[64] They do not use this terminology. While they explicitly state that their goal is to distinguish a stance-independent conception of morality from a stance-dependent conception of morality, the refer to the former as "objective" and use a range of different terms to refer to the latter, including "non-objective," "subjective," "opinion-based," and "preferential." It is hard to fault the lack of precise and consistent terminology here, given that it is difficult to characterize a range of conceptually distinct non-objective views in any simple and straightforward way. While I believe "antirealist" is adequate, a reasonable person might have qualms with this terminology.

[65] Both studies also used an additional paradigm to generate a combined measure of folk metaethical belief. However, the additional measures are not relevant to this analysis. See Goodwin & Darley (2008) for details.

[66] This is not a valid measure of moral realism/antirealism. This question conflates realism/antirealism with cognitivism/noncognitivism (Pölzler, 2018b). It is also highly vulnerable to be interpreted as a question about normative ethics. It is also consistent with asserting something true or false that it is an "opinion or attitude." Indeed, some uses of "opinion" are cognitivistic. As such, the "An opinion or attitude" option is not clearly mutually exclusive with the other response options.

The first option was interpreted as a realist response, while the second and third were interpreted as antirealist responses. Later in the survey, participants were asked to explain what they thought the source of the disagreement between themselves and the previous participant could be. Specifically, they were asked:

*Give us your thoughts about why it is that there is disagreement. What could be its source?*[67]

Goodwin and Darley included this question because they were concerned that some participants might interpret the source of the disagreement in ways that would be inconsistent with the interpretation necessary for their response to reflect a metaethical stance or commitment.[68] It will come as no surprise that I agree with their concerns. This is a laudable step to ensure that participants interpret the question as intended, and more researchers should employ questions like this. They also report reassuring results:

> When individuals did not interpret the disagreement in a bona fide way [...] we excluded such data from all foregoing analyses. In fact, only seven out of a total of 102 responses were excluded on these grounds.

If their analysis of participant explanations is correct, this would indicate that less than 7% of participants interpreted the disagreement in a way that would threaten the validity of the disagreement paradigm. If so, there would be little cause for concern.

Unfortunately, my reanalysis of their findings suggests that they underestimated the rate of intended and unintended interpretations. However, we must pause to consider what would constitute a "bona fide" interpretation of the source of disagreement. Goodwin and Darley (2008) provide a handful of examples, then remark more generally that they excluded "[r]esponses in which participants

---

[67] At least, this is what I infer from analyzing their raw data. Geoffrey Goodwin was kind enough to share their data and their original coded analysis of the open responses, and this was the question that appeared in that file.

[68] As they put it: "We were wary of the fact that some individuals might interpret ethical disagreement as indicating that the disagreeing other person might have been thinking of extraordinary extenuating circumstances, or that they had misread the question, or that they may not have understood the words used in a conventional way, and so on [...] The list of such possible caveats to interpreting the prima facie ethical disagreement is large." (p. 1347)

said they needed more information about the context of the events were also excluded for this reason"

(p. 1348, footnote 5). Here is one concrete example of a response that they excluded:

> *A difference in perception of a situation in which gunfire was opened on a crowded city street. I was thinking gunfire from terrorists/criminals; other person may have thought gunfire from police officers to catch a criminal.* (p. 1348, footnote 5)

Finally, they state they did not exclude "responses in which participants said that perhaps the disagreeing other was operating with a different sense of "morally wrong/bad."" While they do not explicitly state what counts as a "bona fide" interpretation, it is not hard to determine what type of interpretation they have in mind, or at least *ought* to have in mind: *fundamental moral disagreements*, i.e., disagreements about basic moral values, rather than differences in nonmoral beliefs or standards (Bush, 2016; Caven, 2015; Moody-Adams, 2009). As Caven (2015) puts it:

> Fundamental moral disagreement involves two agents who are situated in the same context, possess the same non-moral understanding and are both free of errors of inferential reasoning, yet come to different judgements concerning what is the morally best course of action, all-things considered. (p. 3)

For instance, Alex and Sam could disagree about whether a particular killing was justified because Alex believes the killing was an act of self-defense, and Sam believes it was an act of premeditated murder. Both agree that killing in self-defense is morally justified, but that premeditated murder is not. Yet they could disagree about whether the particular event that took place was an act of self-defense or an act of premeditated murder. If so, Alex and Sam do not have a fundamental moral disagreement, because they share the same moral stance towards self-defense and premeditated murder. They just disagree about what actually happened (i.e., the nonmoral facts).

If participants attribute the source of the disagreement between themselves and a previous participant to something other than a fundamental moral disagreement, then there is no way to be sure whether their response reflects a metaethical stance or commitment. For instance, if participants thought that the previous participant who ostensibly disagreed with them was not thinking of the same

moral issue as they (the participant) were, then whether the previous participant's judgment was "correct" or "mistaken" would have no bearing on realism and antirealism. After all, if I think premeditated murder is wrong, and you inform me that someone else thinks killing in self-defense is not wrong, then both the judgment that they are mistaken *and* the judgment that neither of us is mistaken are consistent with realism and antirealism. In short, participants *must* have regarded the disagreement between themselves and a prior participant to be a fundamental moral disagreement. If they did not, then their response does not reflect a metaethical stance or commitment.

***Participants.*** Responses were initially collected from the participants who participated in Experiment 1 (*n* = 50) and Experiment 2 (*n* = 76) in Goodwin and Darley (2008).[69] Experiment 1 consisted of 50 Princeton University undergraduates (28 females, 22 males) who participated in the study for course credit, while Experiment 2 included 76 students drawn from two populations: (1) 71 Princeton University undergraduates (49 females, 29 males) who participated for course credit and a second sample of five theological seminary students who participated for $10 (these samples were collapsed into a single sample for analysis).[70] No other demographic details were provided.[71] Experiment 2 likewise included two responses from each participant, resulting in a total of 152 responses. This resulted in a total of 252 responses across Experiments 1 and 2 (corresponding to Studies 1A and 1B, respectively).

---

[69] Details about responses are based on what was reported in Goodwin and Darley (2008).

[70] The rationale for including five seminary students was to increase the number of participants likely to attribute their moral standards to a divine source.

[71] Respondents in Experiment 1 answered two open response questions. In Goodwin and Darley's original article, they report 102 responses, and appear to list 51 subjects. However, only 50 subjects are listed in the data. Upon examination, participants are numbered 1 through 51, but respondent #23 is missing from the dataset. Thus, it is unclear whether there were 50 participants who each responded twice, resulting in 100 responses, or whether there were 51, resulting in 102 responses. Since two of these responses were not available to us, I only coded and analyzed the 100 available responses. Goodwin and Darley make no explicit reference to this inconsistency in the data reported in their original article, though they reference having only 50 participants, yet later report 102 written responses (they do reference "missing data" in a note under Table 1, on p. 1347).

***Procedure***. I do not have complete access to the original study materials and did not conduct the original study myself. As such, this reanalysis of the data was conducted based on data collected following the procedures presented in Goodwin and Darley (2008), and my inferences about what that procedure likely entailed. From the way the study is described in the text, and from the materials I received, it would appear participants filled out printed surveys, and that their responses were subsequently transcribed. After completing the primary measures in the study, participants appear to have completed additional follow-up questions. The specific question analyzed here was included among these questions and appeared as a question separated from others by a bullet, and followed by several empty underlined lines in which participants could write a response, like this:

- Give us your thoughts about why it is that there is disagreement. What could be its source?

  _____

  _____
  _____
  _____.

Written responses were presumably transcribed. The original data I received was in an Excel spreadsheet, so I presume the written responses were either transcribed directly to it or transcribed elsewhere and then copied and pasted to it. Interpretation rates for Experiment 1 and Experiment 2 were analyzed separately.

***Measures.*** Interpretation rates for all items in Study 1A were coded by myself and David Moss. Interpretation rates for all items in Study 1B were coded exclusively by myself. Participants provided a written response to the question outlined above. All responses were coded in the two ways described above. Each item was coded as either an intended ("1") or unintended ("0") interpretation and as being clear ("1") or unclear ("0").

These codes were then combined to form one of four codes:

  *1 | 1 = Clearly intended interpretation*
  *0 | 1 = Clearly unintended interpretation*
  *1 | 0 = Unclear intended interpretation*
  *0 | 0 = Unclear unintended*

A second coder (David Moss) was recruited and conducted the same analysis of the data for Experiment 1. A clearly intended interpretation was one that indicated the participant attributed the source of disagreement to a *fundamental moral disagreement* (see above; see also Bush, 2016; Caven, 2015; Moody-Adams, 2009). A clearly unintended interpretation attributed the source of disagreement to something other than a fundamental disagreement, such as the other person misunderstanding the question or imagining a different moral situation than the participant. Unclear interpretations (1 | 0 or 0 | 0) failed to clearly convey either an intended or unintended interpretation.

After initial coding, I reviewed the data to identify sources of disagreement. We attempted to resolve all disagreements via discussion. In some cases, we reached an agreement and coding disputes were resolved. In other cases, we maintained our disagreement, and in those cases responses were not changed by either of us. No other coders participated in the analysis of Experiment 2. Thematic analysis for both Experiment 1 and Experiment 1 was conducted using the method described above, though neither David nor any other coders were involved in this process.

**Study 1A Results**

***Interpretation rates.*** Overall, 25% (*n* = 25) provided clear intended interpretations. A one sample proportion test without continuity correction was conducted to determine whether the proportion of responses coded as clearly interpreted was significantly less than 0.5, and it was significantly less than 0.5, $\chi2(1, N = 100) = 25$, $p < 0.001$.[72] With 95% confidence, the percentage of responses coded as

---

[72] I report a chi-squared test statistic ($\chi2$) because all analyses were conducted in R using prop.test(). Although prop.test() conducts a chi-squared test, $\chi2$ is equivalent to the square of the Z-statistic. I provide the following example in chapter

clearly interpreted was less than 32.72%. In addition, 44% (*n* = 44) of participants responded in a way that was coded a *clear unintended* response, while the remaining 31% (*n* = 31) of responses were *unclear*. David Moss coded 20% (*n* = 20) of responses as clear intended, 57% (*n* = 57) as *clear unintended*, and 33% (*n* = 33) as *unclear*. Results may be seen in **Figure 4.1**.

**Figure 4.1**

*Interpretation rates for Study 1A*
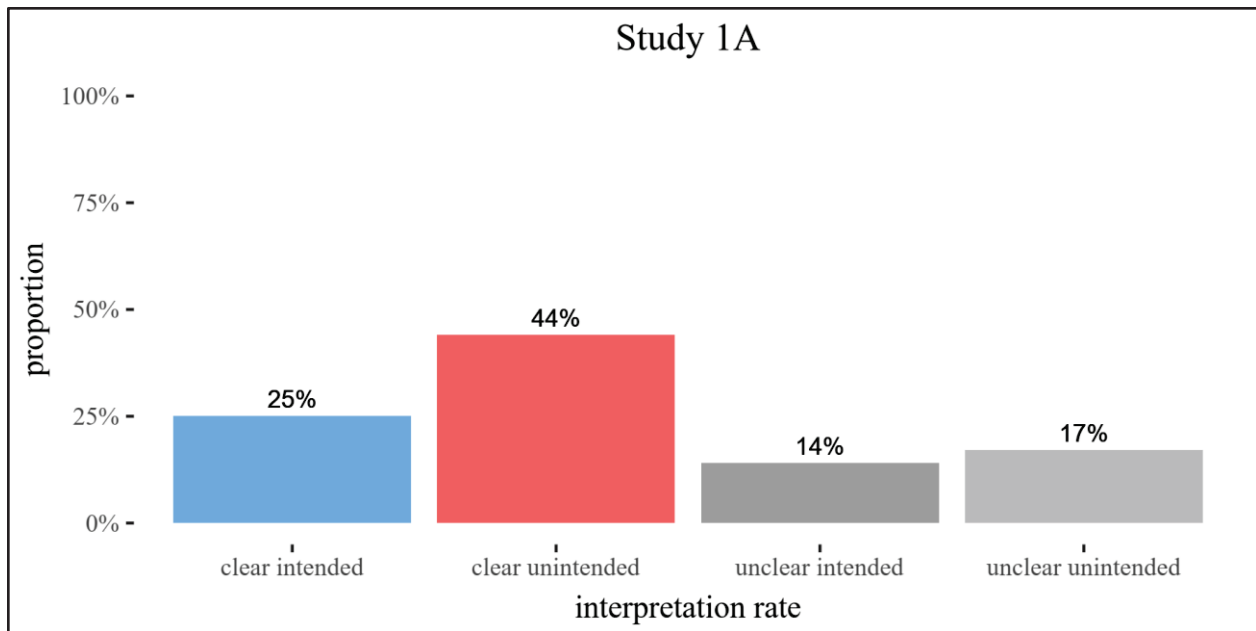


**Interrater reliability.** Initial coding was conducted independently by myself and one other coder, David Moss. David Moss is an experienced researcher competent in both qualitative research and metaethics who has collaborated with me on numerous projects and is familiar with the literature both on the psychology of metaethics and critiques of the assumption that folk metaethics is uniform and determinate (as expressed by e.g., Gill, 2009 and Loeb, 2008). David has also previously conducted

---

five as well as here: the formula for a one proportion Z-test is $Z = (p-p_0) / \sqrt{p_0(1-p_0)/n}$. For instance, the Relativism 1 condition in Chapter 5, Study 1A, this would be $Z = (0.24-0.5)/\sqrt{(0.5(1-0.5)/50} = -3.676955$. $(-3.676955)^2 = 13.52$. Radziwill (2015) explicitly commented on this fact, observing that a one sample proportion test in R technically conducts a chi-squared test, noting that "The prop.test function doesn't even do a z test. It does a Chi-square test, based on there being one categorical variable with two states (success and failure)!"

independent qualitative research on folk metaethics (in the form of interviews). David and I have worked together extensively on the psychology of folk metaethics and normativity more generally (see Moss & Bush, 2022). In fact, David proposed the coding scheme used here. Nevertheless, coding with a non-adversarial collaborator has the disadvantage that we share similar expectations about the data and are thus likely to be biased in the same direction.

After coding, results were compared, and disagreements were resolved through discussion. Some disagreements remained unresolved. After attempting to resolve disagreements, interrater reliability was assessed via Cohen's Kappa (McHugh, 2012). I calculated interrater reliability in both a collapsed and uncollapsed form. The collapsed form involved converting all responses that were clearly unintended and unclear into a single category and distinguishing this category from clearly intended interpretations. This allowed me to assess agreement specifically on which items were instances of clearly intended interpretations, while ignoring disagreements about whether responses that were not clearly intended failed to be categorized this way because they were clearly unintended, unclear intended, or unclear unintended. This resulted in a binary coding system (1 = *clear intended*, 0 = *not clear intended*). The uncollapsed condition did not collapse any of the four categories when comparing interrater reliability (1 = *clear intended*, 2 = *clear unintended*, 3 = *unclear unintended*, 4 = *unclear unintended*). Interrater reliability was calculated using the irr (version 0.84.1) package in R. Interrater reliability for uncollapsed coding of responses was moderate K = 0.759, (84% agreement). Interrater reliability for collapsed coding of responses was high K = 0.857 (95% agreement).[73]

**Study 1B Results**

***Interpretation rates.*** 36% (*n* = 55) of participants interpreted the source of disagreement in a *clear intended* way. This was significantly less than 0.5, $\chi2(1, N = 152) = 11.61$, *p* < 0.001. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 42.79%. 52% (*n* =

---

[73] I report "moderate" and "strong" following the guidelines outlined in McHugh (2012).

79) interpreted in a *clear unintended* way, and the remaining 11.8% (*n*= 18) of responses were *unclear*.

Results may be seen in **Figure 4.2**.

**Figure 4.2**

*Interpretation rates for Study 1B*
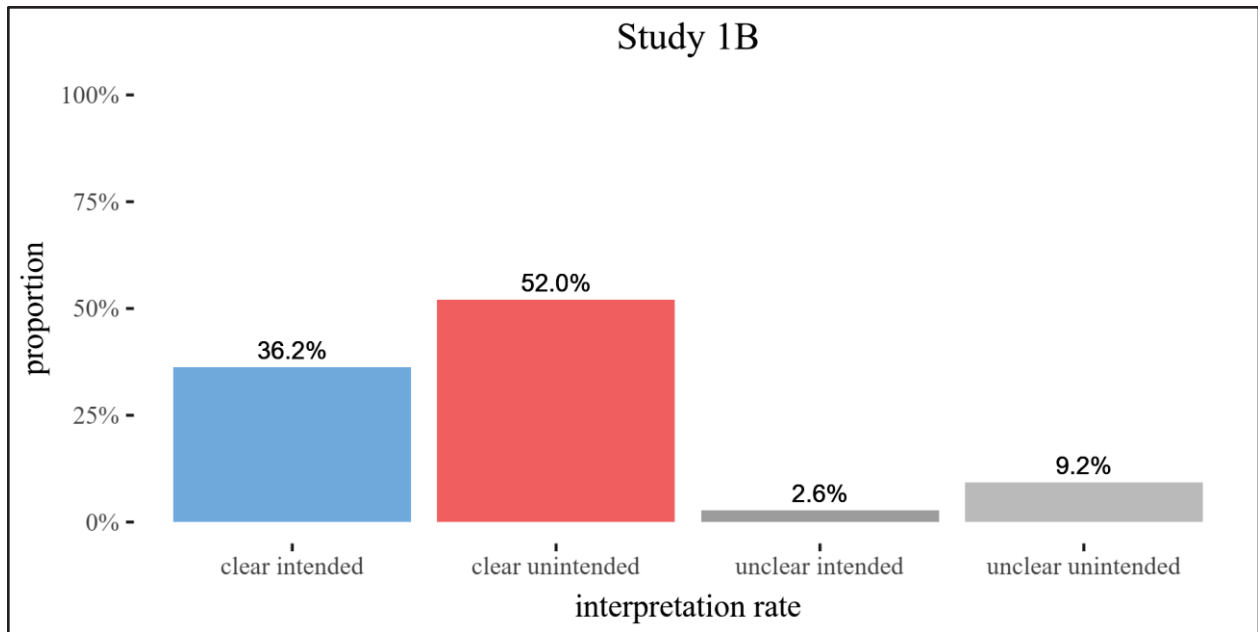


**Discussion**

The proportion of *clear intended* interpretations was very low in both studies. Both coders judged less than half of participants to clearly interpret the question as intended, and both coders judged around half of participants in both samples to have interpreted the source of disagreement in a way that was clearly unintended. Even if it was charitably assumed every participant whose response was unclear interpreted the source of disagreement as intended, this would still indicate that about as many people did not interpret the source of disagreement as intended as those who did. This is an extremely serious methodological shortcoming with the results of this particular study. This is because interpreting the source of disagreement as intended is *necessary* for the validity of the disagreement paradigm. If the participant does not regard the disagreement between themselves and another person to result from

different moral standards, then their responses to the main measures cannot tell us whether they think that there is a stance-independent moral fact about the moral issue in question. Nor does it tell us that they think there is no stance-independent moral fact. Rather, their response is effectively a response to a different question entirely. As a result, their response simply cannot tell us one way or another about their metaethical standards or commitments.

Our goal in reanalyzing Goodwin and Darley's findings was to evaluate the accuracy of their analysis, which purportedly revealed that only a small number of participants interpreted the source of disagreement in a way that would challenge their findings. In particular, they reported that about 7% of participants (7 out of 102) in Study 1A did not interpret the source of disagreement as intended. The present findings suggest that the true proportion is almost certainly much higher than this. Indeed, I can provide more than 7 examples of participants who very clearly interpreted the source of disagreement in an unintended way. In fact, I can provide *dozens* of such examples, and this would reflect what may be a *lower bound* on the proportion of participants who interpreted the source of disagreement in an unintended way, since this isn't even accounting for the unclear interpretations. This reanalysis of this data strongly suggests that Goodwin and Darley underestimated the degree to which participants attributed the source of disagreement to something other than a genuine difference in moral belief.[74]

Overall, these findings support the conclusion that a substantial proportion of participants did not interpret the source of disagreement as intended in the studies conducted by Goodwin and Darley (2008). This is consistent with the more general concern that, when participants are presented with

---

[74] I had no precise predictions about the overall proportion of people who would not interpret the source of disagreement as intended. While I expected the number to be substantial, I had no distinct hypothesis (e.g., "more than half"). The expectation (which was mine, not David's) that fewer than half of participants would interpret questions as intended arose as a result of the coding for this study, analyzing subsequent open response data, and further consideration about the implausibility of high rates of intended interpretation. This also resulted in a reanalysis of the original data in light of the greater experience I developed while analyzing other datasets.

versions of the disagreement paradigm similar to the one employed by Goodwin and Darley, that many will not interpret the source of disagreement in a way that is necessary for the disagreement paradigm to be valid. In short, these findings challenge the validity of at least some forms of the disagreement paradigm.

**4.6 Study 1C**

It's possible that the low clear intended interpretation rates in Study 1A and 1B were due to idiosyncratic features of the participants who were sampled or to the methods employed by Goodwin and Darley (2008). To address this, I assessed interpretation rates using a new version of the disagreement paradigm and found the same general results in a new sample. This sample differed from Study 1A and 1B in that I sampled participants on Amazon's Mechanical Turk, rather than sampling from a student population. In addition, the materials were presented in an abstract, simplified form with minimal instructions. The goal was to eliminate any potential sources of ambiguity introduced by the instructions employed by Goodwin and Darley, and by focusing on an abstract, unspecified moral disagreement rather than on specific, concrete moral disagreements about particular issues (e.g., abortion), I minimized the risk of normative conflations influencing participant responses. Interpretation rates for all items in Study 1C were coded exclusively by myself. In spite of these changes, the proportion of participants coded as *clear intended* was 23% ($n = 23$), which was significantly less than 0.5, $\chi^2(1, N = 100) = 29.16$, $p < 0.001$. See **Supplement 3**, section ***S4.4.1*** for full details.

**4.7 Study 2: Interpreting other people's responses to metaethics stimuli**

Another way to assess whether people interpret questions about metaethics as intended is to evaluate how they interpret other people's responses to questions about metaethics. One advantage to adopting this approach is that the participant does not have to interpret the question about metaethics directly, but is instead presented both the question and a response to that question that did interpret the

question as intended. This provides additional context that minimizes potential ambiguity or uncertainty about what the question is asking and provides a direct and explicit articulation of a metaethical stance. For comparison, suppose you asked a participant:

*Do you like to go to the bank?*

One problem with this question is that it is ambiguous. Some participants may interpret this as a question about financial institutions, while others may interpret it as a question about riverbanks. If we then asked the participant why they answered the question the way they did, and they say something like:

*I like to go to the bank because it means I've got enough money to take some out!*

If we were interested in whether people like going to riverbanks, such a response would be unhelpful, since this participant interpreted the question to refer to financial institutions rather than riverbanks. When participants respond this way, we cannot know whether they like going to riverbanks or not. Yet suppose we instead presented participants with the following:

*A respondent was asked the following question:*

*Do you like going to the bank?*

*John: "No. I don't like going to the bank. I usually have to wait in line and it's really boring. Plus, I can do most of my banking online, anyway."*

*In your own words, what do you think the respondent means in the statement above?*

It would be difficult for participants to *not* interpret John's response to refer to financial institutions rather than riverbanks. As such, we could be confident that *if* people understand what banks (understood as financial institutions are) are, and *if* they understand what it means to not like going to the bank, then they are likely to understand John's response. In short, providing another person's response can help disambiguate what a question is asking, and by focusing on the *response* to the

question, rather than the question itself, we may increase our chance to ensure that participants are in a good position to understand the concepts we are asking about.

Study 2 assesses whether people understand specific expressions of moral *realism* and moral *relativism* (a specific form of antirealism). *Realism* refers to the belief that there are stance-independent moral facts, while *relativism* refers to the claim that moral facts are only true or false relative to the standards of individuals or groups.[75] Interpretations of question and response pairings reflecting realism and relativism were each tested with two distinct question-response pairings.

One pair of questions present a direct question about realism or relativism and a person's response to the question. A second pair of questions present a version of the disagreement paradigm, and either a realist or relativist response to it. In both cases, participants were asked to explain in their own words what they think the *response* to the question means. Thus, I did not ask participants to explain what they think the question is asking, nor did I ask them how they would respond to the question. Rather, my exclusive focus was on how they interpret someone else's response to the question, with such responses intended to reflect either realism or relativism. For all four measures, I expected a majority to not interpret the question in a clearly intended way.

**Methods**

**Participants**. Participants consisted of 449 adult US residents on Amazon's Mechanical Turk (208 males, 191 females, 1 other, 53 unreported, $M_{age}$ = 34.6, $SD_{age}$ = 10.9, age range = 19-70). Since the *direct* and *disagreement* conditions are conceptually distinct, results from each will be analyzed separately.

---

[75] As always, relativizing moral standards to individuals and groups is a simplification. It is always possible for moral standards to be thought of as true or false relative to some other standard of evaluation, such as different religious systems or different species. While such forms of relativism are conceptually possible, they are probably not common in practice. These various ways moral claims can be relativized are also not necessarily mutually incompatible. While someone could think that moral claims can only be relative to individual standards, they could think that they can be relativized to the standards of both individuals *and* groups, or even individuals, groups, species, religious systems, and so on. One could, for instance, think moral standards can be true or false relative to any real or hypothetical standard of evaluation in a profligate way, that is, without consideration for what type of evaluative standard it is. I, for one, find this *more* plausible than more narrow construals of relativism.

***Procedure***. All participants were randomly assigned to one of four conditions. Two conditions represented realism, and two conditions represented relativism. In addition, two conditions presented a direct question about realism or relativism followed by a response, and two questions presented a version of the disagreement paradigm followed by a response. This resulted in four conditions:

1. *Direct realism*
2. *Direct relativism*
3. *Disagreement realism*
4. *Disagreement relativism*

After receiving instructions that informed participants that they would be given an anonymized response from a participant in a previous survey, they were then presented with a question and a response, and asked to explain in their own words what the previous respondent meant by their answer to the question. In other words, they were *not* asked to interpret the question, but the other person's *response* to the question). In all conditions participants were presented with the following instructions:

### Instructions

Recently, a diverse range of American citizens participated in a survey about moral attitudes conducted by the Pew Research Center. You will now be presented with the question and response of one of these participants. The question and respective answer will be selected at random. The name of the respondent was changed to preserve his or her identity.

Please read carefully. Afterwards, you will be asked a series of questions about the respondent. There are no right or wrong answers, please just provide your thoughtful first response.

Participants were then told that the respondent was asked to respond to a question. Participants were given that person's response and told to think about the person who answered the question. After they were presented with these instructions, participants were asked a series of questions (not reported here). Later in the study, they were asked the following question:

*In your own words, what do you think the respondent means in the statement above?*

Since this question was presented on a page that appeared after the initial instructions were presented, reminder text featuring the question and response were presented at the top of the screen. The specific

wording used for each of the four conditions (*direct realism*, *direct relativism*, *disagreement realism*, and *disagreement relativism*) is featured on **Table 4.1.** All of the answers "John" provided were written by the research team, and were not actual responses written by previous participants.

*Measures.* Interpretation rates for all items in Study 2A and Study 2B were coded exclusively by myself. All participants were presented with a text box where they could write a response. Although other stimuli were included and other measures were collected in the original study, the only measure analyzed here consists of the written responses participants provided. All responses were coded following the procedures outlined in the introduction. As such, there were two measures: *interpretation rates* and *thematic analysis*. Interpretation rates were coded by judging whether each response was clear or unclear, and whether it reflected an intended or unintended interpretation. Themes were also coded for each response.

**Results**

As expected, most participants did not clearly interpret any of the responses to questions about realism or relativism as intended. Across all conditions, the proportion of participants coded as *clear intended* was less than half: direct realism 5.4% ($n = 6$), direct relativism 12.4% ($n = 14$) disagreement realism 1.8% ($n = 2$), and disagreement relativism 25.9% ($n = 25.9\%$). The proportion of participants coded as clearly interpreted was significantly less than 0.5 for direct realism, $\chi2(1, N = 112) = 89.29$, $p < 0.001$, for direct relativism $\chi2(1, N = 113) = 63.94$, $p < 0.001$, for disagreement realism $\chi2(1, N = 112) = 104.14$, $p < 0.001$, and for disagreement relativism $\chi2(1, N = 112) = 26.04$, $p < 0.001$, 95%. With 95% confidence, the percentage of responses coded as clearly interpreted for the direct realism condition was less than 10.03%, while it was less than 18.39% for direct relativism, 5.25% for disagreement realism, and 33.21% for disagreement relativism. These results indicate that most participants did not clearly interpret the question as intended. In addition, many participants interpreted all four responses in a *clear unintended* way: direct realism 48.2% ($n = 54$), direct relativism

46.0% ($n = 52$), disagreement realism 28.6% ($n = 32$), disagreement relativism 33.9% ($n = 38$). All results are summarized in **Figure 4.4**. Overall, participants appeared less likely to provide a clear intended interpretation for items expressing realism than for relativism, and more participants offered unclear responses for the disagreement conditions than for the direct conditions.

**Table 4.1**

*Description of conditions for direct realism/relativism and disagreement realism/relativism*

| Direct realism | Direct relativism |
| --- | --- |
| The respondent was asked the following question: | The respondent was asked the following question: |
| **Do you think there is a single correct answer as to whether individual moral statements are true or false?** | **Do you think there is a single correct answer as to whether individual moral statements are true or false?** |
| **John:** | **John:** |
| **"There is one standard of moral truth by which all people should be judged".** | **"There is no single standard of moral truth. Different societies must be judged by different moral standards".** |
| Think about the person who answered this question. | Think about the person who answered this question. |

| Disagreement realism | Disagreement relativism |
| --- | --- |
| The respondent was asked the following question: | The respondent was asked the following question: |
| **When two people disagree about a moral issue, do you think they can both be correct, or must at least of them be incorrect?** | **When two people disagree about a moral issue, do you think they can both be correct, or must at least of them be incorrect?** |
| **John:** | **John:** |
| **"When people disagree about a particular moral issue there can be at most only one correct answer".** | **"When people disagree about a particular moral issue each can be correct according to their own moral standards".** |
| Think about the person who answered this question. | Think about the person who answered this question. |

**Figure 4.4**

*Interpretation rates for Study 2A (by condition)*

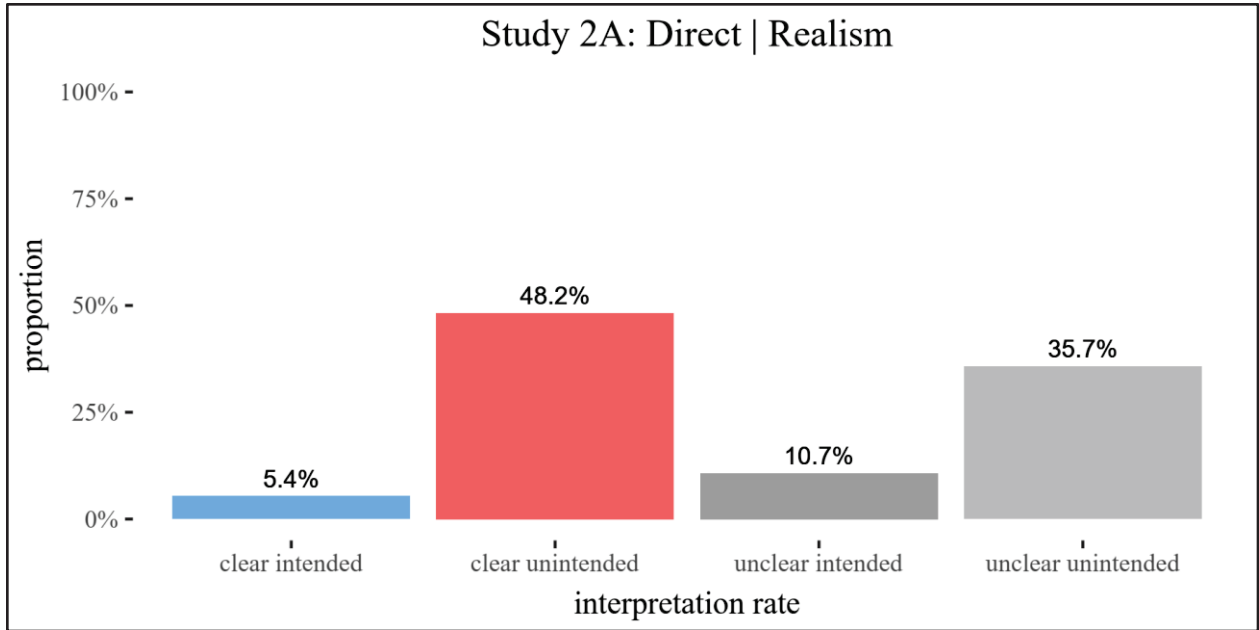*Figure 4.4.1 Study 2A: Direct | Realism*



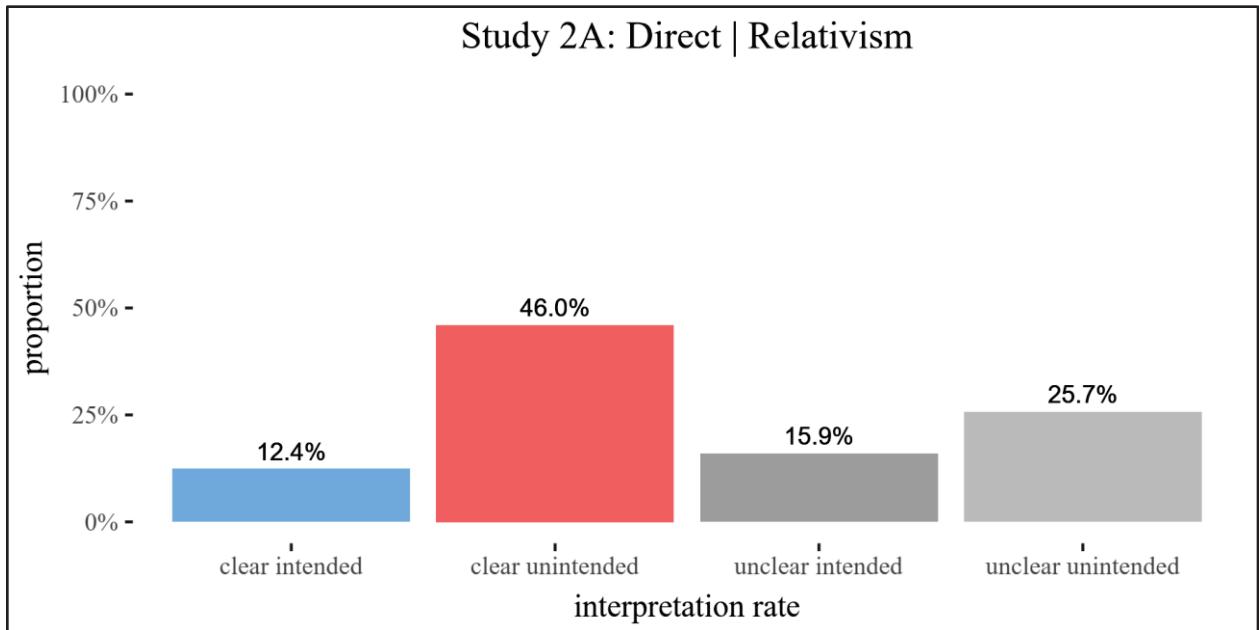*Figure 4.4.2 Study 2A: Direct | Relativism*

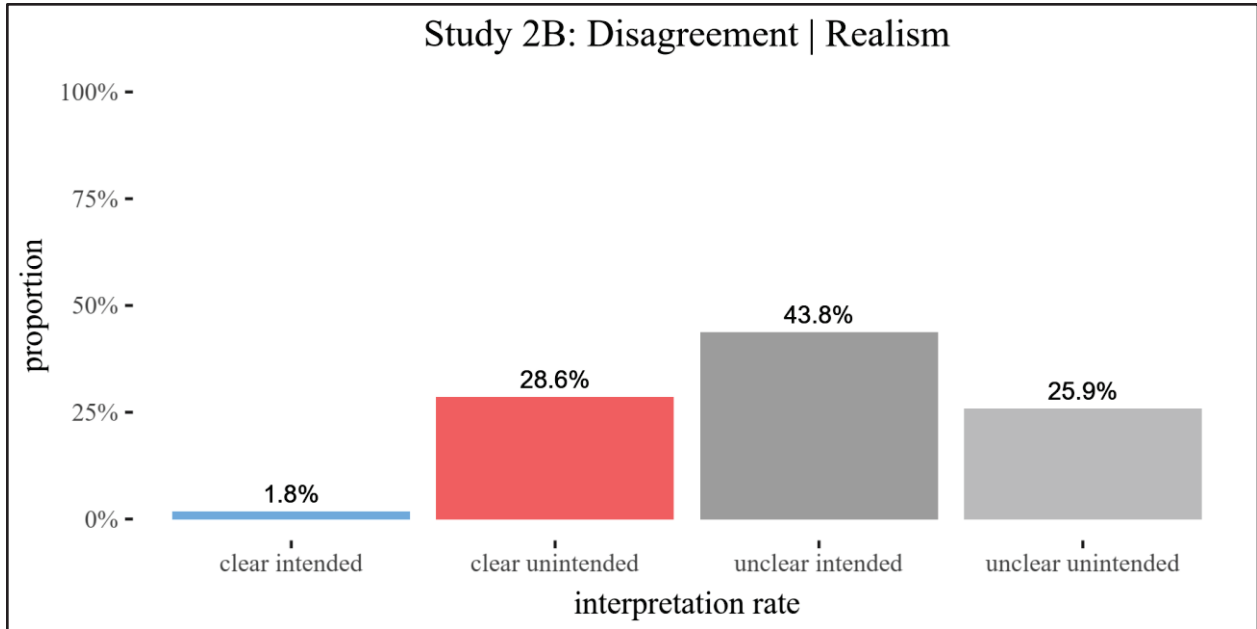*Figure 4.4.3 Study 2B: Disagreement | Realism*



*Figure 4.4.4 Study 2B: Disagreement | Relativism*
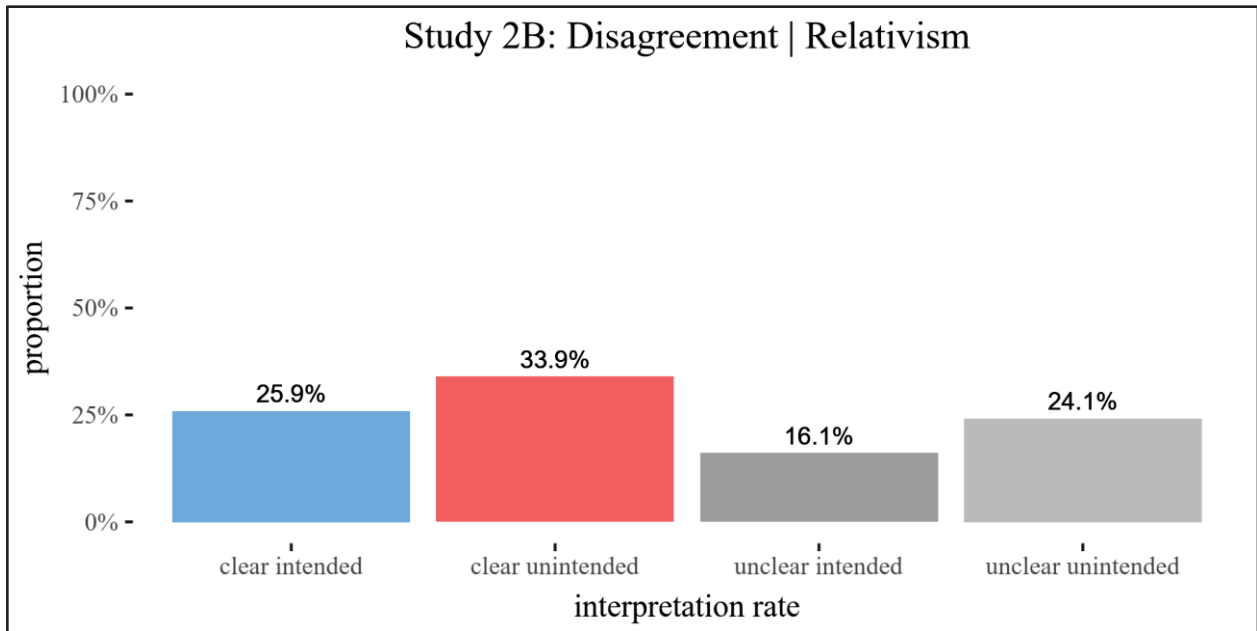
**Discussion**

Overall, clear intended interpretation rates were very low, ranging from 1.8% - 25.9%, while clear unintended interpretations were very high, ranging from 28.6% - 48.2%. Even if every unclear interpretation were an instance of an intended interpretation, it would still turn out that about a quarter to half of participants across conditions did not interpret what they were asked as intended. While there are some reasons for why this may have occurred that are the result of researcher error, there are nevertheless numerous instances in which participants clearly interpret expressions of metaethical stances in ways that have nothing to do with the intended metaethical stance, or about metaethics at all. This is not what we'd expect if we were asking questions that were easy to interpret. These findings indicate that when you present people with another person's response to a question about metaethics, many participants struggle to interpret that response as an expression of the metaethical stance it was intended to reflect.

Nevertheless, these results suffer from several serious shortcomings. First, the higher rate of clear unintended for the direct conditions may be due to at least two factors. First, the direct realist condition may be due to a methodological problem with the item, which may have inaccurately expressed *universalism* and various *normative* concerns rather than *realism*. Likewise, participants may have interpreted the first part of John's response in the *direct relativism* condition as descriptive for understandable reasons. After all, it states that "There is no single standard of moral truth." This *could* plausibly be read as a descriptive claim. In both cases, the low rates of intended interpretations may be due to misoperationalization rather than the inherent ambiguity that may be present even if realism and relativism were expressed more accurately. As such, the *direct* conditions provide at best only limited evidence. Ironically, I failed to accurately represent realism in a study that was motivated in part by criticism of other researchers for failing to operationalize realism.

There are also shortcomings with the disagreement conditions. In the realist condition, many participants restated John's response without elaborating or expanding on the remark. Such responses do not present a clear intended interpretation, but this may be due to inherent limitations in asking people what a response means. Suppose you were given someone's response, and asked, "what do you think this respondent means in the statement above?" An understandable reaction is to attempt to *restate* the position as clearly as you can, rather than *interpret* it in a way that draws out the background assumptions or intentions of the speaker. If so, you'd be complying with the formal instructions of the task in a way that is not diagnostic of whether you interpreted it as intended. After all, if John thinks that if two people disagree that at least one person must be mistaken, to respond by saying that he thinks that there's only one correct answer is *completely accurate*, yet it tells us nothing about whether you understood the remark to convey a realist stance towards the issue. Strangely, the *repeat* theme was not very common in the relativist condition.

One way of circumventing these interpretive difficulties would be to present participants with a variety of specific and directed questions that provide a more comprehensive and focused way to probe participants for how they interpreted a statement. For instance, you could ask participants to match a statement to one or more statements that most closely reflect what it means or ask participants how much they agree or disagree that someone who made that statement meant to convey a variety of claims, or is likely or unlikely to have particular beliefs or attitudes.

There are two more shortcomings with these conditions. One is that they only describe morality in abstract terms, rather than referring to any specific moral issues. One rationale for considering specific, concrete moral issues is that the added context may facilitate enhanced comprehension. After all, people are probably more familiar with judging specific moral issues rather than thinking about morality as a whole. A second shortcoming is that the preceding studies have, so

far, only considered realism and relativism. However, relativism represents only one form of antirealism, which excludes both error theory and noncognitivism.

Study 3 addresses these shortcomings by presenting metaethical statements in a concrete context and including a noncognitivism condition. In addition, I used statements that more accurately convey the central distinction between realism and relativism in a way that does not conflate realism with universalism or normative claims.

**4.8 Study 3: Charity & noncognitivism**

This study assesses interpretation rates in the context of a study that was otherwise focused on assessing attitudes about charitable giving. My goal was to assess interpretation with respect to a concrete moral issue. I used several other measures that were not open response questions, which I discuss in **Chapter 5**. Discussion here will only concern the three open response questions presented to participants towards the end of the study.

The primary rationale for employing questions in a charitable context was to assess interpretation rates with respect to a concrete moral issue, to include an item that reflected noncognitivism (in addition to items reflecting realism and relativism), and to present these statements in a simple, alternative format that may prompt a higher rate of intended interpretations and, if it did not, would show that changes in how one conveys metaethical positions does not substantially alter the low rate of clear intended interpretations.

There is considerable precedent for asking questions about noncognitivism. As Beebe (2015) notes, early formulations of the disagreement paradigm did not include a response option for noncognitivism. As a result, early studies on folk metaethics required participants to choose between response options that presupposed cognitivism, even though this may not reflect how ordinary people think about specific moral issues or morality in general. Yet when Beebe included a noncognitivist option alongside the standard realist and relativist responses to the disagreement paradigm, many

participants opted for this response. It was even the most common response for some moral issues! More recently, Davis (2021) found that noncognitivism was the most common response using a novel, more sophisticated version of the disagreement paradigm, averaging 34.2% of responses across the five subdomains of the Moral Foundations Questionnaire (p. 17; Graham et al., 2011), while Pölzler and Wright (2020a, 2020b) found that noncognitivist responses were common across a variety of novel paradigms, including 23% of participants in an abstract version of the disagreement paradigm, as well as 30% of respondents in their "theory" task, 3% in their comparison task, and anywhere from 9%-35% across a variety of concrete moral issues when presented with a truth-aptness task.

While these rates do not comprise a majority of respondents, they show that noncognitivist response options are often favored by a substantial minority of participants. As such, there is little justification for excluding them using a forced choice paradigm that presupposes cognitivism. In this study, I added an additional noncognitivism condition to assess whether participants interpreted a claim about noncognitivism as intended. I included additional realism and relativism items as well, in order to assess whether different ways of conveying realism and relativism would result in greater clear intended interpretation rates.

**Methods**

*Participants.* Participants consisted of 367 adult US residents on Amazon's Mechanical Turk (166 males, 134 females, 1 other, 66 unreported, $M_{age}$ = 37.4, $SD_{age}$ = 11.7, range 19-73).[76]

*Procedure.* All participants participated in all conditions. Participants were given a brief set of instructions describing the purpose of the study:

> **Instructions**
>
> In this study you will be asked a series of questions about charities. Please read the questions carefully and answer using the options provided. There are no right or wrong answers, please just provide your thoughtful first response.

---

[76] One participant reported an age of 5.

Participants were then asked to respond to a series of questions about their level of agreement with a range of statements reflecting metaethical stances towards charitable giving. Participants were presented with a comprehension check to assess their understanding of the notion of something being "objectively true." Next, all participants were presented with the three open response questions analyzed here, in order of realism, relativism, and then noncognitivism. The three questions may be seen in **Table 4.2**. Order was not varied across participants. Finally, demographic data was collected.

**Table 4.2**

*Wording for conditions in Study 3*

| Condition | Wording |
|---|---|
| Realism | In your own words, what does it mean to say that "it is a fact that some charities do more good than others, not a matter of personal beliefs or values"? |
| Relativism | In your own words, what does it mean to say that "the truth about which charities do the most good depends on the beliefs and values of each individual"? |
| Noncognitivism | In your own words, what does it mean to say that "there is no fact of the matter about which charities do the most good"? |

*Measures.* Interpretation rates for all items in Study 3A, Study 3B, and Study 3C were coded exclusively by myself. Measures consisted of written responses to the three open response questions. All responses were coded for interpretation rates and themes.

**Results**

As expected, most participants did not clearly interpret any of the responses to questions about realism, relativism, or noncognitivism as intended. 9.0% ($n = 33$) provided a clear intended response in the realism condition, 11.4% ($n = 42$) provided a clear intended response in the relativism condition, and 0.5% ($n = 2$) in the noncognitivism provided a clear intended response in the noncognitivism

condition. The proportion of participants coded as clearly interpreted was significantly less than 0.5 in the realism condition, $\chi^2(1, N = 367) = 221.32$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 14.17%. It was also significantly less than 0.5 in the relativism condition, $\chi^2(1, N = 367) = 218.23$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 14.46%. Finally, it was significantly less than 0.5 in the noncognitivism condition $\chi^2(1, N = 367) = 359.04$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 1.63%. 42.8% ($n = 157$) expressed a clear unintended response in the realism condition, 43.1% ($n = 158$) provided a clear unintended response in the relativism condition, and 41.7% ($n = 153$) in the noncognitivist position. Results can be seen in **Figure 4.5**. Overall, these findings support the hypothesis that ordinary people did not understand questions about metaethics as intended.

**Figure 4.5**

*Interpretation rate for Study 3 (by condition)*

*Figure 4.5.1 Study 3A: Realism*
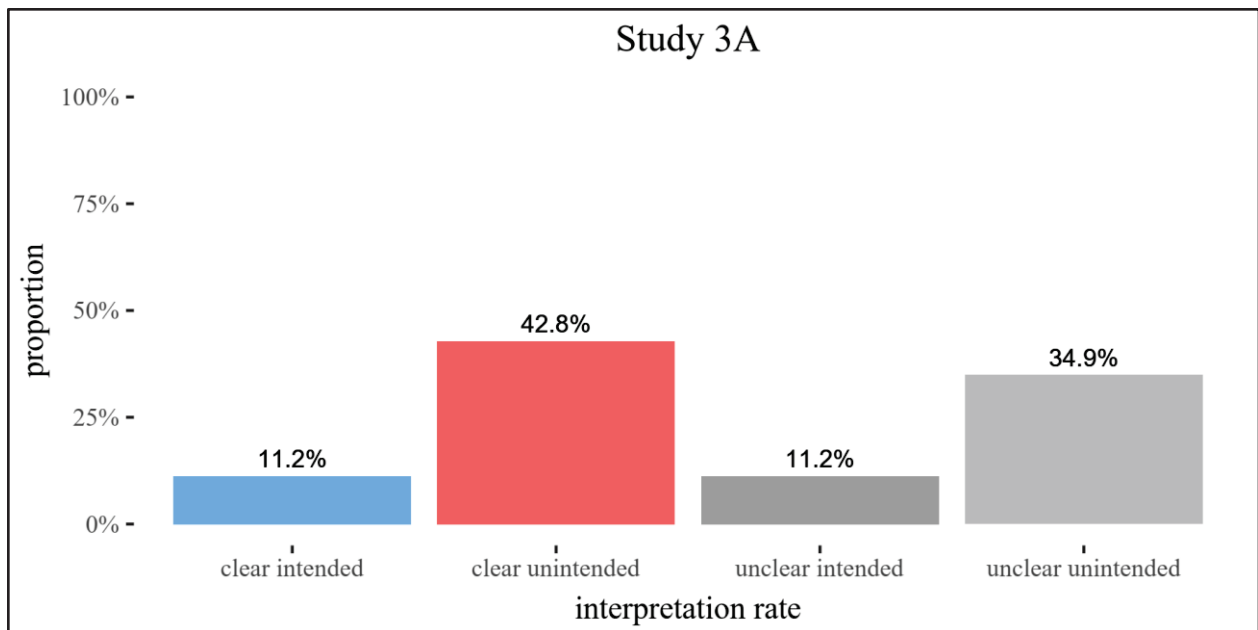
*Figure 4.5.2 Study 3B: Relativism*
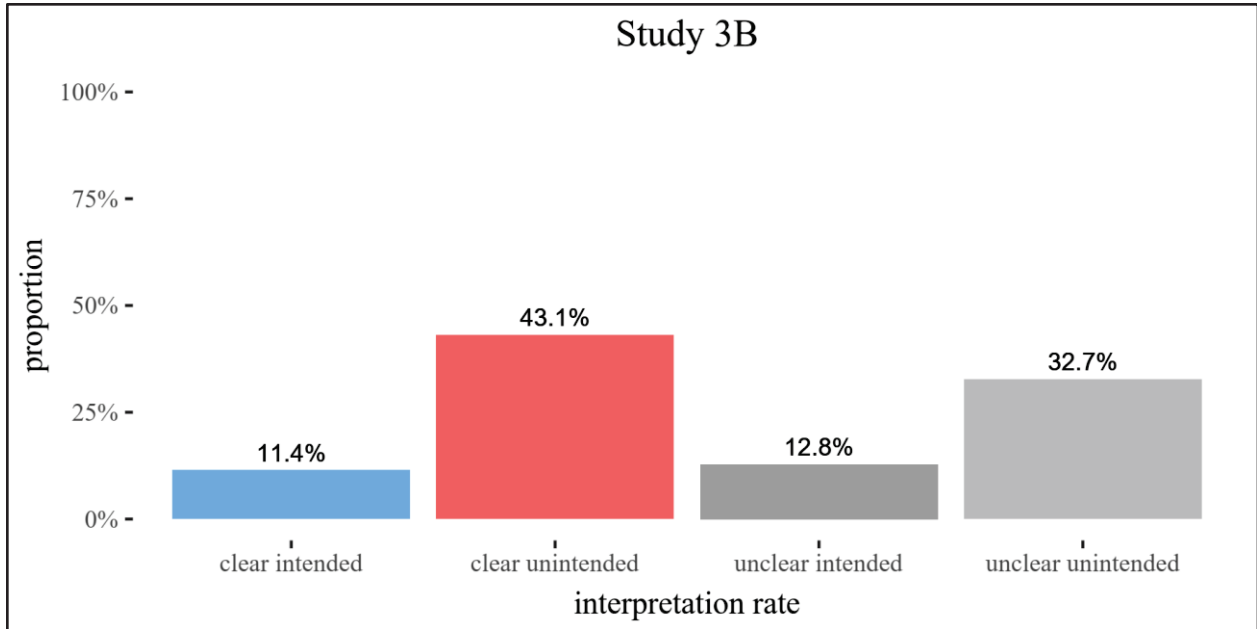


*Figure 4.5.3 Study 3C: Noncognitivism*
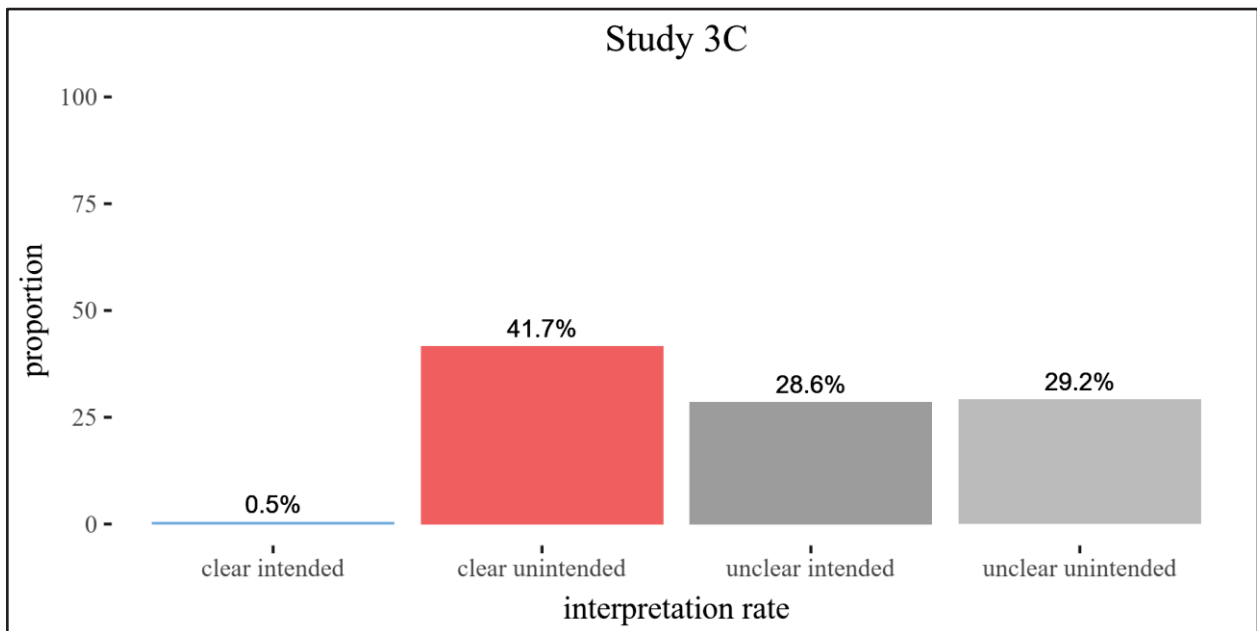
**Discussion**

Clear intended interpretation rates were very low, while clear unintended interpretations very common. In all three conditions, more than 40% of participants exhibited clear unintended interpretations. Even if every participant coded as having an unclear interpretation interpreted questions as intended, these findings still support the conclusion that most people struggle to interpret questions about metaethics as intended. This provides additional support for the conclusion that ordinary people struggle to understand questions about metaethics as intended. In addition, these findings are consistent with the possibility that most ordinary people have no determinate metaethical stances or commitments, since this could explain why people. Nevertheless, there are significant limitations with these findings.

**4.9 Study 4: Direct realism and relativism**

One way to assess whether ordinary people are moral realists or antirealists is to directly ask them, using clear and explicit metaethical language. Fisher et al. (2017) take this approach in their third experiment (pp. 1127-1128). Fisher and colleagues report a series of studies which purportedly demonstrate that people who engage in cooperative interactions with other people were less likely to endorse moral realism than participants who engaged in competitive interactions with other people. They found that participants assigned to the cooperative condition had lower mean realism scores. This would indicate that engaging in a cooperative interaction caused participants to be less likely to endorse moral realism. However, Fisher et al. recognize that participants may not have interpreted the disagreement paradigm as intended, so they designed a study with a different measure of metaethics that simply directly asked participants whether, for a series of concrete moral actions, there was an objectively true answer as to whether those actions should be allowed.

This is a laudable step in the right direction, since they explicitly acknowledge the possibility that questions about metaethics may be susceptible to unintended interpretations that could threaten

the validity of the studies in question. The primary dependent variable used in their previous studies was a version of the disagreement paradigm adapted from Sarkissian et al. (2011). Participants were told: "Earlier studies show that people take opposite positions on the issue of [issue] Given that people have opposite views, at least one side must be wrong" (p. 1123). Participants were asked to express their level of agreement (1 = Strongly disagree, 7 = Strongly agree). Fisher and colleagues note that people may have agreed with the statement, not because they are relativists, but because there could be multiple ways of conforming to the same objective moral standards. As they put it:

> [I]t may not be clear that we are asking about the metaphysical issue as to whether there is an objective truth about the question under discussion. Instead, participants might interpret the measure as simply asking whether this is a question for which there is only one correct answer versus multiple different correct answers. (For example, if the question had been "Please name a prime number between 1 and 10," there would be multiple different correct answers, but any given answer would still be objectively true or false.) (p. 1127)

Their solution is to assess whether they find similar results using a *direct paradigm. Direct paradigms* are any paradigms that explicitly ask participants about their metaethical views using metaethical language, e.g., "Do you think there are objective moral facts?" or asking people whether they agree or disagree that "moral realism is true." Most studies appropriately avoid questions like these because their validity depends on participants understanding technical philosophy terms in a way consistent with how philosophers understand these terms.[77] In this particular instance, Fisher et al. ask participants the following question:

> *"Consider the following question: 'Should [topic] be allowed?' Please tell us whether you think there is an objectively true answer to this question."*
>
> *[1 = definitely no objective truth, 7 = definitely an objective truth]* [p. 1127]

---

[77] Arguably, these studies also rely on the questionable assumption that philosophers themselves agree on the meaning of these terms. They don't. However, there may be enough overlap that this would render such questions at best only a bit noisy, not necessarily meaningless or invalid

Note the explicit use of the term "objectively true." Whether this is a valid measure of folk metaethical belief turns on whether participants understand "objectively true" in the intended respect, which Fisher et al. make quite clear: a *metaethical* respect regarding the *metaphysical status* of moral truth. At the start of the paper, they also offer some account of what they mean by *objective*: "There is a correct answer made true by the features of the topic, and any other answer must be wrong" (p. 1120). This is not an ideal characterization of objectivism, but it is clear enough that they are referring to *moral realism*. Thus, whether their measure is valid would turn on whether people understand "objective truths" about morality to refer to stance-independent normative facts about what is morally right or wrong.

Fisher and colleagues ran their version of the *direct paradigm* and reported that they replicated their previous findings. Specifically, they found that the mean level of agreement that the moral issues had "objectively true" answers was significantly lower in the cooperative condition than in the interaction condition. Thus, they found a significant effect in the same direction as their earlier studies, with participants assigned to the cooperative condition expressing reduced endorsement of moral realism. Given that this study found a significant mean difference in the same direct predicted by their previous studies, Fisher and colleagues concluded that participants interpreted their earlier measures as intended, which resolves any concerns that participants did not interpret the disagreement paradigm as intended. Specifically, they state that their result "replicates our main finding from the previous experiments and suggests that participants were correctly interpreting the original measure" (p. 1128).

There are several serious problems with this claim. First, note that they found a statistically significant difference in the same direction across a handful of studies. But it does not follow that because several studies had a statistically significant result in the same direction that *therefore* those studies must be measuring the same thing, nor does it follow that they are measuring the intended psychological construct. For comparison, imagine if I ran two studies. In one study, I asked if people

prefer A over B. In a second study, I ask if people prefer X over Y. If I predict that people would prefer A in the first study, and X in the second study, does it follow that therefore A and X are measures of the same thing? No. While two measures yielding similar results in line with one's predictions provides some minimal corroborating evidence that both measures are valid, such evidence cannot be decisive, nor even very strong. The problem is that, in the absence of extraneous reasons to consider the measures valid, it is possible that two measures that reliably yield similar results are measuring the same phenomenon, or two closely correlated phenomenon, but that the phenomenon they're measuring isn't the phenomenon researchers are intending to measure. In short, two measures cannot pull one another up by their mutual bootstraps alone. To illustrate why, suppose you ran the following two studies:

---

**Study 1**

---

| **Condition A** | **Condition B** |
|---|---|
| Please rate your agreement with the follow statement: | Please rate your agreement with the follow statement: |
| *It is usually good to be nice to people.*<br>[1 = Strongly disagree, 7 = Strongly agree] | *It is usually good to be mean to people.*<br>[1 = Strongly disagree, 7 = Strongly agree] |

**Study 2**

---

| **Condition A** | **Condition B** |
|---|---|
| Please rate your agreement with the follow statement: | Please rate your agreement with the follow statement: |
| *Pizza tastes good.*<br>[1 = Strongly disagree, 7 = Strongly agree] | *Dirt tastes good.*<br>[1 = Strongly disagree, 7 = Strongly agree] |

Imagine you ran both of these studies. What would you predict about the mean scores for level of agreement? Chances are you'd expect mean agreement in **Condition A** in both studies to be significantly higher than mean agreement in **Condition B**. Yet these studies clearly measure something different. Even if the size of the difference in both cases were about the same, this wouldn't be good

evidence that these questions all measure the same psychological construct. Just the same, the fact that Fisher et al. (2017) show that two studies yield similar results *does not* demonstrate that those studies were measuring the same thing.

Furthermore, even if those studies were measuring the same thing, it does not follow that because the studies yielded similar results that they were measuring the *intended* construct. It could simply be that people interpreted the disagreement paradigm in unintended ways, *and* that they interpreted the *direct paradigm* in unintended ways. And since the reasons *why* people do not interpret questions about metaethics as intended is *not random*, both studies could simply recapitulate the same background problem of unintended interpretation. Fisher et al. seem to have made a catastrophic error in reasoning: they forgot that two wrongs don't make a right.

Setting these methodological shortcomings aside, which suggest that Fisher and colleagues are very far from having demonstrated the validity of their measures, we can simply assess whether participants interpret claims about moral objectivity in a way consistent with Fisher et al. (2017). Such findings would demonstrate not only that one purported demonstration of the validity of metaethics paradigms is mistaken, but that direct paradigms are likely going to be uniformly invalid. That is, if ordinary people reliable interpret references to morality or moral claims being "objectively" true in ways that are inconsistent with moral realism, this would suggest that directly asking people about moral realism using terms like "objective" or "objectively" isn't a valid way of measuring folk metaethical stances or commitments.

In the following study, I sought to refute Fisher et al.'s claims by testing whether ordinary people understand the term "objective" with respect to claims about moral truth in a way consistent with their studies and more generally in a way consistent with moral realism. I tested this by testing the hypothesis that a majority of people would not provide a clear intended interpretation of what it means for the truth of a moral claim to be objective. I also tested whether people would understand

the claim that moral truth is "relative," as well, to head off direct paradigms that directly ask about moral relativism, as well. I was also curious whether participants might respond differently to different moral issues, so for exploratory purposes I included a pair of concrete moral issues. This resulted in six conditions: two abstract conditions which asked whether the truth of a moral claim was "objective" or "relative" in the abstract, without specifying any particular moral issues, and four additional conditions, where I varied both whether I asked about whether the truth of a moral claim was "objective "or "relative," with respect to the either the claim that "murder is morally wrong," or that "abortion is morally wrong." Also note an advantage over the previous study, where I did not explicitly use the term "moral." By using "moral" in these studies, it would be less likely for participants to not interpret the issues in question as moral issues. This is especially important for the "abortion" condition, since Wright, Grandjean, and McWhite (2013) found that only 51% of participants classified "1st trimester abortion" as a moral issue (7% categorized it as a social issue, and 41% classified it as a personal issue).

**Methods**

**Participants.** Participants consisted of 300 adult US residents on Amazon's Mechanical Turk. 459 participants initially participated in the study. However, 159 participants did not complete the study and were dropped from analysis.[78]

**Procedure.** Participants were asked a single open response question and presented with a text box. Across all conditions participants were asked what it means to say that something is "objective" or "relative" in their own words. Wording for all conditions may be seen in **Table 4.3**. No other data was collected.

---

[78] This very high rate of noncompletion is likely due to participants being asked to complete an open response question with no other stimuli. My intent was for participants to write only a short response which should take less than a minute, but the size of the text box and background assumptions about expectations of the task may have led participants to believe I required more significant input. This may undermine data quality.

**Table 4.3**

*Wording for stimuli in Study 4 (by condition)*

| Condition | Wording |
| --- | --- |
| Abstract \| Realism | In your own words, what does it mean to say that moral truth is objective? |
| Abstract \| Relativism | In your own words, what does it mean to say that moral truth is relative? |
| Concrete \| Realism \| Murder | In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is objective? |
| Concrete \| Relativism \| Murder | In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is relative? |
| Concrete \| Realism \| Abortion | In your own words, what does it mean to say that the truth of the moral claim "abortion is morally wrong" is objective? |
| Concrete \| Relativism \| Abortion | In your own words, what does it mean to say that the truth of the moral claim "abortion is morally wrong" is relative? |

***Measures.*** Interpretation rates for all items in Study 4A and Study 4B were coded exclusively by myself. Interpretation rates and thematic analysis were coded for all responses. A clear intended interpretation for realist conditions would consist of understanding the notion that moral truth is "objective" in a way consistent with moral realism, i.e., that there are stance-independent moral facts. A clear intended interpretation of the relativist conditions would entail recognizing the notion that moral claims can be true or false, but only relative to the standards of different evaluative standards (such as those of different people or cultures).

**Results**

Across all concrete conditions, a majority of participants did not provide a clear intended interpretation: abstract|realism 14% ($n$ = 8), abstract|relativism 18.2% ($n$ = 8), concrete|murder|realism 18% ($n$ = 9), concrete|murder|relativism 6.1% ($n$ = 3), concrete|abortion|realism 15.8% ($n$ = 9), concrete|abortion|relativism 2.3% ($n$ = 1). Aggregating

across conditions, 11.2% ($n = 22$) of participants interpreted the meaning of objective or relative as intended. In the abstract|realism condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 57) = 29.49$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 23.24%. In the abstract|relativism condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 44) = 17.82$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 29.49%. In the concrete|murder|realism condition, condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 50) = 20.48$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 28.50%. In the concrete|murder|relativism condition, condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 49) = 37.74$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 14.36%. In the concrete|abortion|realism condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 57) = 26.68$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 25.26%. In the concrete|abortion|relativism condition, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 43) = 39.09$ $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 9.77%. A majority of participants provided a clear unintended interpretation for all four conditions: murder|realism 60% ($n = 30$), murder|relativism 65.3% ($n = 32$), abortion|realism 63.2% ($n = 36$), and abortion|relativism 74.4% ($n = 32$). Aggregating across conditions, 65.3% ($n = 130$) of participants interpreted the meaning of objective or relative as intended. These results indicate that a majority of participants did not interpret any of these items as intended, and a substantial number clearly interpreted them in unintended ways. Results are summarized in **Figure 4.6.**

**Figure 4.6**

*Interpretation rates for study 5 (by condition)*

*Figure 4.6.1 Study 4A: Abstract | Realism*
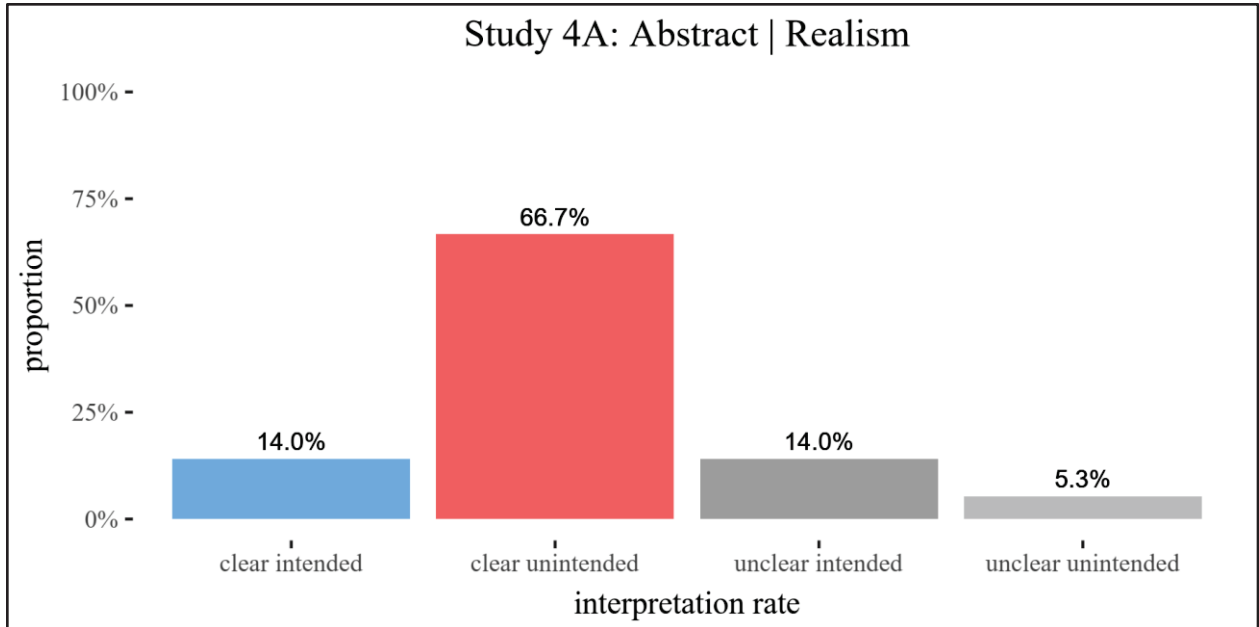


*Figure 4.6.2 Study 4A: Abstract | Relativism*
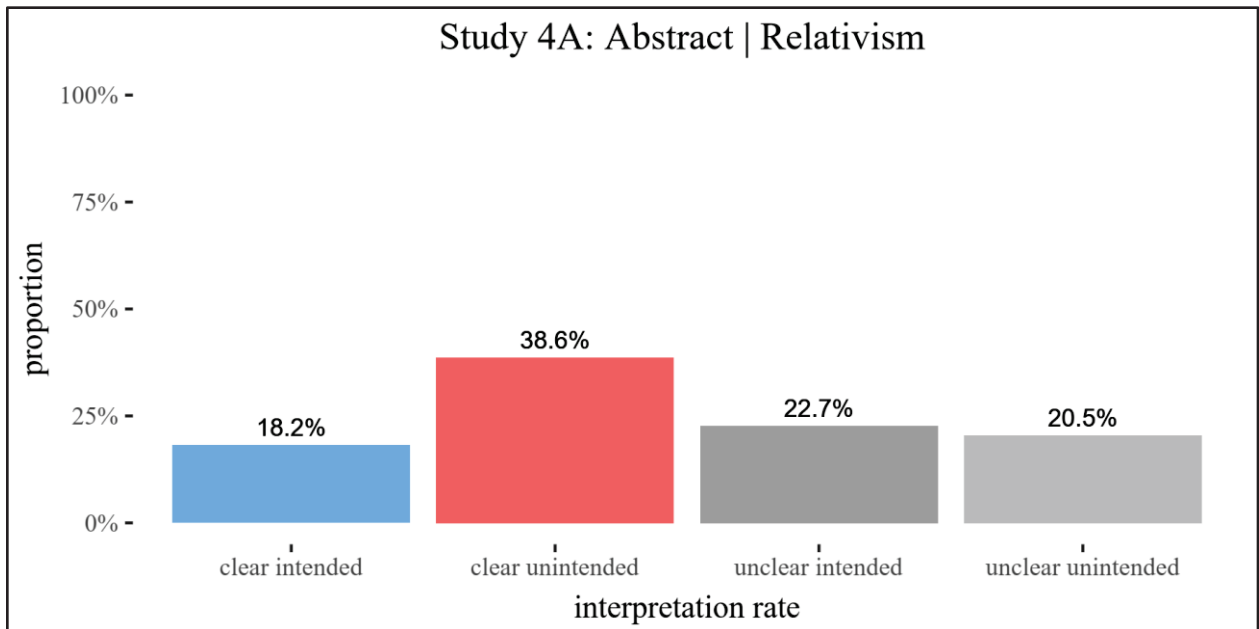
*Figure 4.6.3 Study 4B: Concrete | Realism | Murder*
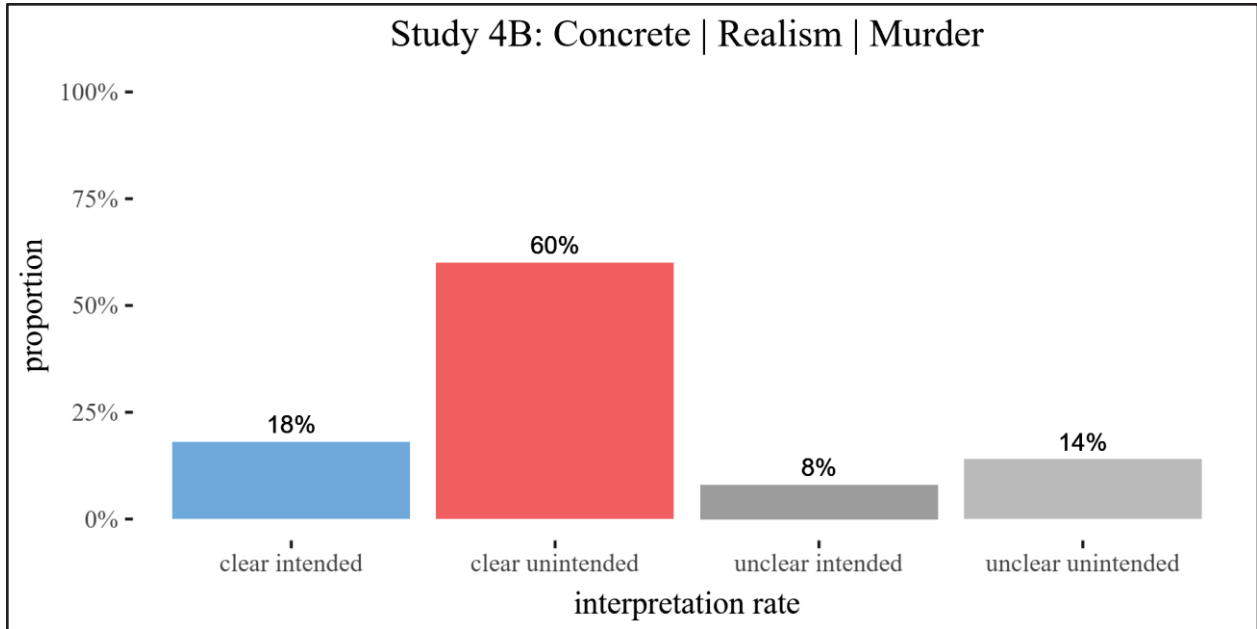


*Figure 4.6.4 Study 4B: Concrete | Realism | Murder*
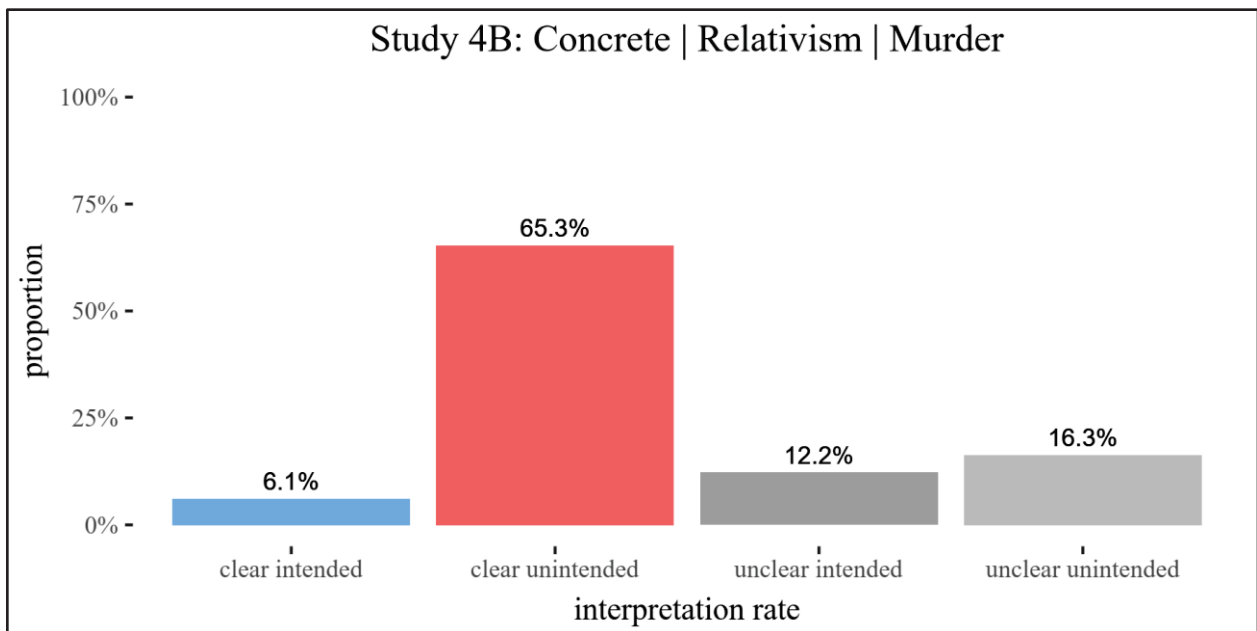
*Figure 4.6.5 Study 4B: Concrete | Relativism | Murder*
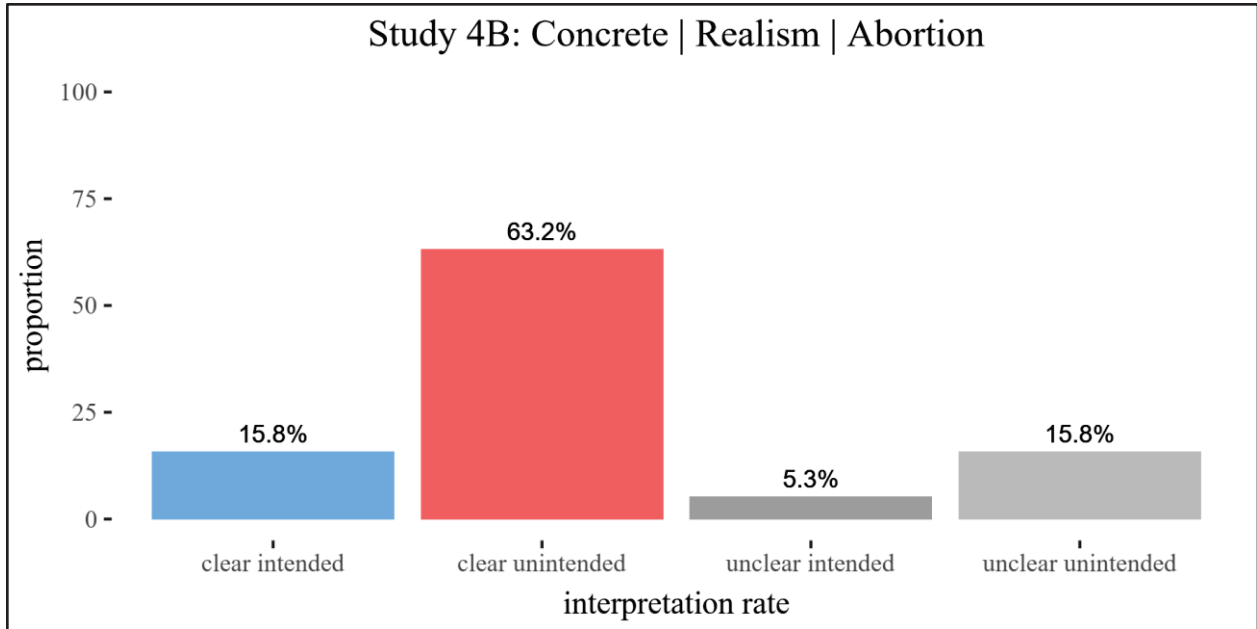


*Figure 4.6.6 Study 4B: Concrete | Relativism | Abortion*
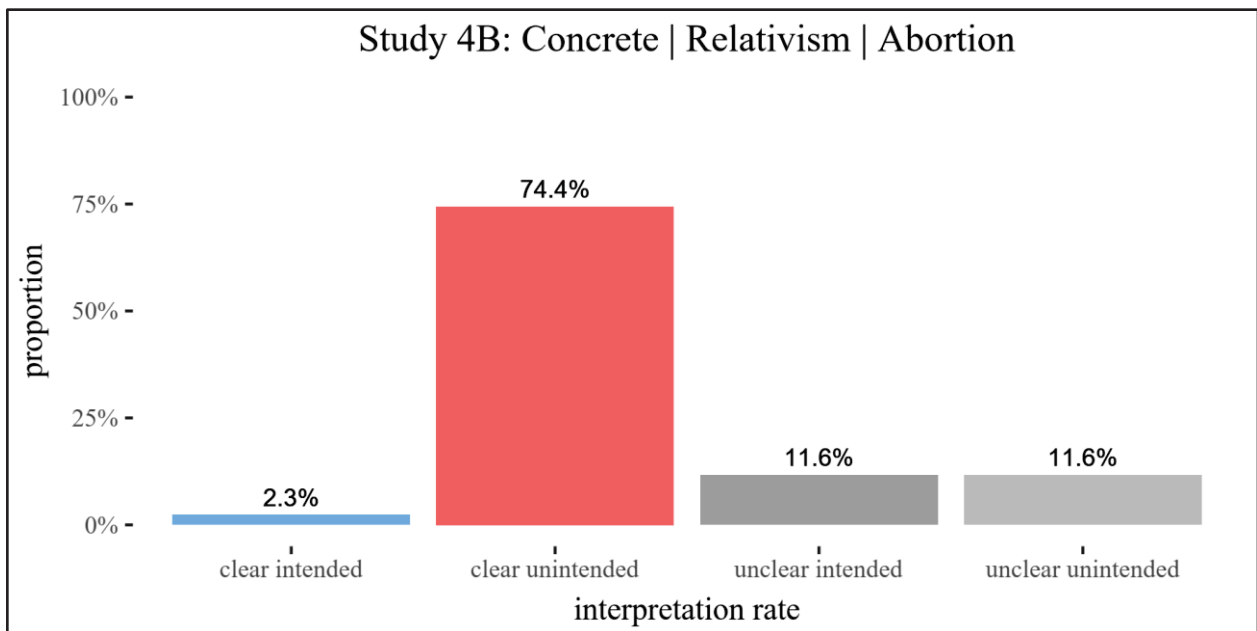
**Discussion**

Across all conditions, the proportion of participants who interpreted direct questions about what it means for moral truth to be objective or relative exhibited incredibly low clear intended interpretation rates. The proportion of clear intended interpretation rates ranged from 2.3% - 18.0%. Conversely, clear unintended interpretation rates were extraordinarily high, even compared to analysis of other open response questions. Clear unintended interpretation rates ranged from 65.3% - 74.4%. Thematic analysis corroborates these findings, lending support to the low rate of clear intended interpretations and high rate of clear unintended interpretations being attributable to a coherent and plausible pattern of specific unintended interpretations. For instance, many participants appeared to misread the term "objective" to mean "relative" or "subjective," while others conflated realism with absolutism or a rigid view of morality (as indicated by the *black and white* theme), or being biased. The most common themes across all three relativism conditions likewise support precisely those conflations I have identified as plausible candidates for ways in which ordinary people conflate relativism with other considerations: in all three conditions, the *descriptive* and *context* themes were in the top three most common themes. This suggests that participants are not merely responding *randomly*, completely clueless as to what I was asking them. Rather, it points to a reliable, recurring pattern of active conflation between relativism and other concepts. This is precisely what we should expect if people do not understand direct questions about relativism.

**4.10 Study 5: Interpretation of the Moral Relativism Scale (MRS)**

Collier-Spruel et al. (2019) constructed the Moral Relativism Scale (MRS) to, not surprisingly, measure beliefs about moral relativism. The MRS consists of ten items intended to reflect moral relativism or the negation of moral relativism. A participant's overall score is interpreted as the degree to which they endorse moral relativism. The MRS has many virtues. The paper which presents the scale documents that the scale was validated over the course of nine studies with over 3,200 participants. It

was measured alongside numerous other constructs with an eye towards demonstrating its predictive validity. The final pool of items loaded well onto a single factor ($\lambda \geq 0.60$), and any items with a 10th grade reading level or higher were dropped to ensure wording was not too complicated for some readers. In addition, the items were evaluated by a panel of 11 experts in moral philosophy and psychology and only items that passed muster were retained.[79] Sure enough, the items exhibit far greater face validity than previous scale items. The efforts to develop this scale are a remarkable piece of work and could serve as a model for many of the steps critical to devising a high quality scale.

In spite of its virtues, I am skeptical that the items on this scale are interpreted as intended often enough for it to serve as a valid measure of folk relativism (see **Supplement 3**). Here, my only interest is in evaluating whether participants interpreted items on this scale as intended. There are several differences between the approach taken here and the approach taken in previous studies where I collected my own data (as opposed to Study 1, where I reanalyzed existing data). First, I used, the closest approximation, the precise wording of the questions used in Collier-Spruel et al., in order to assess how participants interpreted precisely those questions actually used in the scale. Second, I employed two measures: both a request to explain *why* they answered the question as they did, and to explain what they believe the items used in the scale mean. This provided two separate measures of interpretation rate.

Finally, Collier-Spruel et al. (2019) developed two scales, a moral relativism scale, and a moral tolerance scale, in order to evaluate the relationship between relativism and tolerance. I collected the same type of data for the 10-item tolerance scale as well. My goal was to analyze interpretation rates

---

[79] I do wonder whether the people described were, in fact, experts. We are told that the expert panel included professors, postdocs, and graduate students who studied philosophy and psychology and that are all "researchers of morality" (p. 4). This does sound like a promising panel. Yet as I have demonstrated *ad nauseum* throughout this dissertation, metaethics is *difficult*. Expertise in moral philosophy and psychology does not necessarily confer expert judgment about the degree to which a given item plausibly reflects relativism. In my experience, even professional philosophers conflate or misconstrue relativism. I fear that even researchers more competent than I am lack the requisite degree of discernment, borne not of general knowledge but of direct and constant experience with considering the validity of metaethics measures to judge the quality of an item.

for the tolerance items. However, after an initial perusal and a second look, it became apparent that virtually no participants appeared to interpret questions about tolerance in unintended ways, nor to explain what the tolerance items mean in unintended ways. Perhaps it would be rhetorically persuasive to analyze and compare the results of the tolerance items, but I am not sure this would be appropriate. While some responses were unclear, there were almost no straightforward instances in which people appeared to interpret questions about tolerating other people in unintended ways. Yet because I am not as experienced or knowledgeable about the concept of "tolerance," I am not sure that I have the experience necessary to evaluate such responses. This suggests to me that proper open response analysis may require deep immersion in the academic literature and considerable experience analyzing open responses. Since I lack both types of experience, I opted not to analyze the results. However, results are available to anyone who wishes to have a look, and I suspect (or, rather, *expect*) that most people will come away with the impression that participants have no problem interpreting questions about tolerance. This serves, in an indirect way, to reinforce the validity of the methods used here. It could have turned out that open response questions provide such ubiquitously low quality indications of the degree to which participants interpreted questions as intended that there would be no distinctive shortcoming in how people responded to questions about metaethics. But this brief foray into how people interpret other questions provides hints that this is probably not the case. The primary prediction was that fewer than half of participants across all conditions and in aggregate would interpret questions about relativism as expected, and that this would be reflected both in their explanations for why they answered as they did, and their explanations for what they interpreted each item on the MRS to mean.

**Methods**

**Participants.** Participants consisted of 156 adult US residents on Amazon's Mechanical Turk. The complete study included items from both the moral relativism subscale (MRS) and the moral tolerance

scale (MTS) consisted of 353 participants. However, only the results of participants in MRS conditions were analyzed. No demographic data was collected.

**Procedure.** All participants were asked to rate their level of agreement with an item on the MRS on a 5-point scale. Wording was as follows:

Please rate how much you agree or disagree with the following statement.

**One's own culture determines whether that person's actions are "right" or "wrong."**

○ (1) strongly disagree ○ (2) Disagree ○ (3) Neither agree nor disagree ○ (4) Agree ○ (5) Strongly agree

Next, they were asked to explain why they chose the response they did to the multiple choice question, and were then asked to explain what the item they were asked means. In both cases they were given a text box. Order of the two open response questions was not varied. The two questions appeared as follows:

Please explain why you chose this response.

```



```

---{page break}---

In your own words, please explain what this statement means:

**[Item] [Example: One's own culture determines whether that person's actions are "right" or "wrong."]**

```



```

**Measures.** Interpretation rates for all items in Study 5A and Study 5B were coded exclusively by myself. Measures consisted of level of agreement with the item (1 = strongly disagree, 5 = strongly

agree), and written responses to the two open response questions. Results of the two measures are analyzed separately.

### 4.10.1 Study 5A: Explanation

**Results**

As expected, most participants did not interpret items in a clear intended way when summing all conditions. Overall, 28.2% ($n = 44$) of participants provided a clear intended interpretation. This was significantly less than 0.5, $\chi 2(1, N = 156) = 35.10$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 32.45%.[80] Results are summarized in **Figure 4.6.**

**Figure 4.6**

*Interpretation rates for Study 5A: Explanation*



---
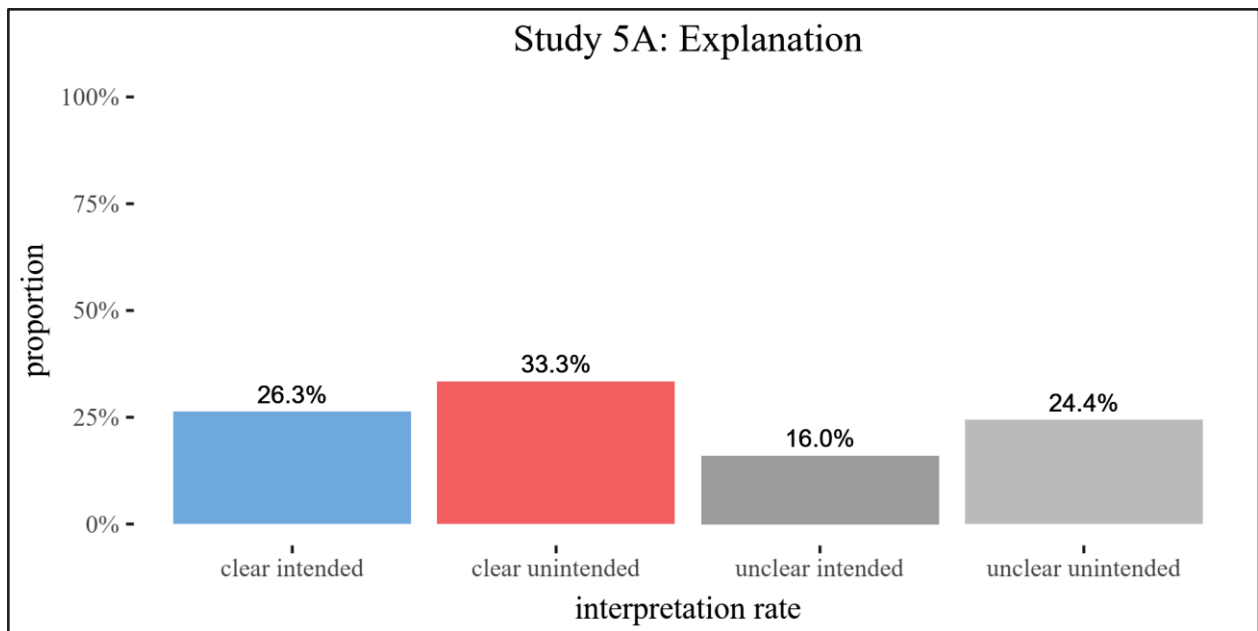
[80] I aggregated responses across all items rather than analyzing each separately because there were too few participants per item (14-17 per item). Future studies should assess response rates for individual items.

### 4.10.2 Study 5B: Meaning

As expected, most participants did not interpret items in a clear intended way when summing all conditions. Overall, 21.8% ($n = 34$) of participants provided a clear intended interpretation. This was significantly less than 0.5, $\chi^2(1, N = 156) = 47.41$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 28.37%. Across all items, a majority of participants did not interpret those in a clear intended way, with the proportion of participants who provided clear intended responses ranging from 5.9% (item 1, $n = 17$) to 42.9% (item 6, $n = 14$). Overall, clear unintended interpretations were high, with 32.1% of participants across all items expressing a clear unintended interpretation ($n = 50$). Clear unintended interpretations were moderate to high. Results for individual items are available on **Figure 4.7.**

**Figure 4.7**

*Interpretation rates for Study 5B (aggregated)*
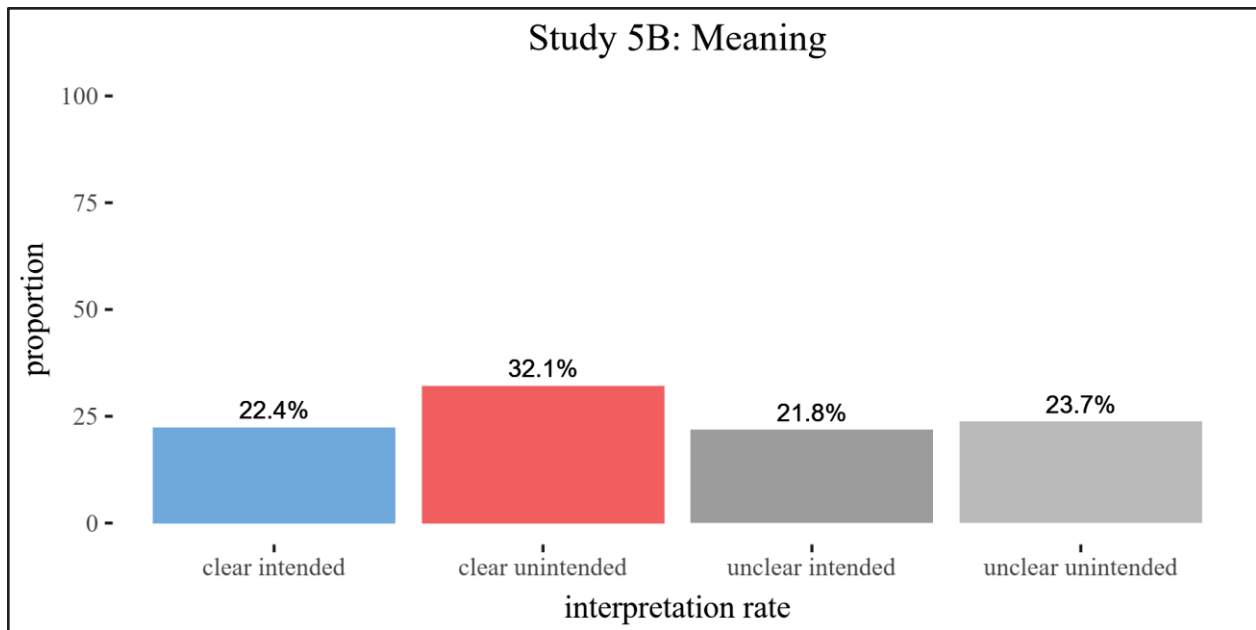
**Discussion**

Consistent with previous studies, most participants did not tend to interpret items on the MRS in a clear intended way. However, they did perform *better* than previous studies, and as we will see, for all subsequent studies. This may be attributed to the many steps taken to enhance the validity of the items used in the MRS. Notably, the MRS is the only scale to have undergone a rigorous validation process and the benefits may be revealed in the improved rate of clear intended interpretations]. Nevertheless, clear intended rates were still very low, and clear unintended interpretation rates were still very high. If anything, the low rate of clear intended interpretation rates even following such strenuous efforts to create a valid set of measures is a *stronger* indication that participants struggle to interpret such questions as intended, and provides somewhat better support for metaethical indeterminacy.

**4.11 Study 6: The New Metaethics Questionnaire (NMQ)**

Study 6 employs the same methods as Study 5 to assess items on a 3-item scale (YB3) that appears in Yilmaz & Bahçekapili (2015b) and the New Meta-ethics Questionnaire (NMQ) employed in Yilmaz & Bahçekapili (2015a). Interpretation rates for all items in Study 6 were coded exclusively by myself. The same pattern of extremely low clear intended rates for all conditions (which ranged from 11.4% to 22.5%), and all proportions of clear interpreted interpretations were significantly less than 0.5. For full discussion, see **Supplement 4**, section *S4.4.2*.

**4.12 Study 7: Folk Moral Objectivism scale (FMO)**

Study 7 employs the same methods as Study 5 and Study 6 to assess items on the Folk Moral Objectivism (FMO) scale that appears in Zijlstra (2019). Interpretation rates for all items in Study 7 were coded exclusively by myself. The same pattern of extremely low clear intended rates for all conditions (which ranged from 12.4% to 19.8%), and all proportions of clear interpreted interpretations were significantly less than 0.5. For full discussion, see **Supplement 4**, section *S4.4.3*.

**4.13 General discussion**

Over the course of this chapter, I have endeavored to provide a comprehensive body of evidence in support of two conclusions:

(1) Most participants do not understand questions about metaethics as intended
(2) This could be explained by widespread folk metaethical indeterminacy: people do not understand questions about metaethics because they don't have metaethical stances or commitments

It is far easier to provide evidence for (1) than for (2). Studies 1A through 1C reveal that people frequently do not attribute descriptions of a moral disagreement to differences in moral standards, even though this is not only the intended interpretation, but a necessary condition for the validity of the paradigm. Study 2 showed that when people are asked to assess other people's interpretations of questions about metaethics, many still fail to interpret those questions as intended. This study included both a response to a version of the disagreement paradigm, and direct questions about realism and relativism, expanding the range of questions in addition to providing a novel approach to assessing how people interpret questions about metaethics. One advantage to this study is that it circumvents the concern that questions about metaethics are ambiguous. By providing a response to such questions that was specifically designed to reflect an intended interpretation, and to thereby reflect the kinds of metaethical rationale consistent with an intended interpretation, the fact that participants still reliably fail to interpret these questions as intended provides converging evidence that participants do not interpret questions about metaethics as intended.

The low rate of clear intended interpretations and the comparatively high rate of clear unintended interpretations recurred across all studies: Study 3 demonstrated that participants struggle to interpret questions about metaethics in the concrete context of charitable giving, Study 4 showed that participants did not interpret the terms "objective" and "relative" as intended in both concrete and abstract cases, suggesting that studies that employ explicit metaethical language are likely failing

to reliably prompt intended interpretations. Finally, Studies 5, 6, and 7 show that participants do not interpret the items used in a variety of scales designed to assess folk metaethics as intended. When asked to explain why they agreed or disagreed with items intended to reflect their metaethical stances or commitments, the reasons given by most participants did not clearly indicate that they interpreted these questions as intended. Supporting this low rate of intended interpretations, asking participants what the items on these scales meant likewise revealed that few participants could clearly articulate what these statements meant.

Taken together, these findings represent multiple, converging lines of evidence across a wide variety of measures, all of which suggest that most ordinary people do not interpret questions about metaethics as intended. This conclusion is supported by thematic analysis, as well, which revealed recurring themes among unintended interpretations that made sense of and could readily account for why interpretation rates were so low. For instance, many participants appeared to conflate questions about relativism with descriptive claims about the differences in the moral standards of different people or cultures. This is precisely the kind of unintended interpretation we would expect of people engaging with the stimuli but failing to interpret it as intended (rather than, e.g., random, disorganized responses that might instead suggest disinterest or inattention).

At the very least, these findings challenge the validity of the disagreement paradigm, explicit paradigms (e.g., Fisher et al., 2017) , and existing scales used to assess folk metaethics. Nevertheless, there are considerable limitations to these findings. Before we can even address the degree to which these results support (2) metaethical indeterminacy, we must first acknowledge and, where possible, address these limitations.

### Limitations

Given the unusual nature of my analysis, I have opted to provide a more extensive section on limitations than is typical. See **Supplement 4**. Here, I will provide a brief outline of the main limitations of the studies reported here.

(1) *Generalizability.* All studies were conducted on US participants, mostly on MTurk. This significantly limits the generalizability of the results reported here.

(2) *Lack of demographic data*: In several instances I did not gather demographic data. This prohibits analysis of potential demographic differences within samples.

(3) *Potentially poor response quality*: Most studies were conducted in contexts in which participants may have had suboptimal incentive to engage with content. This may have resulted in reduced response effort, which could compromise the quality of the data.

(4) *Risk of bias and erro*r: With the exception of Study 1A, all data was coded exclusively by myself. This introduces significant risk of bias, idiosyncrasy, and error.

(5) *Introspective access*: Participants may have competence with metaethical concepts but are unable to convey that competence in a research context. If so, these findings may overestimate the number of unintended interpretations.

(6) *Accounting clear intended interpretation rates*: A significant minority of participants do offer clear intended interpretations. This must be explained in a way consistent with indeterminacy for indeterminacy to remain a viable hypothesis.

(7) *Tension between indeterminacy and validity*: Low clear intended interpretation rates could be explained by *poor validity* or *indeterminacy*. However, the less valid a study, the worse indeterminacy serves as an explanation. Since I argue for both, there is some tension in my explanation of findings.

(8) *Untested paradigms*: Interpretation rates for some measures have not yet been tested. In addition, some of those that have been tested had few participants per item.

***Future directions***

Like the limitations outlined above, I will only briefly reference future directions here. See **Supplement 4** for a fuller discussion.

(1) *Acquire additional coders*: The most important step is to identify people competent to code the same datasets. Especially adversarial coders disinclined to share my biases.

(2) *Improve generalizability*: All studies were of MTurk participants or college students in the United States. Future research should be directed at non-WEIRD populations and subpopulations of interest, e.g., insular religious communities.

(3) *Improve rigor*: Participants could be assessed under more rigorous conditions, e.g., greater incentives, less time constraints, detailed instructions, or following training paradigms.

(4) *New or more detailed questions*: small changes to the present method could be made, e.g., researchers could ask different questions or request longer and more detailed responses.

(5) *Larger sample size for scale items*: Validity of scale items should be tested with a larger number of responses per item to better estimate interpretation rates for individual items.

(6) *Comparison to expert populations*: The same measures could be tested among expert populations to assess whether interpretation rates substantially improve.

(7) *Explore nonmoral domains*: Similar methods could be used to assess interpretation rates for realism/antirealism about other domains, including aesthetics, personal preferences (in e.g., food, music), social conventions, prudential norms, epistemic norms, and descriptive claims (e.g., science, math, religion), among others.

(8) *Employ richer qualitative methods*: Researchers could conduct a richer variety of qualitative measures of folk metaethics, such as interviews.

(9) *Employ corpus methods*: Researchers could employ corpus methods to assess large bodies of real world text. This could be used to assess e.g., the proportion of metaethical language used, the contexts in which metaethical claims are made, and whether people employ metaethical language consistently and in a way consistent with academic philosophy (see Chartrand, 2022 for discussion on corpus methods in experimental philosophy). This has recently been done (Stojanovic, 2022)

**4.14 Conclusion**

In this chapter, I introduced a novel method for assessing interpretation rates for questions about metaethics. This mixed methods approach involves the acquisition of open response data, coupled with a coding scheme designed to assess the overall rate at which participants appear to clearly interpret an item in an intended or unintended way. Across all studies, significantly fewer than half of participants interpreted questions about metaethics as intended. Extremely low rates of intended interpretations were found for every major scale and paradigm tested, as well as a variety of novel paradigms introduced by my colleagues and me. This suggests that existing methods for measuring folk metaethical stances and commitments are not valid. If most people do not know what they are being asked, and if a sizable minority, often accounting for around half of the participants, appear to have clearly interpreted what they were asked in an unintended way, results will consist largely of noise.

I have argued that there are two main factors contributing to the low rate of intended interpretations: many of the stimuli and questions are ambiguous or exhibit low face validity. As a result, many of the measures used to assess folk metaethics are not valid measures of folk metaethics. Second, I have argued that ordinary people don't have determinate metaethical stances or commitments, and that this partially explains the low rate of clear intended interpretations. Even correcting for ambiguity and reducing the poor face validity will be insufficient: people continue to interpret questions and statements about metaethics in a variety of unrelated ways even under conditions in which I have sought to clarify the relevant metaethical concepts. Far from existing in tension with one another, a large part of the reason researchers struggle to convey metaethical concepts in ordinary language, and a large part of the reason why interpretative variation, and unintended interpretations, remain high is *because* people lack the conceptual resources to respond to questions about metaethics appropriately. That items routinely fail to adequately convey metaethical concepts may be due, in part, to correctable researcher error. But such errors are themselves a

symptom of a deeper issue: metaethical distinctions are sophisticated, technically, metaphysically top-heavy notions that are simply absent from the way ordinary people speak and think.

Unfortunately, qualitative data is unlikely, on its own, to convince optimists about the prospects of empirical folk metaethics or anyone else who may be skeptical of my approach. In the next chapter I turn to a handful of quantitative approaches to shed further light on folk metaethics, which go some way in providing corroborating evidence in support of metaethical indeterminacy and the invalidity of existing measures of folk metaethics.

## CHAPTER 5:
## Multiple Choice & Indeterminacy Paradigms

### 5.0 Introduction

My goal in this chapter is to provide converging evidence that most ordinary people don't interpret questions about metaethics as intended. This chapter is divided into two sections. First, I employ multiple choice questions and Likert scale items designed to assess how people interpret questions about metaethics. This approach provides a straightforward quantitative alternative to assess interpretation rates for questions about metaethics. The second section presents participants with a set of questions about nonmoral issues: in particular, their views on how to interpret quantum mechanics and mathematical platonism. The goal of these studies is to illustrate that people can appear to express a determinate philosophical stance towards an issue even when they plausibly lack such a position. Furthermore, these studies illustrate *spontaneous theorizing* by illustrating how forced choice paradigms create the artificial appearance of a determinate position where none existed prior to participating in the study.

### 5.1 Multiple choice paradigms

In the previous chapter, I assessed interpretation rates using open response questions. Written responses provide a rich body of data that can provide insights not only into interpretation rates, but the particular ways participants interpreted various metaethics stimuli. Yet there are limitations to this approach.

One shortcoming with open response questions is that people may have an implicit competence with metaethical concepts, but the structure and wording of open response questions is insufficient to render those concepts salient in a way that manifests in people's responses. Using open response questions to evaluate how people interpret questions about metaethics potentially suffer

from the very problem they are ostensibly designed to assess: people may interpret open response questions in unintended ways. If so, participants coded as having unclear or unintended responses may be competent with the relevant metaethical distinctions, but fail to demonstrate this. Ironically, then, open response studies may suffer from the same shortcoming as the measures they were designed to debunk.

It is also possible that ordinary people may lack the relevant vocabulary to clearly convey the relevant metaethical concepts, even if they have them. For instance, a person might recognize the distinction between stance-independent and stance-dependent facts, but lack a clear or recognizable terminology to express this. In study 2 of the previous chapter, I asked participants to explain what a *response* to a question about metaethics meant, rather than the question itself. This goes some way in mitigating these concerns, since the response uses precisely the kinds of vocabulary I (and hopefully other competent coders) would consider a clear way of conveying the relevant metaethical distinctions. Yet this doesn't go far enough. There are no guarantees that people will interpret a response to questions about metaethical questions as intended, either.

There is no easy solution to these problems. Nor is any one method likely to decisively settle the matter. Our best bet is to provide mutually corroborating lines of evidence using a variety of distinct methods and theoretical approaches that converge on the same conclusions. To this end, I developed a simple and straightforward set of studies that eschew coding open response in favor of measures subject to conventional quantitative analysis. In all studies, participants were presented with a *target statement* that reflected a realist or antirealist metaethical position. Then they were presented with a set of statements. *One* of these statements matched the intended meaning of the target statement (the *matched statement*), while five statements did not match the meaning of the statement, but were designed to appear as plausible candidate meanings (the *decoy statements*). These decoy statements were designed to reflect the common ways ordinary people are inclined to interpret statements about

metaethics in unintended ways. Matched and decoy statements are presented in three formats. In studies 1A through 1D, I used a multiple choice question. Finally, in study 1E, I use a Likert scale that allows participants to express how much they agree or disagree that both the matched and decoy statements match the meaning of the target statement.

The rationale behind these studies is simple: if people interpret statements reflecting metaethical distinctions as intended, we should expect them to be able to match the meaning of these statements with other statements that convey the same meaning as the target statement. If only a small portion of people correctly match the target statement with the matched statement, this suggests that people struggle to interpret the target statement as intended. All target statements were drawn directly from stimuli used in previous studies, were adapted from the stimuli used in previous studies, or were designed to convey to my satisfaction the relevant metaethical distinctions. Thus, if few people match the target statement with the matched statement, this challenges the validity of existing measures and raises doubts about the ease with which researchers can measure folk metaethical stances and commitments using conventional survey methods.

### 5.1.1 Study 1A: Moral relativism scale (MRS) items

Currently, the Moral Relativism Scale (MRS) is the most rigorous and well-developed metaethics scale (Collier-Spruel et al., 2019). Collier-Spruel and colleagues chose from a pool of items rated by a team of experts for accurately representing relativism (or its negation) and that exhibited conventional psychometric virtues, e.g., loading well onto a single factor, had a high Cronbach's alpha, and that exhibited adequate predictive and convergent validity, since scores on the MRS were associated with tolerance and other predicted associations, and with other measures of relativism, respectively (but see Maul, 2017). This made it an ideal candidate for an initial attempt to assess interpretation rates using a multiple choice paradigm.

The MRS consists of ten statements. Participants express how much they agree or disagree with each statement on a five-point scale. Seven of the items express relativism in some form, while three items express realism, universalism, or both, and are reverse-coded, such that a low score on these items is interpreted as a greater relativism score. A composite of one's response to all ten questions may be taken as one's overall endorsement of moral relativism. If these items are valid measures of relativism, then participants presented with items on the MRS should be able to match the meaning of those items with another statement that also expresses relativism (or the rejection of relativism). I predicted that participants would struggle to match items from the MRS with their intended interpretation, and that fewer than half would choose the correct response.

**Methods**

*Participants*. Participants consisted of 201 adult US residents on Amazons' Mechanical Turk. I intended to recruit 200 participants from Amazon's Mechanical Turk. 206 participants began the study. However, 5 did not complete the study, and were dropped from subsequent analysis.

*Procedure*. Across all conditions, participants were first presented with a set of instructions that explained what they'd be asked to do in the study. These instructions appeared on a page without stimuli to encourage participants to read the instructions carefully. Then they were randomly assigned to one of the conditions. Each of these conditions repeated the same instructions, which explained that they would be presented with a statement and asked to choose the statement that had the same meaning, *not* to choose a statement they agreed with the most.

Then all participants were presented with a statement in quotes and bold text. This statement, and the response options that followed, differed across conditions. They were then asked to choose the statement that "is the best interpretation of the meaning of the statement," and were given a set of six statements to choose from. Order for statements was randomized. Two conditions presented a

statement expressing relativism, and two expressed statements intended to express the rejection of relativism. Here is a sample item (the correct answer is underlined):

> Below, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*
>
> We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.
>
> **"The viewpoint of one's culture determines whether their actions are morally right."**
>
> ○ <u>The truth about whether an action is morally right can only be judged according to the standards of different cultures</u>
>
> ○ People from different cultures can be justified in holding different moral beliefs
>
> ○ Each culture adapts its views about what is morally right to fit with their traditions and meet local needs
>
> ○ Since each culture has different views about what is morally right, a person's actions will be judged as moral by some cultures and immoral by other cultures
>
> ○ Different cultures have the right to hold different moral beliefs, so it isn't acceptable to impose our moral values on them
>
> ○ The attitudes of each person's culture influence whether their actions are judged as moral or immoral

As this sample item illustrates, participants could choose one of six response options. One item reflects the intended interpretation (the *correct* response), while the other five do not reflect an intended interpretation (*incorrect* responses). The correct and incorrect responses are unique to each item to provide more context and appear as relevant as possible to the statement used. The *incorrect* responses were intended to reflect a variety of conflations I anticipated would be most plausible for the target statement. For instance, in order from top to bottom, the five *incorrect* responses listed above include: (1) epistemic claim that we're *justified* in holding contrary moral beliefs (*epistemic conflation*), (2) the notion that different people conform to the same moral rules in different ways (*evaluative standard ambiguity*), (3) a descriptive claim about how people will be judged (*descriptive conflation*), (4) a normative claim

about people being entitled to their beliefs (*normative conflation*), and (5) a causal or genealogical claim. Although not originally anticipated, (5) was a very common theme that emerged from analyzing open response questions. The same attention was given to tailoring appropriate alternatives to all other statements, in both this and all subsequent studies.

There are some difficulties with providing an accurate response option for the reverse-coded items that were intended to express the rejection of relativism. However, these problems were due to the inherent misoperationalization of these items. Both items were ambiguous and conveyed more than one concept. These were the items selected:

"There are moral rules that apply to everyone regardless of personal beliefs."

"The same moral standards should be followed by people from all cultures."

The first statement appears to convey *both* universalism *and* realism, since it describes moral rules "applying to everyone" (i.e., universalism) and doing so "regardless of personal beliefs" (i.e., realism). Unfortunately, *neither* the rejection of universalism *nor* the rejection of realism necessarily entail relativism.[81,82] The second item conveys both universalism and a *normative* claim about what people "should" do. Rather than design a "correct" response option that accurately conveyed the full content of these items, I opted for response options that expressed realism since this is the standard

---

[81] The contrast to realism, understood as stance-independence, would be stance-dependence, but stance-dependence and relativism are not identical and don't entail one another. Conversely, universalism could be construed as the view that there is a single correct moral system rather than multiple correct moral systems. This may be a more appropriate contrast, and future studies should assess interpretation when the correct response conveys universalism exclusively. Unfortunately, *if* that were the intended contrast, then this item isn't face valid because it includes a remark suggesting realism. Note that this item was approved by a panel of experts, despite conflating realism and universalism. This isn't a critique. I'd like to think I'm an expert and I've conflated realism and universalism in my own research as well. Rather, I want to draw attention to the fact that people who invest their professional careers in this topic *still* conflate these distinctions. Are we really expecting participants with no experience and little incentive to think carefully about this topic to avoid these conflations?
[82] Note also that agreeing with this item is consistent with cultural relativism, since it only states that rules apply regardless of personal beliefs, not the standards of one's culture. The other statement conveys a rejection of cultural relativism. It's worth noting that rejecting subjectivism doesn't entail universalism or realism, since you could be a cultural relativist, while rejecting cultural relativism does not entail universalism or realism, since you could be an individual subjectivist. Thus, neither of these items accurately captures a proper continuum between relativism and the rejection of relativism.

contrasting concept presented in the literature. However, an argument could be made that the optimal

contrast would be some form of universalism (or more specifically some type of non-indexicalism).[83]

Our primary goal was to assess whether most participants would choose the correct response.

This sets the threshold at 50%, which is remarkably low. If only half of the participants in a study

interpreted an item as intended, this would raise serious questions about whether the measure in

question was valid. Of course, it is possible that the "correct" response is ambiguous or confusing or

difficult to interpret. If so, participants who interpreted the original statement as intended may still

choose an incorrect response. By setting the bar very low, I hoped to buffer against such errors.

Nevertheless, I did my best to select a correct interpretation that accurately reflected the intended

metaethical position even if the original statement did not do a great job of conveying that position.

***Measures***. The primary measure consisted of a single multiple choice question with six response

options. The *correct* response option matched the meaning of the target statement. The remaining five

*incorrect* response options did not match the meaning of the target statement. All responses were coded

as either "1" = correct or "0" = incorrect. The four target statements may be seen below. Each set of

six response options was individualized to the particular target statement.

> *Relativism 1*
> Different people can have opposing views on what is moral and immoral without anyone being wrong.
>
> *Relativism 2*
> The viewpoint of one's culture determines whether their actions are morally right.
>
> *Realism 1*
> There are moral rules that apply to everyone regardless of personal beliefs.
>
> *Realism 2*
> The same moral standards should be followed by people from all cultures.

---

[83] E.g., I might have used something like "there is a single correct standard of moral values" as a correct response.

## Results

A one sample proportion test without continuity correction was conducted for each of the four conditions to determine whether the proportion of participants who selected the correct response was significantly less than 0.5. For Relativism 1, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi 2(1, N = 50) = 13.52$, $p < 0.001$.[84] With 95% confidence, the percentage of correct responses was less than 35.10%. For Relativism 2, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi 2(1, N = 51) = 18.84$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 30.18%. For Realism 1, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi 2(1, N = 48) = 18.75$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 29.59%. For Realism 2, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi 2(1, N = 52) = 15.08$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 33.87%. The frequency and percentage of correct responses is featured on **Table 5.1**.

## Discussion

Across all four conditions, less than half of participants successfully matched an item from the MRS with its intended metaethical interpretation. This held for items representing relativism and the rejection of relativism. In fact, participants did only marginally better than chance (i.e., 16.67%).[85]

---

[84] As noted in chapter 4 (where I provide a similar comment), all one sample proportion tests were conducted in R using prop.test(). However, rather than conducting a one sample proportion test directly, prop.test() conducts a chi-squared test. Fortunately in these cases $\chi 2$ is equivalent to the square of the Z-statistic one would obtain by conducting a one sample proportion test using the standard formula. To demonstrate this, note that the formula for a one proportion Z-test is $Z = (p-p_0) / \sqrt{p_0(1-p_0)/n}$. For this study, this would give us $= (0.24-0.5)/\sqrt{(0.5(1-0.5)/50} = -3.676955$. $(-3.676955)^2 = 13.52$, which is equivalent to the $\chi 2$ reported in the main text.

[85] And for Relativism 1, the pattern of responses was not significantly different from chance. A chi-square goodness-of-fit test was performed to determine whether people were equally likely to choose each of the six responses. The pattern of responses was not significantly different from an equal distribution across all responses for Relativism 1, $\chi 2(5, N = 50) = 5.20$, $p = 0.392$. However, given the relatively small number of participants per condition this study is probably underpowered.

These findings strongly indicate that very few participants matched the selected items from the MRS scale with statements intended to accurately reflect the intended metaethical concept.

**Table 5.1**

*Proportion of correct responses (by condition)*

| Item | Correct | | Incorrect | | Total |
|------|---------|------|-----------|------|-------|
| | **%** | **n** | **%** | **n** | |
| Relativism 1 | 24.0 | 12 | 76.0 | 38 | 50 |
| Relativism 2 | 19.6 | 10 | 80.4 | 41 | 51 |
| Realism 1 | 18.8 | 9 | 81.2 | 39 | 48 |
| Realism 2 | 23.1 | 12 | 76.9 | 40 | 52 |

*Note. n* = 201.

### 5.1.2 Study 1B: Fisher et al.'s explicit measures

Fisher et al. (2017) attempt to assess folk realism by *explicitly* asking people whether morality is objective. Participants were presented with questions about a variety of topics and asked whether they think there is an "objectively true" answer to those questions. Here is the wording that they used:

> Consider the following question: "Should [topic] be allowed?" Please tell us whether you think there is an objectively true answer to this question. (p. 1127)

Responses were measured using a 7-point Likert scale (1 = definitely no objective truth, 7 = definitely an objective truth).[86] Topics included a variety of issues, including "same-sex marriage, marijuana legalization, teaching evolution in school, abortion, and violence in videogames" (p. 1126). According

---

[86] Note that technically this is a measure of *confidence* in realism, not whether the participant endorses realism or not. It's not clear that someone who doesn't think realism is "definitely true" is *less of a realist* than someone who does. I remain puzzled that researchers uncritically treat degrees of epistemic confidence as an indication of "how much" you believe something. A person can be extremely committed to a view but be hesitant to maintain that it's "definitely" true. For instance, a passionate religious believer may be open to doubt and choose a 6, while an ambivalent religious person could choose 7 because they don't have any doubts. It's not clear what it would mean to say the second person believes "more" or "more strongly."

to Fisher et al., one advantage of this approach is that "it removes any ambiguity about whether we really are asking about whether there is an objective truth about the topic" (p. 1127).

Unfortunately, this is a shockingly naive approach to "removing ambiguity." Far from removing ambiguity, asking people whether there is an "objective truth" to a moral issue presumes that participants will interpret this term in a way consistent with researcher intent and with one another. As I demonstrated in chapter 4, they do not. "Objective" is an ambiguous phrase that has multiple possible meanings. Far from eliminating ambiguity, Fisher et al. amplified it. The term "objective," as it is used in philosophy, is a piece of technical jargon with a narrow meaning. Everyday uses of "objective" include notions such as "unbiased," "capable of being measured using publicly available criteria," and so on. The mistake Fisher and others make is to imagine that ordinary people's conception of "objective" matches the technical way the term is used in academic philosophy.

The goal of this study was to assess whether people would reliably match the meaning of an adapted version of the stimuli Fisher et al. (2017) used. Since participants were explicitly asked to select the best interpretation of the meaning of a statement, it would have been unnecessarily complicated to use Fisher et al.'s exact stimuli, since it included a question within a broader statement. Asking them to interpret a statement within a statement would be unnecessarily cognitively demanding. "Consider the following question" also seemed redundant and unnecessary for assessing whether people interpreted Fisher et al.'s stimuli as intended. To resolve these concerns, I adapted the wording of the original stimuli to make the statement as clear as possible. I chose one of the moral issues that appeared in Fisher et al.'s study, frontloaded an explicit metaethical statement, and then brought in the concrete moral issue I selected and used the phrase "objectively true" to match the precise term used by Fisher et al. Although Fisher and colleagues did not employ a similar question for relativism, I also devised a novel question that explicitly asked about moral relativism as well. Once again, my goal was to assess whether most participants would match the meaning of these statements

to another statement that expressed realism and relativism, rather than a statement that conveyed some other unintended concept.

**Methods**

*Participants*. Participants consisted of 301 adult US residents on Amazon's Mechanical Turk (124 females, 176 males, 1 other, $M_{age}$ = 38.2, $SD_{age}$ = 11.4, age range = 19-74). I intended to recruit 300 participants from Amazon's Mechanical Turk. 302 participants began the study. However, one participant did not complete any of the measures, and was dropped from subsequent analysis.

*Procedure*. This study employed the same procedure as Study 1A. All participants were presented with the same instructions, which clarified that they would be asked to match the target statement with the statement that most closely matched its meaning. Participants were then randomly assigned to either the realism or relativism condition. They were given reminder instructions, the target statement, and a set of six response options, where they could select the statement that most closely matched the meaning of the target statement. Finally, participants were presented with the comprehension check as study 1B and I collected data on age and gender.

*Measures*. The primary measure consisted of a single multiple choice question with six response options. The *correct* response option matched the meaning of the target statement. The remaining five *incorrect* response options did not match the meaning of the target statement. All responses were coded as either "1" = correct or "0" = incorrect. In addition, a comprehension check was introduced to assess whether participants interpreted the task as intended (i.e., to select the statement that matched the meaning of the target statement). The A sample item with the correct response is underlined is presented below:

In response to the question on the previous page, which of the following were you doing?

○ Selecting the single statement which I think best matches the meaning of the bolded sentence
○ Selecting the single statement that I most agree with
○ Selecting every statement that I agree with
○ Selecting every statement which I think best matches the meaning of the bolded sentence

*Explicit | Realism*
"Moral truth is objective, so there is an objectively true answer to whether violence should be seen as an appropriate response in conflicts."

*Explicit | Relativism*
"Moral truth is relative, so the truth about whether violence should be seen as an appropriate response in conflicts is relative."

Finally, participants were given a comprehension check and demographic data (age and gender) was collected.

**Results**

A one sample proportion test without continuity correction was conducted for both conditions to test whether the proportion of participants who selected the correct response was significantly less than 0.5. For the realism condition, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi2(1, N = 149) = 53.16$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 26.05%. For the relativism condition, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi2(1, N = 152) = 96.61$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 16.09%. The frequency and percentage of correct responses is featured on **Table 5.2**. 80.1% (N = 241) of participants passed the comprehension check. Excluding participants who did not pass the comprehension check did not meaningfully change any results.

**Table 5.2**

*Proportion of correct responses (by condition)*

| Item | Correct | | Incorrect | | Total |
|---|---|---|---|---|---|
| | **%** | **n** | **%** | **n** | |
| Explicit \| Realism | 20.1 | 30 | 79.9 | 119 | 149 |
| Explicit \| Relativism | 11.2 | 17 | 88.8 | 135 | 152 |

Note. *n* = 301.

**Discussion**

Once again, most people did not interpret questions about metaethics as intended. In this case, participants did not interpret statements explicitly conveying the notion that morality is "objective" or "relative" in a way consistent with the metaethical positions these terms are supposed to correspond to. These results were found in a larger sample than previous studies, and were not meaningfully influenced by the failure of many participants to pass a comprehension check. This casts serious doubt on Fisher et al.'s claim that they successfully eliminated ambiguity by asking whether morality is objective. Although it was not a part of Fisher et al.'s measures, people also struggled to understand what it would mean to say that moral truth is relative. Even if people struggle to interpret explicit use of metaethical terms as intended, they may still interpret other, indirect ways of expressing realism and antirealism. In the next study, I assess an attempt to prime participants with realism and antirealism. Once again, my goal is to assess whether participants match the statement with its intended metaethical meaning. Of course, I only assessed interpretation rates for a single item (violence as a method of resolving conflicts), so it remains possible participants would interpret "objectively true" differently in reference to some other moral issue, or if the item were worded differently. However, absent any positive indication that people do interpret terms like "objective" and "relative" as intended, the onus remains on those who do believe participants interpret explicit questions about whether morality is objective or relative as intended to furnish supporting evidence.

### 5.1.3 Study 1C: Young and Durwin's measures

Young and Durwin (2013) conducted a study that assessed whether priming people with realism and antirealism would influence charitable giving. The details of that study are unimportant for our purposes. Our focus is whether the priming conditions that they used could plausibly serve as methods for inducing belief in realism and antirealism. Y&D employed two conditions with metaethics primes. In both conditions, a canvasser approached people on the street, pitched a charity, and attempted to solicit donations for the charity. In the course of this pitch, canvassers used either a realism or antirealism prime that consisted of asking participants a simple question:

> Realism: "Do you agree that some things are just morally right or wrong, good or bad, wherever you happen to be from in the world?"

> Antirealism: "Do you agree that our morals and values are shaped by our culture and upbringing, so there are no absolute right answers to any moral questions?" (p. 303)

One obvious problem with these items is that they are not face valid representations of realism or antirealism. The realism condition would be better interpreted as an expression of universalism. It could also be read as a descriptive claim, i.e., that people *consider* some of the same things to be right or wrong everywhere in the world. The antirealism item is much worse. It actually contains two claims: an uncontroversial descriptive claim about the origins of our moral standards that most people would probably agree with, coupled with the denial of "absolute" right answers to moral questions. Unfortunately, "absolute" does not clearly convey stance-independence, but could instead mean all sorts of things: an undeniable answer, an exceptionless moral rule, and so on. Nevertheless, it's still possible that participants would match these priming conditions with the appropriate metaethical stance. The items used in this study were taken verbatim, with the sole exception of replacing the question mark with a period and treating them as statements rather than questions. Once again, my goal was to assess whether most participants would match the target statement to the correct metaethical position.

**Methods**

***Participants***. Participants consisted of 202 US residents on Amazon's Mechanical Turk (91 females, 110 males, 1 other, $M_{age}$ = 41.1, $SD_{age}$ = 11.8, age range = 22-71). I intended to recruit 200 participants from Amazon's Mechanical Turk. 202 participants completed the survey.

***Procedure***. I employed the same measures as those employed in Study 1B and Study 1C, using new target statements and response options.

***Measures***. I employed the same measures as Study 1B: all participants were given one target statement, and were asked to choose which of the six response options matched the meaning of the target statement. This was followed by a comprehension check and demographic questions.

> *Y&D | Realism*
> Some things are just morally right or wrong, good or bad, wherever you happen to be from in the world.

> *Y&D | Antirealism*
> Our morals and values are shaped by our culture and upbringing, so there are no absolute right answers to any moral question.

**Results**

A one sample proportion test without continuity correction was conducted for both conditions to test whether the proportion of participants who selected the correct response was significantly less than 0.5. For the realism condition, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi^2$(1, N = 100) = 27.04, $p$ < 0.001. With 95% confidence, the percentage of correct responses was less than 31.65%. For the antirealism condition, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi^2$(1, N = 102) = 28.59, $p$ < 0.001. With 95% confidence, the percentage of correct responses was less than 31.06%. The frequency and percentage of correct responses is featured on **Table 5.3**. 82.7% (N = 167) of participants passed the comprehension check. Excluding participants who did not pass the comprehension check did not meaningfully change any results.

**Table 5.3**

*Proportion of correct responses (by condition)*

| Item | Correct | | Incorrect | | Total |
|---|---|---|---|---|---|
| | **%** | **n** | **%** | **n** | |
| YB3 \| Realism | 24.0 | 24 | 76.0 | 76 | 100 |
| YB3 \| Antirealism | 23.5 | 24 | 76.5 | 78 | 102 |

Note. *n* = 202.

**Discussion**

Once again, most participants did not interpret questions about realism and antirealism as intended. Notably, a very high proportion of participants chose specific incorrect responses. In the realism condition, 37% of participants chose the response option "There are some actions that are considered morally wrong in all societies." In the antirealism condition, 49% of participants chose the incorrect response option, "Each person's views about what is morally right or wrong is shaped by their unique background and experiences, so we frequently reach different conclusions about moral questions." Both of these claims are *descriptive*, with the latter including an additional claim about the etiology of moral belief: that we reach different conclusions *because* of our different backgrounds and experiences. This suggests that participants don't find the target statements so confusing that they simply guess at their meaning, which would typically result in a more random response pattern. Rather, they appear to systematically favor incorrect response options. This lends support to the conclusion that people reliably conflate claims about metaethics with unintended concepts.

### 5.1.4 Study 1D: Bush and Moss's face valid metaethical items

So far, all of the target statements were drawn from or based on existing folk metaethics paradigms. However, these items all suffer very poor face validity. As a result, we could explain the low rate of intended interpretations as an indication not that people don't have determinate metaethical stances or commitments, but as evidence that researchers have consistently failed to properly operationalize

171

folk metaethics stimuli. To address this possibility, I adapted items that had been developed for an unpublished folk metaethics scale and used these as target statements. My goal was to craft items that would reflect more face valid representations of realism and relativism. I opted for relativism rather than a more general way of expressing antirealism since it seemed less plausible that people would interpret the denial of stance-independent moral facts as intended as readily as they'd interpret a statement expressing relativism. That is, I wanted to choose items that I thought had a reasonable chance of enjoying a higher rate of intended interpretation. I chose the following two items:

> *B&M | Realism*
> There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

> *B&M | Antirealism*
> The truth of all moral claims can vary depending on the moral standards of different individuals or cultures.

## Methods

*Participants*. Participants consisted of 201 US residents on Amazon's Mechanical Turk (88 females, 113 males, $M_{age} = 40.0$, $SD_{age} = 11.9$, age range = 19-81).We intended to recruit 200 participants from Amazon's Mechanical Turk. 201 participants completed the survey.

*Procedure*. I employed the same measures as those employed in previous studies, using new target statements and response options.

*Measures*. I employed the same measures as Studies 1B and 1C: all participants were given one target statement, and were asked to choose which of the six response options matched the meaning of the target statement. This was followed by a comprehension check and demographic questions.

## Results

A one sample proportion test without continuity correction was conducted for both conditions to test whether the proportion of participants who selected the correct response was significantly less than 0.5. For the realism condition, the proportion of participants who selected the correct response was

significantly less than 0.5, $\chi2(1, N = 100) = 4.00$, $p < 0.023$. With 95% confidence, the percentage of correct responses was less than 48.22%. For the antirealism condition, the proportion of participants who selected the correct response was significantly less than 0.5, $\chi2(1, N = 101) = 64.96$, $p < 0.001$. With 95% confidence, the percentage of correct responses was less than 15.88%. The frequency and percentage of correct responses is featured on **Table 5.4**. 87.1% ($n = 175$) of participants passed the comprehension check. Excluding participants who did not pass the comprehension check, 41.8% ($n = 38$) of participants chose the correct response, which resulted in the realism condition no longer being significantly less than 0.5, $\chi2(1, N = 91) = 2.47$, $p = 0.058$. With 95% confidence, the percentage of correct responses was less than 50.38%. No other results were meaningfully changed.

**Table 5.4**

*Proportion of correct responses (by condition)*

| Item | Correct | | Incorrect | | Total |
|------|---------|---|-----------|---|-------|
| | % | n | % | n | |
| B&M \| Realism | 40.0 | 40 | 60.0 | 60 | 100 |
| B&M \| Relativism | 9.9 | 10 | 90.1 | 91 | 101 |

Note. $n = 201$.

**Discussion**

Most participants in the relativism condition did not interpret the target statement as intended. Likewise, most participants in the realism condition did not interpret the target statement as intended, unless I removed participants who failed the comprehension check (N = 10). Once I did so, I could no longer reject the null hypothesis that the proportion of people who select the correct response is equal to or greater than 0.50. This suggests that it's possible that the true proportion of people who interpret the realism condition as intended is equal to or greater than half. I discuss this possibility in the general discussion. In short, this outcome is plausibly due to inadequate power to detect small

differences. With the upper bound of the confidence interval just barely over 50%, these results are not good evidence most people interpreted the item as intended. Overall, these findings are consistent with the conclusion that people do not reliably interpret questions about metaethics as intended. Even if around half of the participants did interpret one of the items as intended, such an item would still be a valid measure of realism. And if this is the best we can achieve, then efforts to convey realism to ordinary people via single-sentence measures are probably hopeless.

### 5.1.5 Study 1E: Likert scale paradigm

One shortcoming with multiple choice paradigms is that they require participants to choose just one of the response options available. There are at least three limitations with measuring interpretation rates this way. First, it is possible that if participants believe more than one of the response options matches the meaning of the target statements. Second, it is possible that participants believe *none* of the response options match the meaning of the target statement, but if so they have no way to express this. Third, this method does not allow us to assess *how well* one or more statements match the meaning of the target statement. It is possible, for instance, that some participants only slightly agree that a particular response option matches the meaning of the target statement, while others strongly agree that it does so. Requiring participants to express a categorical judgment prevents us from assessing such variability.

Study 1E addresses all of these shortcomings by presenting participants with a Likert scale that allows them to express level of agreement with each response option separately. This allows participants to judge that more than one of the response options match the meaning of the target statement, or that none of them do, and it also allows them to express level of agreement. This study drew on four items from the MRS (two realism, two antirealism) that were used in Study 1A and the two items I created that were used in Study 1D (one realism, one antirealism). These six items (three

realism, three antirealism) served as the target statements. All three antirealist conditions represented either cultural relativism or subjectivism.

In addition to assessing how well the response options originally devised for the multiple versions of these items matched the meaning of the target statement (the one correct *matched* response and five incorrect *decoy* responses), I also asked participants how well each of the other two alternative target statements reflecting the same metaethical position (i.e., either realism or antirealism) matched the meaning of the target statement, in order to assess whether participants would judge that items ostensibly intended to measure the same psychological construct would mean the same thing.[87] I will refer to these as the *alternative* statements.

It's possible that participants would be significantly more likely to judge that the target statement matched the alternative statements and the matched statement than the decoy statements. This would not be a good indication of the validity of the target statement, however. Such a tendency could be significant but marginal, and a marginal tendency to favor the intended interpretations of an item would not be an indication that a measure is valid, only that it performs better than chance. As such, my expectation was that there would be no reliable tendency, overall, for participants to demonstrate an *overwhelming* tendency to judge that the target statement would match the meaning of the alternative statements or the matched (i.e., correct) statement, when compared to the decoy (i.e., incorrect) statements. Instead, I expected the general pattern to be one in which participants

---

[87] Note that realism and antirealism could be depicted as complex constructs with multiple subdimensions (e.g., antirealism could be divided into noncognitivism, relativism, and error theory), so it's possible participants may not think that items that express e.g., individual subjectivism mean the same thing as cultural relativism. Such judgments would be reasonable, since individual subjectivism and cultural relativism *don't* mean the same thing. Notably, the scales typically used to measure people's metaethical attitudes don't draw such distinctions, and instead treat measures of e.g., cultural relativism and individual subjectivism as measures of the same underlying construct. For instance, Collier-Spruel et al. (2019) use the following two items as measures of "relativism":

(1) *The viewpoint of one's culture determines whether their actions are morally right*
(2) *Each person is the final authority on whether his or her actions are morally corre*ct

Taken at face value, these items are not consistent. If cultures determine moral truth, individuals can't be the final authority, and vice versa.

demonstrated no clear and decisive tendency to match the meaning of the target statement to the other target statements or the correct response, when compared to incorrect responses.

**Methods**

*Participants*. Participants consisted of 604 US residents on Amazon's Mechanical Turk (313 females, 288 males, 2 other, $M_{age} = 42.1$, $SD_{age} = 13.0$, age range = 20-80). I planned to collect 600 participants in total (100 participants per condition). A total of 604 participants completed the survey.

*Procedure*. Each participant was assigned to one of six conditions. In each condition, participants were presented with one target statement. They were asked to judge "the degree to which you agree or disagree that the additional statements match the meaning" of the target statement, for each of eight statements (1 = Strongly disagree, 7 = Strongly agree). For the realism conditions, these eight statements consisted of the other two realist target statements, the correct matching statement, and five incorrect statements. For the antirealism conditions, the eight statements consisted of the other two antirealist target statements, the correct matching statement, and five incorrect statements. For each condition, the correct and incorrect statements were designed to most closely reflect the meaning of that statement (for the correct statement) and to most closely reflect plausible unintended interpretations.

*Measures*. Measures consisted of level of agreement with the eight statements described above on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree), a comprehension check, and demographic questions. The comprehension check appeared after the main measures and assessed whether participants interpreted the task as intended (i.e., that their level of agreement indicated how well each statement matched the meaning of the target statement, *not* how much they agreed with that statement).

**Results**

For each condition, I ran linear mixed model analysis in Jamovi (version 2.3.13) using the GAMLj module (version 2.6.6) to estimate differences in levels of agreement between the statements participants were asked to match to the target statement (this included both alternative target statements, the matched statement, and the five incorrect "decoy" statements). For instance, for condition 1 I included $r\_1$ as the dependent variable, where $r\_1$ is level of agreement that each of the statements matches the meaning of the target statement and added the target statement ($c\_1$) as a fixed effect. I included participant (ID) as a random effect (random intercept only). The model specification was as follows: $r\_1 \sim 1 + c\_1 + ( 1 \mid ID )$. The same model was used for conditions 2 through 5. Using the GAMLj module, I also assessed the estimated marginal means and 95% confidence intervals for responses to each of the eight statements. Linear mixed models are often used for more sophisticated designs. However, as Magezi (2015) observes, "it is important to realize that the use of LMMs is by no means restricted to complex grouping designs, and can also be used for experimental psychology studies with a single grouping factor of participant or subject" (p. 2; see also Singmann & Kellen, 2019).

Although a linear mixed model would allow tests of whether agreement with the other target statements and the correct response were significantly greater than the incorrect responses, as I noted above, the goal of this analysis is primarily descriptive and exploratory (see above), and is primarily intended to provide evidence that participants do not overwhelmingly match target statements that reflect a particular metaethical position (e.g., realism) with responses that match their meaning (i.e., the other target statements and the correct response). As such, a significant difference between mean response levels for these statements and incorrect statements would not, by itself, be informative. Nevertheless, I conducted pairwise comparisons both without correction and with Bonferroni correction for all pairs of statements. Results are available in **Supplement 5**. Finally, since the same

general pattern occurred across all conditions, I report only the results for Condition 1 in the main text, and provide results for the remaining five conditions in **Supplement 5**.

Graphs displaying the estimated marginal means with 95% confidence intervals are featured in **Figure 5.1**. This figure displays the estimated marginal mean level of agreement for each of the eight statements across all six conditions along with 95% confidence intervals. Means above 4 indicate a tendency to agree that the statement matched the meaning of the target statement, while means below 4 indicate a tendency to disagree that the statement matched the meaning of the target statement. Moving from left to right for each condition, the first two items along the x-axis (*a_1* and *a_2*) are the alternative target statements that are intended to reflect the same metaethical position (i.e., either realism or antirealism). The next item, *m_1*, is the matched statement, which is intended to reflect the intended interpretation of the target statement. Finally, *d_1* through *d_5* are the decoy statements, which reflect unintended interpretations. Thus, higher mean scores for *a_1*, *a_2*, and *m_1* indicate that participants correctly matched the target statement to its intended interpretations, while higher mean scores for *d_1* through *d_5* indicate that participants incorrectly matched the meaning of the target statement to unintended interpretations. At a glance, it is clear that participants showed no general tendency to overwhelmingly favor the intended interpretations (the alternative or matched statements) over the unintended interpretations (the decoy statements) for *any* of the target statements. The closest any condition came was for BM Realism 1. The estimated marginal mean for *a_1* was 5.74, 95% CI [5.42, 6.05]. However, this overlaps with *d_1* and *d_5*. For every other condition, the alternative target statements and the matched statement did not overwhelmingly outperform other items in level of agreement. Indeed, the alternative and matched statements often underperformed compared to decoy statements.

**Figure 5.1**

*Estimated marginal means with 95% CIs (by condition)*
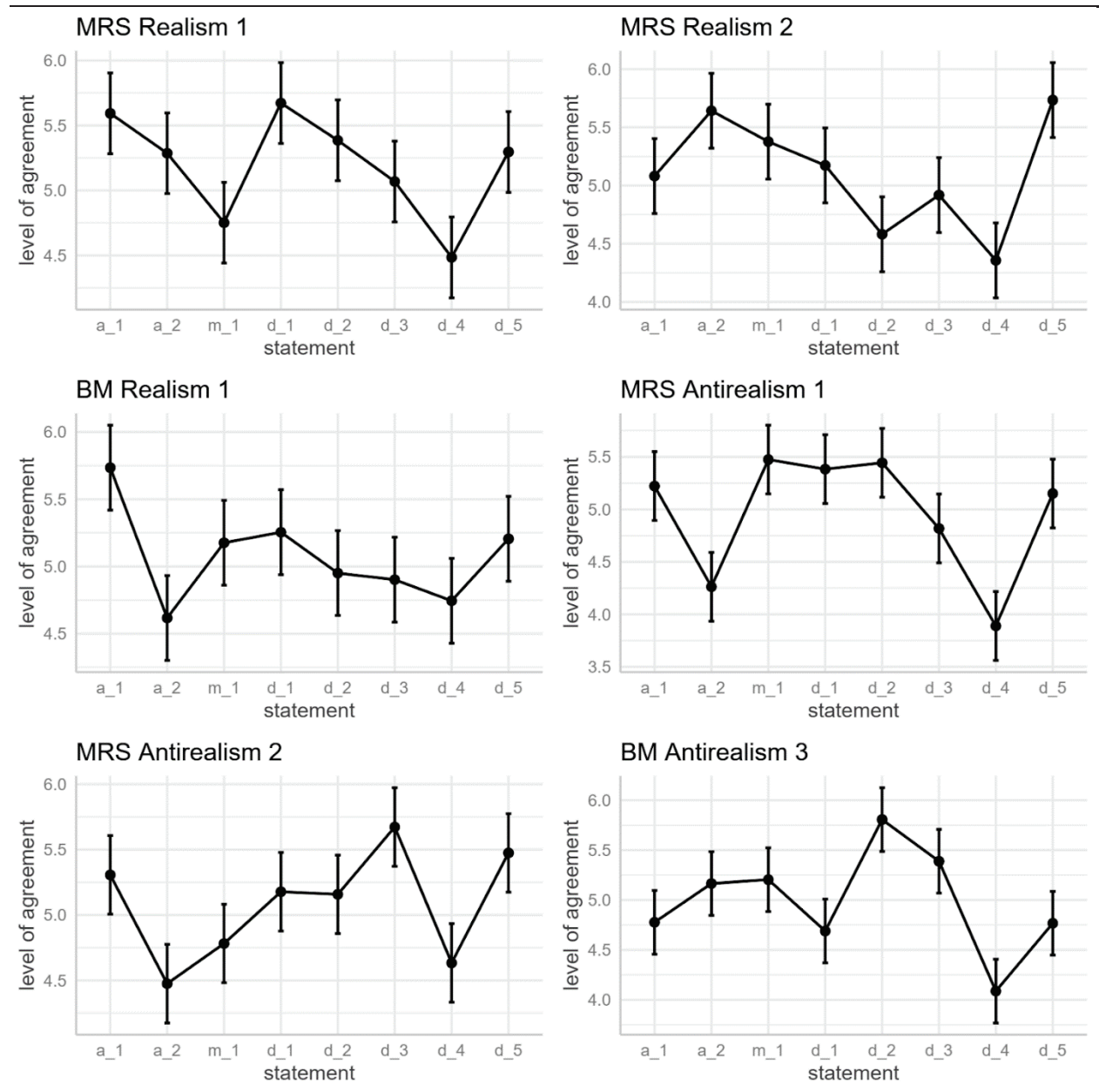


*Note.* Going from left to right, "a_1" and "a_2" refer to the two alternative target statements, "m_1" refers to the *matching* (or correct) statement, and "d_1" through "d_5" refer to the *decoy* (or incorrect) statements. As such, higher scores for the first three items are indicative of interpreting the target statement as intended, while higher scores for the five items on the right indicate a failure to interpret target statements as intended.

For MRS Realism 1, MRS Realism 2, MRS Antirealism 2, and BM Antirealism 3, the statement with the highest estimated marginal mean level of agreement was one of the decoy statements, while for MRS Antirealism 1 the difference between the matched statement and the first and second decoy statements (i.e., $d\_1$ and $d\_2$) was negligible. However, one of the most noteworthy patterns across all conditions is that participants *tended to agree that every statement matched the meaning of the target statement*. Across 48 opportunities to judge how well a statement matched the meaning of the target statement, only *one* item fell below the midpoint, d_4 in MRS Antirealism 1, and only just barely ($M = 3.89$). This may be attributable in part to acquiescence (Podsakoff et al., 2003). If so, it may be that a general tendency to agree obscures evidence that participants interpreted target statements as intended.

Participants did not perform well on the comprehension check. Only 69.2% ($n = 418$) participants chose the correct response, suggesting that many participants may not have performed the task as intended, and may have, e.g., been expressing their level of agreement with statements, rather than judging how well statements matched the meaning of the target statement. However, results did not substantially change when such participants were excluded from analysis.

Under ideal conditions, we would expect participants to agree that the alternative and matched statements match the meaning of the target statements, and to *disagree* that the decoy statements match the meaning of the target statement. Yet this is simply not what we find. A consistent tendency to agree that decoy statements match the meaning of the target statement threatens the validity of the target statements as measures of people's metaethical stance because it indicates that people do not interpret such statements as *only* expressing a metaethical claim. If an item on a metaethics scale is interpreted to express both a metaethical claim *and* one or more non-metaethical claims (e.g., descriptive or normative claims), then we'd be unable to know whether people's level of agreement with such items can be attributed to their metaethical views, or to the unintended non-metaethical content they attribute to the claim. In short, these findings provide further evidence that people do

not reliably interpret items intended to reflect metaethical claims as intended. Participants do not reliably interpret such claims as exclusively *metaethical* claims. Instead, they seem to judge that such items may also reflect a variety of unintended meanings.

**Discussion**

Once again, results indicate that participants do not appear to interpret items on metaethics scales as intended. In order for items on metaethics scales to serve as valid measures of realism and antirealism, participants should exhibit a reliable and overwhelming tendency to match the meaning of those items with statements that are intended to reflect the same metaethical concept, and a consistent tendency to recognize when a statement doesn't match the meaning of the item. Participants did not do this. Instead, they frequently judged that non-metaethical statements matched the meaning of the target statement. This provides further evidence that participants do not interpret items on metaethics scales as intended.

One limitation of these findings is that many participants failed the comprehension check. Although results did not substantially change when such participants were excluded from analysis, it's possible that many participants were expressing their level of agreement with the statements, rather than expressing how well those statements matched the meaning of the target statement. This could explain why participants exhibited a general tendency to agree with all statements. Although this may be due in part to acquiescence, it could be that the statements were, on average, the sorts of claims people tend to agree with. If so, it's worth noting that participants were evenly divided between realist and antirealist conditions, yet they exhibited a similar tendency to agree with items in both conditions. If participants tended to agree with statements in both realist and antirealist conditions, this would suggest a general tendency to endorse statements that reflect realism *and* to endorse statements that reflect antirealism, at least when they are presented in isolation. It may be the inconsistency between

statements reflecting realism and antirealism are more salient, or only become salient at all, when they are presented side-by-side.

Future research could improve on this design by making the nature of the task more explicit to participants. For example, this could be achieved by restating the purpose of the task for each item, e.g., participants could be asked to express their level of agreement that "[statement] means the same thing as [target statement]" or they could be asked for each statement how well it matches the meaning of the target statement, with a Likert scale ranging from 1 = Very poorly to 7 = Very well. Participants could also be given a training exercise on an unrelated topic at the start of the task to test their understanding. There may be still other ways to ensure participants interpret the task as intended.

## 5.2 Study 2: Indeterminacy in non-metaethical domains

In the previous section, I demonstrated that participants reliably fail to match statements that express realism, antirealism, and relativism to other statements that are intended to represent the same concepts. These findings suggest that people struggle to interpret folk metaethical stimuli as intended. However, this only serves to corroborate findings from the previous chapter: people don't interpret questions about metaethics as intended. In this last study, I show that it's possible to generate the superficial appearance of genuine patterns of folk philosophical stances even where it is unlikely any exist, and propose that such results are best explained as a result of *spontaneous theorizing*. Evidence of spontaneous theorizing in non-metaethical domains supports the possibility that spontaneous theorizing could also explain apparent patterns of realism and antirealism in folk metaethics research.

Anyone familiar with research on folk metaethics will observe that we find consistent, replicable patterns of results. If metaethical indeterminacy is true, this poses a challenge. If metaethics stimuli were completely unfathomable, we might expect an approximately *random* distribution of responses, or at least a pattern that could be readily explained by factors other than the construct of interest, e.g., acquiescence bias (Podsakoff et al., 2003).

Yet this isn't what we find. How can the hypothesis that ordinary people have no determinate metaethical stances or commitments account for this? The previous chapters provided evidence that most folk metaethics research relies on invalid measures. As a result, the pattern of results we observe is consistent with people having determinate metaethical stances or commitments, but systematically interpreting metaethical stimuli in unintended ways. In other words, existing findings are consistent with the possibility that people have determinate metaethical stances or commitments, but existing studies have simply failed to measure them.

I have yet to develop a method for ruling this explanation out. It may be that no simple set of studies could do so, and that it will take many converging lines of evidence to resolve the matter. Yet training paradigms (e.g., Wright, 2018; Pölzler & Wright, 2020a; 2020b; see **Chapter 3**) represent one method for breaking through this impasse. If we can ensure that participants interpret what we're asking as intended, then we can be sure that their answers genuinely reflect their metaethical stances or commitments because we've deliberately constructed studies that successfully *cause* the intended interpretation.

Unfortunately, this trades one dilemma for another: once we induce people to think about folk metaethics, their responses could reflect a stance or commitment they developed as a result of the training, rather than a stance or commitment they held prior to participating in the study. If so, then we could not make inferences about these participants to the population they were drawn from, since they would no longer be members of that population. Recall that the target population of folk metaethics is ordinary people. Ordinary people are, by definition, people without adequate philosophical training. How could we study such a population by recruiting people from that population, and then providing them with adequate philosophical training? That would make no more sense than attempting to study people who've never been to Japan by flying them to Japan for the study.

But suppose Pölzler and Wright (2020a; 2020b) convincingly argue that training paradigms can successfully induce participants to interpret questions about metaethics as intended, but don't require providing enough philosophical training that participants no longer qualify as ordinary people. That is, suppose we are dealing with a population of *ordinary people* who *understand questions about metaethics as intended*. When faced with such a situation, we can no longer dismiss results as the product of unintended interpretations. We'd have to accept that whatever answers people provide *do* reflect determinate metaethical stances or commitments.

Regrettably, even this would be insufficient to conclude that ordinary people have determinate metaethical stances or commitments. This is because training paradigms point to another possibility: participants may have held no determinate stance prior to participating in a study, but form a position in virtue of participating in the study. That is, they may engage in *spontaneous theorizing*. Spontaneous theorizing occurs *whenever a participant who held no determinate stance or commitment prior to participating in a study is induced to develop or express a stance or commitment due to the experience of participating in the study itself.*

This is not to say people genuinely develop a full-fledged philosophical position in the span of a few minutes that they will carry with them for the rest of their lives. Such theorizing may be transient: the context of a given task may prompt a person to favor one position over another and respond accordingly for the purposes of the task, yet this may represent only a fleeting inclination towards the view on superficial consideration, could quickly be buried in one's unconscious or forgotten altogether, and in any case may make no meaningful difference in how a person thinks, speaks, or acts after participating in the study. That is, spontaneous theorizing doesn't require that people become full-blown converts to a particular position; they need only think about it *enough* that they form enough of a position on the matter to meaningfully respond to stimuli, regardless of whether this has any long term psychological implications for their stances or commitments.

Spontaneous theorizing poses a serious threat to training paradigms. If we cannot tell whether responses reflect stances or commitments held prior to participating in a study, or were formed during the course of the study, training paradigms may lack external validity since we cannot be sure whether the pattern of responses obtained in a training paradigm tells us anything about the metaethical views of people outside the context of the study. Yet spontaneous theorizing also poses a threat to folk metaethics research in general. Such research typically relies on forced choice paradigms: participants are *required* to express a realist or antirealist stance. As a result, all possible responses are interpreted that participants hold a particular metaethical stance. When a participant has no recourse other than to select one of the available response options, even if they have no determinate position and even if they do not interpret the stimuli as intended, they may be disposed to respond in nonrandom ways. This may result in a systematic disposition to favor some response options over others, independent of whether those response options accurately reflect the putative psychological construct researchers are studying. Such inclinations may be too insignificant to reflect a determinate stance or commitment, e.g., superficial features of one response option may make it more appealing than another, such as one statement seeming more affirmative or positive ("Yes, we have free will") than the alternative ("No, we don't have free will"), even if the participant doesn't understand what's meant by "free will." Yet it's also possible that participants *do* interpretation questions as intended at least *some* of the time. When this occurs, conventional folk metaethics research cannot distinguish instances of spontaneous theorizing from instances in which a participant expressed a stance or commitment held prior to exposure to the study stimuli.

In the previous chapter, we saw that clear intended interpretation rates are not *literally zero*. Some people do appear to interpret questions about metaethics as intended. And while we cannot tell which people interpreted questions as intended rather than simply guessed correctly in the studies reported in this chapter, it's possible that at least *some* participants interpreted statements about realism,

antirealism, and relativism as intended. While some of these cases may represent genuine instances of determinate folk metaethical views held prior to participation, some may also be the result of spontaneous theorizing. This would make sense of why many people don't interpret metaethical stimuli as intended, but *some* do, without requiring us to propose that *everyone* who appears to interpret questions about metaethics as intended has a determinate metaethical view.

Spontaneous theorizing can explain why at least some people appear to have determinate stances or commitments in the context of a study: because some people *do* have a determinate stance or commitment. Yet this explanation is consistent with folk indeterminacy. So long as the determinate stances or commitments result from participating in the study, they will have no necessary implications for how people think *outside* the context of the study. The goal of this final study is to illustrate that even in instances where, on *a priori* grounds, we have good reason to believe people don't genuinely understand certain concepts and distinctions, they may nevertheless develop a perspective on the matter via participation in the study.

This study also serves as a rebuttal to a recent critique of metaethical indeterminacy put forward by Pölzler and Wright (2020a).[88] Pölzler and Wright intend to reject indeterminacy with

---

[88] Their characterization of indeterminism strikes me as questionable, or at least incomplete:

"Indeterminism: Moral sentences and judgements are indeterminate with regard to the cognitivism/non-cognitivism distinction. That is, they are correctly analyzed by both theories (like men with a certain amount of hair may be correctly described both as being bald and as not being bald)" (p. 23).

My concern is with the notion that competing theories provide a *correct* analysis. This makes it seem like both accounts *win*, when it might be more apt to say the accounts are *not more explanatorily adequate than the other*. In his initial formulation, Gill (2009) characterized indeterminacy as the claim "that some parts of ordinary moral discourse give us *no reason to prefer* an analysis that involves one meta ethical commitment over an analysis that involves the commitment that has traditionally been taken to be its meta-ethical competitor" (p. 216). This does *not* sound like the claim that both analyses are literally correct. For comparison, suppose two scientific hypotheses were equally consistent with available evidence. Would we conclude that both hypotheses were *correct*? No. We'd simply conclude that both were equally adequate explanations. Gill (2009) is even more clear about this when he elaborates on indeterminacy later in the article:

"According to the Indeterminacy Thesis (which is the "I" of the IV Thesis), many parts of our moral thought and language provide no good answers to the questions that were central to much of 20th century meta-ethics, vindicating neither relativism nor absolutism, neither internalism nor externalism, etc. The Indeterminacy Thesis holds that the relationship between some instances of ordinary moral discourse and these meta-ethical debates is analogous to the relationship between ordinary arithmetic and debates in the philosophy of mathematics. There is no fact of the matter as to whether

respect to whether ordinary people are cognitivists or noncognitivists, rather than realists or antirealists but their critique applies equally well to the realism/antirealism dispute. In particular, they claim that "Many versions of indeterminism, hybrid expressivism, and incoherentism predict that 'truth-apt' and 'not truth-apt' classifications will be distributed fairly equally for each statement" (p. 24). Yet they argue, because their results showed that a significant majority of participants judged every statement to be truth-apt, the "distribution of responses was quite unequal" (p. 24). In particular, they found that about three quarters of their participants favored noncognitivist response options across conditions. Since this unequal distribution is inconsistent with what we'd expect if indeterminism were correct, their findings purportedly cast doubt on indeterminism.

There aren't "many versions" of indeterminism or incoherentism, so it's a bit puzzling for them to say that *many* versions of these accounts would make such predictions. Yet the more troubling problem feature of these claims is that it's not at all clear that indeterminism does, in fact, predict that responses to questions about metaethics would be "fairly equally distributed" for any given set of statements. An equal distribution is what we might expect if people were completely indifferent between two accounts, and had no inclination whatsoever for one position over another. The result would be that nothing would bias participants towards selecting one response over another. However, we should only expect an equal distribution, in practice, if no superficial or extraneous features of a response option would systematically incline participants to favor one position over its alternative. Such conditions *might* be met in practice. Yet superficial or extraneous features of stimuli may, in practice, prompt a disproportionate number of participants to favor one position over another. Even

---

ordinary mathematic usage is better explained by a Platonist or anti-Platonist conception of number. The way people use numbers in everyday math simply does not contain answers to the questions that animate philosophy of mathematics." (p. 218)

It is clear that Gill does not take indeterminacy to entail both accounts are correct, but that *neither* account is correct. Unfortunately, Pölzler and Wright have mischaracterized indeterminism.

if we would predict an equal distribution under *ideal* circumstances, where all features of the stimuli irrelevant to what is being measured are held constant across response options, *conditions are never actually ideal.*

Psychological characteristics of a given population may systematically incline a subset of a given sample (and, by extension, the population they're drawn from) towards one response over another merely due to features of the participants unrelated to the construct of interest. For instance, acquiescence bias is the well-documented "tendency to answer affirmatively to a question no matter what it's content" (Knowles & Nathan, 1997). Even in instances where people hold no particular position towards the stimuli, they may still exhibit a general tendency to favor affirmative responses. This holds *even when the stimuli in question are meaningless, uninterpretable, or could not possibly represent any legitimate psychological construct.*

Maul provides one demonstration of this, showing that reliable patterns emerged from a set of scale items even when those items asked about nonsense concepts (e.g., *quintessence*), consisted of lorem ipsum, or were simply blank (that is, no question was even asked). Here, I provide another example. Nelson (2013) describes an amusing demonstration of such spurious patterns in a collaboration between Nelson and Meyvis. In the course of developing a manipulation, they presented participants with a set of 20 pictures of animals, and asked them to rate each in terms of its speed "goodness." As Nelson observes, "The latter could be best construed as an evaluation of moral worth. *That is an absurd question*" (emphasis mine). Despite the absurdity of the question, participants answered. After all, as Nelson points out:

> In surveys, most people answer most questions. That is true regardless of whether or not questions are coherently constructed and reasonably articulated. That means that absurd questions still receive answers, *and in part because humans are similar to one another, those answers can even look peculiarly consistent.* (Emphasis mine)

They found a number of amusing results. For instance, people judged tortoises to be the most morally virtuous. However, the most important result was that almost every animal was judged, on average, to be good. Their "scale had a meaningful midpoint, yet all but three animals are above it." Apparently, hyenas, barracudas, and jellyfish were the most diabolical of the set of animals. Interestingly, they also recruited two experts (an ecologist and an evolutionary biologist) to respond to the scale as well. Critically, both judged that "any response would be random," with one stating that "I would probably tie them all in ranking." Of course, this isn't what happened, and both experts likewise gave seemingly nonrandom responses that even correlated with one another ($r = 0.29$). In fact, both agreed with ordinary people that tortoises are highly virtuous.

Notably, both experts expected a response pattern, perhaps because the questions were nonsensical, when this clearly turned out not to be the case. I remain puzzled as to why researchers would think that people would respond randomly to questions, even when the questions are absurd. Simply because a question is nonsensical, it doesn't follow that people will just throw their hands up and choose random responses or uniformly select the midpoint if one is available. While this might hold if stimuli are completely inscrutable, people are still generally motivated to engage with a study, and will often do their best to interpret stimuli in *some* way, even if this requires substituting what's being asked for a more meaningful question. For instance, people may judge tortoises as more virtuous because they are more likable or less dangerous. Cultural associations may also influence people's responses. Tortoises are associated in popular culture with patience (*The Tortoise and the Hare*), and, after all, patience is a virtue. Hyenas, in contrast, are portrayed as villains in iconic movies such as *The Lion King*, and are often depicted as malicious and unappealing.[89] Such associations may influence

---

[89] A report on the conservation status of hyenas notes that "they are viewed with contempt and fear" and observe that "Tourists do not rate hyaenas very highly" (Hofer & Mills, 1998).

people's responses even if people don't literally believe any of these animals are actually more or less virtuous.

This example illustrates how meaningless or nonsensical questions can nevertheless result in meaningful response patterns. However, I wanted to construct a study that demonstrated a nonrandom response pattern that more closely resembled the stimuli used in folk metaethics research, and, in particular, the training paradigms employed by Pölzler and Wright. My goal was to show that if you employ measures that are intended to measure philosophical stances or commitments that couldn't plausibly figure into the way ordinary people speak or think, that you can still obtain results in which a disproportionate number of people favor one position over another.

I opted for competing interpretations of quantum mechanics. Quantum mechanics is a highly successful theory that describes the nature of atomic and subatomic phenomena. It is widely regarded as one of the most advanced, mathematically intense, and philosophically impenetrable fields of study, featuring mind-bendingly complicated equations and seemingly intractable puzzles about the fundamental nature of reality that have vexed many of the world's greatest thinkers over the past century. However precise and replicable its equations may prove, there is still the question of how to make sense of the various subatomic events captured by these equations. While some theorists wish simply to focus on the math, others are concerned about the best way to interpret quantum mechanics, with many taking the view that it ought to provide a "literal description of reality" (de Muynck, 2004, p. 92). This is not the place for an introduction to the topic, nor would I be competent to do so. For our purposes, it will suffice to say that interpretations of quantum mechanics revolve around extraordinarily complex issues in mathematics and philosophy that prompted Feynman to quip that "I can safely say that nobody understands quantum mechanics" (p. 123). In a popular comment on the remark, Mainwood (2020) points out that the quote was delivered during a lecture. According to Mainwood:

Feynman opens this lecture by musing on the fact that as science has progressed, it has moved further and further from our intuitive experience of everyday objects; and that while this is unsurprising given how limited our everyday experience is, it makes the work of understanding and explanation more and more difficult.

Mainwood's point is that Feynman is not claiming that literally nobody understands quantum mechanics, but rather that we cannot understand quantum mechanics "in terms of a simple, familiar model." This is because quantum mechanics represents a broader tendency for scientific discoveries to move further from everyday experience over time. In short, quantum mechanics has radically counterintuitive and bewildering implications.

Given its radically counterintuitive nature, it is incredibly implausible that an implicit commitment to any particular interpretation is implicit the way ordinary people speak or think. And it is nearly as implausible that any appreciable number of ordinary people have studied quantum mechanics well enough to have any particular stance on the matter. As such, it is an ideal candidate for folk indeterminism: there is simply no good reason to think most people have a determinate stance on the matter. In short, we should presume that most ordinary people have no determinate stance on how to interpret quantum mechanics. Nevertheless, you could present participants with a description of different interpretations of quantum mechanics, and a forced choice between two or more response options. If the pattern of responses is not equally distributed across different interpretations, this would indicate that responses could be unequally distributed across response options even if we have strong reasons to suspect ordinary people don't have a determinate stance or commitment on the topic.

This is exactly what I did. I decided to narrow my focus to the two most prominent interpretations of quantum mechanics: the Copenhagen and Many Worlds interpretation. Participants were given a brief explanation of what quantum mechanics was, and a description of each interpretation. Then they were asked to decide which interpretation they thought was correct. Since

the Many Worlds interpretation (MWI) suggests that the universe is constantly bifurcating into countless other universes, I expected ordinary people to find this far more bizarre and implausible than the Copenhagen interpretation, which treats outcomes as probabilistic. Although I am skeptical of the measures used in studies on free will, such studies do suggest that people tend to reject determinism (Nadelhoffer et al., 2014). Even in the absence of data, it strikes me as more plausible that, if confronted with the one of two possibilities: that some events in our universe have random, indeterminate outcomes that cannot be predicted in advance, or that the universe is *literally* in a perpetual state of spawning countless slightly different copies of itself, each of which is populated by everything from slightly different versions of themselves to worlds with honest-to-gosh mustache-twirling supervillain versions of themselves, I figured the former would strike most people as more plausible.

As such, I predicted that a significant proportion of people would favor the Copenhagen interpretation over the Many Worlds interpretation. In addition, I predicted that most people would be unfamiliar with quantum mechanics, would report having no prior position on the matter, would state that quantum mechanics rarely or never comes up in everyday conversations, and that many people would agree that the structure of the study compelled them to respond in ways that didn't reflect their actual position, suggesting that the forced choice paradigm prohibited them from conveying what they actually thought. I asked a handful of other questions, mostly for exploratory purposes, including whether most ordinary people have an opinion on how to interpret quantum mechanics, whether the way ordinary people speak presupposes a particular interpretation, and whether beliefs about quantum mechanics have any practical significance.

### 5.4.1 Study 2: Quantum mechanics

**Methods**

*Participants*. Participants consisted of 200 US residents on Amazon's Mechanical Turk (55 females, 144 males, 1 other, $M_{age}$ = 37.3, $SD_{age}$ = 10.3, age range = 20-72).

*Procedure*. All participants were presented with the same conditions and measures. Participants were presented instructions indicating that they would be presented with different perspectives on an issue, but that the issue may be unfamiliar to them and that no prior knowledge is expected or necessary. This was intended to make it clear that they were not expected to have any prior knowledge of quantum mechanics, which may have discouraged precisely those participants of interest from continuing with the study (i.e., people who aren't familiar with quantum mechanics). Next, participants were given five questions to assess their knowledge and familiarity with quantum mechanics and to assess whether they have an opinion on which interpretation of quantum mechanics is correct. Then participants were presented with a description of the Copenhagen and Many Worlds interpretations of quantum mechanics:

> **Quantum Mechanics**
>
> Quantum mechanics is a branch of physics that deals with atomic and subatomic phenomena. It is a highly successful theory that makes precise predictions and has inspired new technologies. Despite this success, scientists disagree about how to interpret it as a literal description of the world.
>
> One disagreement concerns the nature of some subatomic events. These events appear to have more than one possible outcome, but when we measure them, we only observe one outcome. There are two competing explanations for what occurs when we observe one of these outcomes.
>
> **The Copenhagen Interpretation of Quantum Mechanics**
>
> According to the Copenhagen interpretation, we only observe one outcome because there is only one outcome. According to this view, such events have probabilistic outcomes. For example, a subatomic particle could have a 50% chance of moving to the left, and a 50% chance of moving to the right. We will not know which way it will move until we observe it. Once we observe it, there is an equal chance it ends up moving in either direction, like a coin flip. There is no way to know whether the particle would go to the left or the right in advance,

because the result is random. Thus, the Copenhagen interpretation holds that some subatomic events have random outcomes.

**The Many Worlds Interpretation of Quantum Mechanics**

According to the Many Worlds interpretation, we observe one outcome but there is actually more than one outcome. According to this view, events do not have probabilistic outcomes. Whenever a subatomic particle could move to either the left or the right, instead of it only moving in one or the other direction, the universe splits into two different universes. The particle moves to the left in one of the universes, and to the right in the other universe. Each of those universes contains an exact copy of us and everything else in the universe, with the only difference being the direction the particle moves. In one universe, we observe the particle move to the left, while in the other, we observe the particle move to the right. The reason the particle seems to have moved to either the left or the right is because we are in one of these universes, and cannot observe the other universe. But there is a copy of us in the other universe that saw the particle move in the opposite direction. As a result, the outcome of subatomic events is never random. Instead, all possible ways the particle could move are realized in different universes. Since many subatomic events occur every second, there are countless other universes, each different from the one you are in.

Participants were asked to select which of these interpretations was correct, a series of questions about the presence and practical relevance of different perspectives of quantum mechanics in everyday life. Finally, participants were given a set of six comprehension checks, were asked whether they felt forced to select one of the two interpretations despite it not representing their actual views, and demographic data was collected.

*Measures*. The main measure was a multiple choice question with two response options. Participants were asked "Which interpretation of quantum mechanics do you think is correct?" and were given the following response options:

**The Copenhagen interpretation is correct**
Whenever a subatomic event appears to have more than one possible outcome, only one of those outcomes randomly occurs.

**The Many Worlds interpretation is correct**
Whenever a subatomic event appears to have more than one possible outcome, all possible outcomes are realized. The universe splits into a different universe for each possible outcome, and each of those possibilities occurs in one of those universes.

The familiarity questions included (a) a checkbox featuring eight technical terms related to quantum mechanics (e.g., "quantum harmonic oscillator," "finite potential well"). Participants were asked to select all terms they were familiar with (b) a question about their highest level of education in physics (ranging from "none" to "graduate degree in physics (MS, PhD, etc.)," (c) how well they understand quantum mechanics (1 = Very poorly, 7 = Very well), (d) their level of expertise in quantum mechanics (ranging from "no knowledge" to "expert"), and whether they have an opinion on which interpretation is correct (this question proceeded the description of the Copenhagen and Many Worlds interpretations and the main measures, which appeared on the next page).

After the main measure, participants were asked how often the topic of quantum mechanics comes up in everyday life ("never (It has never come up in conversation)" to "Frequently (several times a week)") and whether most ordinary people have an opinion on how to interpret quantum mechanics (with three options (1) Yes, Copenhagen, (2) Yes, Many Worlds, or (3) No). Participants were also asked whether ordinary people presuppose the Copenhagen or Many Worlds interpretations when they speak (again with the option to say yes and choose one of the two interpretations, or no), and how much of an impact that opinions on quantum mechanics have on everyday life (from "no impact" to "enormous impact").

Comprehension checks consisted of six true/false questions designed to assess whether participants understood the description of quantum mechanics and the Copenhagen and Many Worlds interpretations presented earlier in the study. All six questions are straightforward and simple questions that are based on the instructions. For instance, one item stated that, "According to the Copenhagen interpretation, some events have random outcomes." Participants were then asked whether they felt forced to choose one of the two interpretations:

In this study, you were asked to choose between the Copenhagen interpretation and the Many Worlds interpretation. However, it is possible that you do not agree with either of these positions. Did you feel that you had to choose between one of these two answers, even though it did not accurately reflect what you thought?

○ Yes
○ No

Finally, I collected demographic information (age and gender).

**Results**

In order to assess the judgments of ordinary people, which in this case meant people without substantial training in physics, I excluded all participants who met one or more of the following characteristics from analyses:

(1) Anyone with a bachelor's degree, graduate education, or a graduate degree in physics
(2) Anyone who reported understanding quantum mechanics above the midpoint (i.e., anyone with a score of 5 or more on a 7-point Likert scale)
(3) Anyone who reported having "proficient" or "expert" knowledge of quantum mechanics

13% ($n$ = 26) of participants reported having a degree or graduate education in physics, 13.5% ($n$ = 27) were above the midpoint for reported understanding of quantum mechanics, and 7% reported proficiency or expertise with respect to knowledge of quantum mechanics. Most of these responses overlapped, so this resulted in few exclusions. In total 15.5% ($n$ = 31) chose at least one of these responses, and were excluded from subsequent analyses. However, all analyses were conducted without excluding these participants, which did not result in any meaningful changes. This resulted in a remaining pool of 161 participants.

A one sample proportion test without continuity correction was conducted to test whether the proportion of participants who selected each interpretation was significantly different than 0.5. The proportion of participants who selected the Copenhagen and Many Worlds interpretations was significantly different from 0.5, $\chi2(1, N = 169) = 25.00, p < 0.001$, 95% CI [0.69, 0.76]. Overall, 69.2% ($n$ = 117) of participants reported that the Copenhagen interpretation was correct, while 30.8% ($n$ =

52) reported that the Many Worlds interpretation was correct. Participants who were not excluded due to their knowledge of physics overwhelmingly reported having no prior position, with 97% ($n = 164$) reporting that they had no opinion about which interpretation of quantum mechanics is correct.[90]

53.3% ($n = 90$) of participants reported that quantum mechanics "never" comes up in everyday life, while another 34.9% ($n = 59$) reported that it "rarely" comes up. 11.2% ($n = 19$) reported that it sometimes comes up, 0.6% ($n = 1$) reported that it "often" comes up, and no participants reported that it "frequently" comes up. 72.2% ($n = 122$) reported that most ordinary people have no opinion about which interpretation of quantum mechanics is correct, while 21.3% ($n = 36$) reported that ordinary people favored the Copenhagen interpretation and 6.5% ($n = 11$) reported that ordinary people favored the Many Worlds interpretation. 69.2% ($n = 117$) reported that ordinary people don't presuppose an interpretation of quantum mechanics when making claims about the past, present, or future, while 22.5% ($n = 38$) reported that reported that ordinary people presuppose the Copenhagen interpretation, and 8.3% ($n = 14$) reported that ordinary people presuppose the Many Worlds interpretation. 56.2% ($n = 95$) of participants reported that ordinary people's opinions about how to interpret quantum mechanics have "No impact" on everyday life, 34.3% ($n = 58$) reported that opinions about quantum mechanics have "Very little impact," 8.9% ($n = 15$) reported that opinions about quantum mechanics have "Some impact", 0% ($n = 0$) reported a "moderate" impact, and 0.6% ($n = 1$) reported an "Enormous impact." 34.3% ($n = 58$) of participants reported that they felt forced to choose one of the two interpretations, even though it did not accurately reflect what they thought.

Participants performed reasonably well for most comprehension checks with a majority of participants (71.6%, $n = 121$) getting five or six out of a total of six questions correct ($M = 4.99$, $SD = 1.13$). 43.2% ($n = 73$) got all six questions correct, 28.4% ($n = 48$) got five questions correct, 16.6%

---

[90] Obviously, this difference was significant $\chi2(1, N = 169) = 149.59$, $p < 0.001$, 95% CI [0.93, 0.99]. It remained significant when including participants who were not excluded for knowledge or expertise, $\chi2(1, N = 200) = 106.58$, $p < 0.001$, 95% CI [0.81, 0.91].

(*n* = 28) got four correct, 8.9% (*n* = 15) got two correct, 2.4% (*n* = 4) got two correct, and 0.59% (*n* = 1) got 1 correct, and none got zero correct. 89.9% (*n* = 152) got the correct answer for Question 1, followed by 95.3% (*n* = 161) for Question 2, 76.3% (*n* = 129) for Question 3, 59.8% (n = 101) for Question 4, 85.2% (*n* = 144) for Question 5, and 92.9% (*n* =157) for Question 6.[91]

**Discussion**

A substantial majority of ordinary people (nearly 70%) favored the Copenhagen over the Many Worlds interpretation, despite reporting no substantive training, understanding, or expertise in quantum mechanics and in spite overwhelmingly (97%) reporting that they had no opinion on how to interpret quantum mechanics prior to being asked to select which interpret they thought was correct. This provides a compelling illustration of how participants can be induced to express a stance on a position for which they held no position prior to participating in the study. It's not plausible that virtually all participants would report having no interpretation of quantum mechanics when in fact they did have an interpretation.

This may be a simple demonstration, but it has serious implications. This is because these findings present us with a dilemma: since participants were forced to express that either the Copenhagen or Many Worlds interpretation was correct, we have two options: (1) this does *not* genuinely reflect their views, in which case the forced choice nature of the paradigm has produced an invalid measure, or (2) it *does* represent their position on the matter, *but that position was formed seconds ago in the process of participating in the study*, i.e., it's a result of *spontaneous theorizing*. Each of these

---

[91] Participants removed from analysis due to their self-reported knowledge or training in physics (n = 31) performed about the same on Question 1 (90.3%) and substantially worse on every other question. In order from Question 2 through 6, the proportion of correct responses was 71.0%, 48.4%, 38.7%, 45.2%, and 58.1%. I don't know what to make of this. It's possible that participants who reported expertise were overconfident, did not take the comprehension checks as seriously, or that participants who were more likely to fail comprehension checks were less engaged with the study, and the association between self-reported expertise or knowledge and poor performance on the comprehension checks is actually due to participants being inattentive or disengaged when responding to both sets of questions. If so, then at least some self-reported experts may have not actually been experts. Of course, 31 is also an incredibly small number so these proportions are all likely to be very noisy and not very informative.

possibilities could likewise reflect the results of folk metaethics research: responses may be invalid by forcing participants to choose responses that don't reflect their views *or* they could be engaged in spontaneous theorizing.

60% of participants reported that they did *not* feel forced to choose a response that did not accurately reflect their position, which was significantly different from 0.5, $\chi2(1, N = 169) = 16.62$, $p < 0.001$, 95% CI [0.58, 0.72]. This result suggests that around two thirds of people may have engaged in spontaneous theorizing, with the remaining third forced to choose despite their response not accurately reflecting their views. Such self-reports should be interpreted with caution, since there may be considerable incentive due to e.g., self-presentation concerns, cognitive dissonance, or confabulation to report or mistakenly believe that one has formed a substantive stance on a topic even when one hasn't.

Nevertheless, these findings provide some evidence that when people are given the opportunity to read about a topic, they can readily form an opinion about it, even if they didn't previously hold an opinion. This seems like a sufficiently modest claim that it hardly requires evidence, yet this example still provides a clear illustration of the process occurring with respect to a complicated topic for which most people plausibly held no prior view, a topic adjacent in many ways to the kinds of paradigms employed in folk metaethics and folk philosophical research in general. If participants can feel compelled to answer questions even when response options don't represent their actual position with respect to quantum mechanics, the same could be true for metaethics as well. And if participants can spontaneously develop a position about how to interpret quantum mechanics after being introduced to the topic, they could potentially do so when responding to questions about metaethics as well, especially when they are participating in training paradigms that provide far more extensive instructions and explanations than what was provided in this study.

These results provide evidence that it is fairly easy to identify instances in which ordinary people are unlikely to have any particular position on a topic, but can be readily induced to adopt and express a position on that topic during the course of a study. This challenges Pölzler and Wright's (2020a) presumption that indeterminacy would predict an equal distribution across response options. To my knowledge, I'm the only full-fledged proponent of metaethical indeterminacy studying folk metaethics, and this is *not* what I predict.

These findings also illustrate how forced choice paradigms can coerce many people into expressing a position on a topic even if they don't feel their response accurately reflects their views. Even when over a third of participants responded in this way, a disproportionate number of people still favored the Copenhagen interpretation, as expected. If reliable patterns can emerge from data even when people feel compelled to express views that don't reflect what they think, the same could be true for many studies in folk philosophy, including folk metaethics. In fact, even if we focus only on participants who report that they felt forced to choose between the Copenhagen and Many Worlds interpretations even though neither accurately reflected their position, which represented 34.3% (n = 58) of the sample, 63.8% (n = 37) of participants still favored the Copenhagen interpretation, which was *still* significantly differed from 0.5, $\chi^2(1, N = 58) = 4.41$, $p < 0.036$, 95% CI [0.51, 0.75]. In other words, if we look only at the responses of people who claim that neither response option accurately reflected their views, they *still* exhibited an unequal response pattern.[92] Given the wide variety of biases, errors in design, and factors that may favor a disproportionate number of people reliably favoring one position over another for a variety of philosophical issues, there is simply little reason to suppose that folk indeterminacy with respect to any given philosophical issues should result in responses that are eventually distributed across response options.

---

[92] Results were about the same for participants who reported that they did not feel forced, with 72.1% (n = 80) favoring the Copenhagen response, which remained significantly different from 0.5, $\chi^2(1, N = 121) = 4.41$, $p < 0.001$, 95% CI [0.63, 0.80].

**General discussion**

In studies 1A-1E, I demonstrated that when people are tasked with matching the meaning of items used in prominent folk metaethics paradigms, items adapted from such paradigms, or novel items designed with my collaborators to exhibit greater face validity, with response options that share the meaning of those items, participants reliably fail to do so. Across studies 1A-1D, less than half of participants were able to correctly match items with their intended meaning. The only exception was a single item that, once I removed participants who failed a comprehension check, was close enough to the 50% mark that I could no longer reject the null hypothesis.

In fact, these findings may underestimate how difficult it is for participants to interpret questions about metaethics as intended. Matching one statement to another may be some evidence that people interpreted the target statement as intended. But it is not decisive evidence. In this case, about 42% of people who passed a comprehension check correctly judged the first of these statements matches the meaning of the second better than any of available alternatives::

> "There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise."

> "Moral truth does not depend on cultural standards or personal values"

However, understanding that two statements mean the same thing as one another does not demonstrate that one understands either of those statements in anything more than a superficial respect, much less that one understands the relevant items well enough to meaningfully express a stance on the matter. It could simply be that people have sufficient general linguistic competence to pair appropriately similar statements when prompted to do so. At least some of the consistency in people's response to multiple scale items may be attributable to semantic overlap between the items, rather than successful measures of underlying psychological constructs. As Maul (2017) observes, responses to typical survey items "may exhibit statistical associations explainable purely by overlap in

the semantics of the items, rather than relationships between the underlying target attributes" (p. 7; Arnulf et al., 2014). Given this possibility, it could be that some (if not many) correct responses result from people recognizing semantic associations between the target statement and the correct response, without genuinely understanding the meaning of either.

It would be another thing entirely for people to demonstrate that they understand what the target statement means without any explicit prompting and without the ability to select from among a list of candidate options. These studies are, effectively, comprehension checks. Yet there is an asymmetry with such checks: widespread failure to answer them appropriately may be stronger evidence that people didn't understand the concepts in question than passing such checks indicates that they do. This is because comprehension is best seen as a continuum, with people understanding certain concepts to a greater or lesser extent. Comprehension checks typically affix the threshold somewhere on the low end of this continuum, such that to pass a comprehension check is to demonstrate a fairly shallow degree of understanding. Around half of participants failing to meet even these minimal standards is thus, if anything, still an alarmingly poor performance, and in no way indicates that folk metaethics research is vindicated. If anything, if my best efforts can only reach this point, my meager ability to drag performance from 10-25% up to about double that is *more* conclusive than the failure of earlier studies: at least those items have the excuse of lacking face validity. If even face valid items perform this poorly, efforts to devise appropriate measures of folk metaethics may be hopeless.[93]

---

[93] Of course, the fact that I was able to drag those numbers up at all points to the possibility that future efforts to devise stimuli may succeed where I failed. Note, however, that my other effort failed badly, performing worse than average across the previous item, with only 9.9% choosing the correct response. The ~40% rate for the realism item may be a fluke or an indication that some superficial feature of the item is prompting the higher success rate. Whatever the cause, I have certainly not demonstrated that I can *reliably* produce valid items for various forms of realism and antirealism. But perhaps I could. We cannot rule such a possibility out. Perhaps researchers could use a slightly more elaborate description of realism or antirealism that doesn't prompt spontaneous theorizing but still manages to be sufficiently clear so as to prompt a high rate of intended interpretations. Future studies can address this possibility.

Setting this issue aside, the rest of the findings provide further evidence that people do not interpret questions about metaethics as intended. Critically, they do so without relying on open response questions or qualitative analysis, but still yield surprisingly similar proportions of "correct" responses when compared to the proportion of clear intended interpretations from the previous chapter. Given the superior performance of one of David and I's items, which had greater face validity, the most plausible explanation for the extremely low rate of clear intended interpretations found in the previous chapter is a *combination* of both indeterminacy and invalidity: a considerable portion of participants respond in unintended ways due to researcher error. Items are so poorly designed that the ambiguities, conflations, and underspecificity of these items reliably prompts substantial numbers of participants to interpret these items in unintended ways. At the same time, people have no determinate position on realism or antirealism, so even when they manage to successfully navigate the many hazards of flawed measures, most are still left without any way to express a meaningful position.

This brings us to our last study, which illustrates that, even under such circumstances, people can still exhibit a nonrandom tendency to favor one position over another. As such, researchers cannot point to the fact that a disproportionate number of participants favor a particular metaethical position as evidence that people really do have determinate positions, since there is no good reason to think that such patterns wouldn't persist even if folk metaethical determinacy were correct. This result challenges Pölzler and Wright (2020a). There is little evidence that many ordinary people have an explicit position on how to interpret quantum mechanics, nor evidence that the way people speak or think implicitly commits them to any particular interpretation. In short, there is no "folk quantum mechanics." Yet I did not expect that responses to questions about quantum mechanics would be equally distributed between the Copenhagen and Many Worlds interpretations. On the contrary, I expected an unequal distribution. And that's exactly what I found. Endorsing indeterminacy simply does *not* entail the prediction that people would equally favor competing positions, especially when

those positions are explained to the participant. This is *exactly* what Pölzler and Wright (2020a; 2020b) did. Participants were given extensive instructions and engaged in exercises to train them in the relevant concepts and distinctions. If they didn't have a metaethical position before doing so, such training would virtually guarantee that they developed one in the course of the study.

Yet the outcome of a study similar to the one I conducted on quantum mechanics could reveal quite different results if it were adapted to folk metaethics, *even if* folk metaethical indeterminacy is correct. Here's why. When it comes to quantum mechanics, it's plausible that ordinary people recognize that the topic is one very distant from everyday experience, that it requires extensive technical knowledge (and involves a bunch of scary math), and that it isn't the sort of thing they're *expected* to have any stance towards. For comparison, I don't know anything about veterinary medicine. If my wife showed me a blood sample and asked me whether the best diagnosis was Disease A or Disease B, I'd have no idea which disease was more likely. More importantly, I'd *know* that I have no idea. Just so for quantum mechanics. Is the same true for metaethics? I'm not sure. Moral considerations are far more embedded in everyday life, and metaethical concepts plausibly do arise in everyday discussions often enough that people may have the impression that they understand the relevant concepts and distinctions. If they were asked, ordinary people may report that they don't have any position on realism or antirealism. Yet they could just as plausibly report that they do, even if they don't actually understand these positions.

Granted, realism and antirealism are not nearly as technical and difficult to comprehend the Schrödinger equation, but I suspect they're technical *enough*. Indeed, while referencing the study of folk metaethics, Nichols (2014) states that "Giving a precise characterization of objectivity is itself a major philosophical endeavor" (p. 734). While Nichols goes on to characterize *one* conception of objectivism as, "roughly speaking," a type of attitude-independence, others point out that there are antirealist conceptions of objectivity as well. For instance, Hopster (2017) argues that certain forms of

constructivism can carve a middle path between realist accounts of objectivism and antirealist accounts which reject objectivism, resulting in a kind of antirealist objectivism. And Cohen (2021) draws a distinction between two other forms of objectivism: mind independence and mind nongroundedness. As I observe in **Supplement 3**, characterizing these concepts and distinguishing them from one another is a technical matter, e.g., according to Cohen:

> Grounding theorists generally agree that grounding is an asymmetric, irreflexive, and transitive relation. It is further generally accepted that a ground $\Gamma$ of some fact $p$ has to be such that each member of $\Gamma$ plays some role in noncausally making $p$ the case. Facts that play no role whatsoever in noncausally making $p$ the case cannot be part of what grounds $p$. (p. 183)

I don't even understand most of this, and these remarks from an article adjacent to my own area of specialization. Yet what is clear from articles like these is that Nichols is correct: it would be a considerable philosophical undertaking to develop a satisfactory account of just what it is we're trying to say when we speak of morality being "objective" or attitude-, stance- or mind-independent, with such distinctions breaking down into a variety of distinct concepts, each with a panoply of distinct qualities and implications. Sure, quantum mechanics may be far more complicated. But anyone who thinks metaethics is by any reasonable standard a simple topic characterized by straightforward, categorical distinctions that ordinary people could readily understand is simply unfamiliar with the field.

Yet for some reason philosophers have traditionally seemed convinced that a commitment to realism or antirealism is implicit in the way ordinary people speak and think, and even have beliefs about which metaethical positions are correct. Folk metaethics research isn't about discovering how ordinary people *would* think if they engaged in philosophy, it's about *how they already think*. This presumption enjoys the illusion of plausibility in part due to a lack of familiarity with just how abstract, sophisticated, and distant from everyday considerations various accounts of realism and antirealism actually are. These are highly sophisticated technical theories. It's not plausible ordinary people

endorse any of these views without first learning about them. Yet it is far more difficult to show that a commitment to realism or antirealism is absent from everyday speech. Such a possibility could fly under the radar of conventional survey methods. People need not know the syntactical rules of their language to competently comply with them; likewise, they could be oblivious to their commitment to realism or antirealism if all this amounts to is the best account of their linguistic outputs, without claiming to capture our beliefs. I see little reason to presume such commitments are built into ordinary moral thought. But even if they were, it's unclear whether conventional folk metaethics research would be appropriate for identifying such commitments.

This is because, if our goal is to uncover people's linguistic commitments, survey methods provide only indirect data about our *metalinguistic* judgments. That is, folk metaethics studies, such as Pölzler and Wright's (2020a; 2020b) training paradigms, at best only prompt people to express their theories about how they *think* they (or other people) use words, rather than directly assessing how they actually use those words in practice. Martí (2009) raised this objection with respect to early folk philosophical research on theories of reference conducted by Machery et al. (2004):

> [...] it is important to distinguish carefully between observations that will reveal how people do things (in this case, use names) and observations designed to reveal how they think they do them. The latter will only provide grounds to determine how they are disposed to theorize about their practices, i.e., predict which theories about what they do they are disposed to favour. If we want to test, for instance, whether people use modus tollens when they reason, it may not be the best strategy to ask them, 'John knows that not B, and he knows that if A then B, should John conclude that not A?' That question prompts people to reflect on the principles they regard as correct - their answers will tell us something about their theory of reasoning; but they won't tell us how they really reason. (pp. 44-45)

If Pölzler and Wright's studies are designed to uncover people's linguistic commitments, then it's unclear whether they succeed, since this isn't what they (directly) test for, and we cannot be sure whether people's metalinguistic intuitions about metaethics comport with actual linguistic practice. If, on the other hand, they're designed to capture people's *stances*, i.e., their actual explicit metaethical

beliefs, then these studies would fail for the reasons outlined throughout this dissertation. *Even if* metalinguistic practices *were* an accurate proxy for actual linguistic practice, we'd have no way to confirm this unless *we had some way of directly measuring the metaethical commitments in our linguistic practices in the first place*, which is precisely what we *don't have*. If we did, and the purpose of folk metaethics research were to study metaethical commitments, we'd have no need to do so, because we'd already have the answer! Either way, existing research on folk metaethics provides little in the way of conclusive evidence that most ordinary people are realists or antirealists, or indeed that they have any determinate metaethical stances or commitments at all. In short, just as we have good reason to believe ordinary people have no determinate stances or commitments with respect to quantum mechanics, we likewise have little reason to suppose people have determinate metaethical stances or commitments. At the very least, those who believe people do have yet to furnish compelling evidence to the contrary.

**Limitations**

Although the findings in this chapter support the case against the validity of existing research and lend at least some support for metaethical indeterminacy, they are far from conclusive, and remain subject to a number of objections and limitations. With respect to Studies 1A-1E, one objection is that these studies don't provide direct evidence that participants don't interpret metaethics stimuli as intended. All they show is that participants fail to match the meaning of a given stimulus to some response option intended to reflect the same meaning as that item. There are several ways participants may have chosen "incorrect" responses that are indicative of flaws in the designs of these studies, rather than evidence that the target statements aren't a valid representation of the intended metaethical position.

First, participants were asked to select the statement or statements which are the best "interpretation" of the meaning of the statement. While they were also asked to choose the option which "most closely matches the meaning" of the target statement, *interpretation* can be understood to consist not merely of the task of judging which sentence means the same thing, but to reflect the

*implications* of a given statement, which are often carried by pragmatic considerations. Suppose participants were asked to interpret the sentence "human life begins at conception" in the following exchange:

> Alex: "Do you think abortion is morally wrong?"
> Sam: "Well, I think human life begins at conception."

If asked to interpret Sam, it would be reasonable to infer not only that Sam thinks that "human life begins at conception," but that Sam probably endorses a pro-life stance, and thinks abortion is morally wrong. This illustrates that "interpreting" a statement can go beyond merely identifying its semantic equivalent; it also involves making inferences. People are sensitive to pragmatic implicature; much of what people mean when they say things is carried by implication, and most people are sensitive to e.g., subtext, innuendo, and sarcasm, and recognize that certain claims or turns of phrase can hint at a person's beliefs, values, or motivations. To illustrate, do you think Sam is more likely to be a Christian or an atheist? I suspect most readers would presume Sam is more likely to be a Christian, despite nothing about the meaning of "I think human life begins at conception," necessarily entailing that Sam thinks that e.g., God exists, or that Jesus rose from the dead.

Participants were not directly asked about the implications of a given statement, and instructions did ask them to choose responses which *matched* the meaning of the response option, but instructions may still have misled some participants, and participants may still have felt that it was appropriate to select response options whose content was implied rather than directly conveyed by the target statement. Without knowing how participants interpreted the task, it's hard to know how much of an impact this possibility had on responses. Yet if this problem were present, it's unclear why the effect would be so pronounced and ubiquitous that it would so overwhelm intended interpretations and efforts to match the meaning of statements that the proportion of participants who chose the intended interpretation was often marginally better than chance.

Furthermore, response options were designed to either restate the item or to convey implications or entailments, but were mostly designed to convey what I took to be plausible interpretations of the meaning of the original statement. The response options were not pulled out of a hat, but were crafted to reflect the kinds of conflations I expected people to make, both due to armchair consideration of the kinds of ambiguities that arise in metaethics stimuli, and due to the kinds of responses people given when they were explicitly asked to explain how they interpreted metaethics stimuli. Even so, I cannot rule out the possibility that people didn't interpret the meaning-matching paradigms as intended. It remains an open, and amusingly ironic possibility that I have failed to determine whether people interpreted stimuli as intended because they didn't interpret the stimuli in my studies as intended. Since my assurances that I consider this unlikely don't count for much, I'll leave it to future research to refute or corroborate my findings.

Another limitation with these studies is that they were all conducted on MTurk, among participants who had little incentive to take the task seriously. Pölzler (2021) has shown that insufficient effort responding (IER) can pose a significant threat to the validity of studies in experimental philosophy, and it is worth noting that Pölzler is one of the only people whose research focuses primarily on folk metaethics.[94] Since these studies were incredibly short, this buffers them against IER, but the difficulty of the task may have exacerbated IER. Since around 13-20% of participants failed the comprehension checks in these studies, it's plausible that IER threatened the outcome of these studies, though all results remained significant when excluding participants who failed comprehension checks. However, little research has directly addressed the role comprehension

---

[94] Pölzler estimates IER to account for 10% of responses in self-report surveys in general, yet as low as 5% IER is sufficient to threaten study results. It's less clear what the overall rate of IER is in experimental philosophy, but Pölzler points to a variety of measures that are roughly consistent with around 10% IER, e.g., 10-18% of participants failed various comprehension checks and attention checks across a variety of studies. Around 13-20% of participants failed the comprehension checks in studies 1A-1D, which is roughly comparable.

checks play in assessing IER, so it's unclear how helpful these results are in estimating the presence of IER.

One potential concern with Study 2 is the possibility that many people really do have determinate stances towards the Copenhagen and Many Worlds interpretations. It's possible that quantum mechanics has become a sufficiently popular topic that it has permeated popular culture. Indeed, the study was launched in June of 2022, just one month after the release of *Doctor Strange in the Multiverse of Madness*. Shows like *Star Trek* and *Rick and Morty* have likewise familiarized people with the notion of multiple universes. In fact, the first episode of Season 2 of *Rick and Morty*, "A Rickle in Time," explicitly involves a state of quantum uncertainty resolving by branching into multiple universes. While it's possible enough ordinary people are familiar with different interpretations to have a determinate stance, this isn't a plausible explanation for the results. First, after excluding participants who reported understanding quantum mechanics or having formal training in physics, 22.4% (n = 61) of participants reported that they were familiar with the Many Worlds interpretation, while only 8.5% (n = 23) reported familiarity with the Copenhagen interpretation. These are hardly reassuring numbers. Furthermore, mere familiarity with the *terms* is hardly sufficient to have a view on the matter. Most importantly, almost all participants explicitly stated that they had no position. Yet even if people did report having a position, we should interpret such claims with caution. Pop cultural references to quantum mechanics are hardly adequate to furnish people with a sufficiently substantive or accurate understanding. In fact, the prevalence of cultural references to quantum mechanics overlaps in some ways with references to metaethics that occasionally trickle outside of academia, as when politicians or religious leaders speak of the disastrous consequences of "moral relativism." In both cases, popular understanding of these concepts is likely minimal and riddled with misconceptions. Such misconceptions and associations may even play a significant role in explaining *why* responses to

questions for both topics are nonrandom: people *think* they have some sense of what various technical terms mean, even if they don't.

Finally, while I do not develop on these points in much detail, there remains the question of whether or not, and to what extent, these findings support metaethical indeterminacy. I will address this in greater detail in the conclusion, and will focus here on how these findings support indeterminacy. Studies 1A-1E illustrate that, even when participants are presented with a variety of metaethics stimuli, they reliably fail to match the meaning of these items with the intended metaethical meaning.

**Future directions**

IER may have been a significant factor in the poor performance for Studies 1A-1E. Future studies should address the same or similar questions using methods that increase effortful responding, e.g., incentives for getting the correct response. Such methods could also be expanded to include additional metaethics paradigms that have yet to be tested. It's also possible that people perform poorly with these tasks in general, which would suggest flaws with the method rather than flaws with the stimuli. There are two ways to assess this. First, we could provide the same stimuli to professional philosophers, and see how well they perform. If they perform much better, this would indicate that training does allow people to identify the sentences which match the meaning of the target statement. We could also employ similar matching exercises for topics we expect ordinary people to be competent with, as a control. Once again, if the response rates are higher in these cases, this would suggest that there are distinctive difficulties with interpreting metaethics items as intended.

This method could also be adapted for assessing interpretation in other nonmoral domains, e.g., judgments about taste preferences, aesthetics, epistemic norms, social conventions, and so on. It could even be adapted to other philosophical questions, or to psychological stimuli more broadly. Note, however, that what I've done essentially amounts to glorified comprehension checks. I would

not suggest so much that researchers begin employing the meaning-matching paradigm as a standalone method. Rather, I would suggest instead that researchers follow at least three general guidelines:

(1) *Be more careful about designing stimuli.* This is so general as to be nearly worthless, but the general point stands. Researchers are often inattentive to potential interpretative difficulties with their stimuli. As *G.I. Joe* has taught us, knowing is half the battle. We can't begin to address interpretative difficulties if we're not aware of the problem, and make an active effort to think about the role pragmatics and ambiguity can prompt unintended interpretations. Researchers should also be more wary of the curse of knowledge (Birch et al., 2017) which is difficult to suppress, and the methods employed are often inadequate to mitigate its influence. For instance, Collier-Spruel et al. (2017) had a panel of *experts* evaluate items intended to represent relativism. While these experts may have been good at judging how well these items represented relativism, they did not judge how *nonexperts* would interpret these items, nor would they be an appropriate pool of judges to do so. The only way to know *that* would be to conduct the proper empirical tests. This brings me to the second piece of advice

(2) *Pretest items to assess interpretability.* Researchers should employ methods like the ones reported here *prior* to conducting more extensive research. It's absurd to rely on a set of measurement tools without finding out if those tools actually work. Relying on previously validated measures and scales will not be adequate, since a scale validated in one population may not be valid in another, and purported evidence of validity is often inadequate to ensure that the measures in question actually exhibit good validity (Hussey & Hughes, 2020; cf. Wetzel & Roberts, 2020). Furthermore, researchers often modify or adapt items to a new population without retesting validity (Flake, Pek, & Hehman, 2017).

At the risk of being impertinent, I'll just come out and say what I actually think: researchers have historically been *far too careless* about employing decent measures. They've largely coasted on identifying some purportedly validated measure (or not, as they often just make up *ad hoc* measures on the fly), plopping it into a study, citing some paper that allegedly validated the measure that they

probably didn't read (and probably didn't provide good evidence of validity anyway, since traditional validation procedures are inadequate; see Maul, 2017), and declaring victory. Mindlessly copying-and-pasting measures or conjuring a measure on the fly is barely scientific. There are probably astrologers that put more work into making predictions. In any case, when the topic in question is subtle or complicated, as is often the case with folk philosophy, researchers have an especially strong incentive to pretest items to assess how well people interpret them, and to modify them accordingly. There are already methods for doing this (Hunt, Sparkman, & Wilcox, 1982; Reynolds, Diamantopoulos, & Schlegelmilch, 1993). However, these methods may be insufficient for research on folk philosophy. We may need to develop more rigorous and distinctive methods for pretesting questions related to philosophy, and be open to the possibility that in some cases there aren't psychological constructs that function as analogs to a philosophical account. Finally, pretesting may fall short of ensuring that everyone interprets philosophical questions as intended. This brings me to the final suggestion.

(3) *Include better comprehension checks.* In addition to pretesting, researchers should include comprehension checks in their studies. Yet it's not enough to employ simple multiple choice questions. The comprehension checks that appear in studies often demand little more than minimal attention to the instructions. For instance, the comprehension checks I employed in the quantum mechanics studies simply require recalling the instructions presented earlier in the survey. People can recall such information even if they have an incredibly superficial understanding of the stimuli. In other stances, researchers may collect open response data, only to pass analysis off to people who lack the training, expertise, or appropriate level of skepticism to analyze responses (Bush & Moss, 2020).

Future studies should also be directed at developing on the concept of spontaneous theorizing and devising novel methods for assessing if, when, and how it occurs. I'm not sure what form these studies would take, but providing more robust evidence that spontaneous theorizing can occur, and that it can explain the results of studies in folk philosophy or more generally could go some way

towards recognizing a novel methodological problem that could impede progress in psychology. At a minimum, researchers could begin by more carefully assessing whether participants themselves report having held no position on a topic prior to the study, but developing one over the course of it. Indications of this occurring could perhaps be achieved by e.g., cognitive interviewing (Beaty & Willis, 2007; Drennan, 2003; Madans et al., 2011; Willis, 2004), which incidentally may also be valuable for designing folk philosophical studies in the first place.

**5.3 Conclusion**

I had three goals in this chapter: to provide further evidence for the invalidity of existing measures of folk metaethics, to provide an illustration of spontaneous theorizing, and to provide evidence that nonrandom response patterns can emerge from surveys even when we have strong theoretical grounds for presuming indeterminacy. All three goals were achieved. These findings bolster my argument that existing folk metaethics research relies on invalid measures. In addition, I have shown that studies can yield nonrandom response patterns even if people held no prior stance on an issue. This illustrates the possibility that studies can yield replicable patterns of results that give the appearance of capturing the intended psychological phenomena, even if they do not. I have provided two explanations for this outcome. First, some participants may engage in spontaneous theorizing, developing a position in the course of a study that does not reflect a position they held prior to the study. Second, people may systematically favor one response over another for reasons unrelated to it, accurately representing their position on the matter as a result of the forced choice design typical of most research in folk metaethics. Neither of these explanations are mutually exclusive, but may work in tandem to produce artifactual findings.

While these findings do not directly demonstrate metaethical indeterminacy, they show how existing studies could create the *illusion* of a determinate folk metaethics even if none existed. Given the absence of strong theoretical foundations for supposing that there is a determinate folk metaethics,

the onus should be on those who maintain that ordinary people are realists or antirealists to provide adequate evidence for this claim. Such evidence would, at a minimum, require valid measures that don't prompt spontaneous theorizing. At present, no studies have met these conditions.

## CHAPTER 6:
## Third Person Paradigm & Normative Entanglement

### 6.0 Introduction

One of the most common ways ordinary people interpret questions about metaethics in unintended ways is by interpreting them as questions about their *first-order*, or *normative* moral standards (these terms may be used interchangeably). In other words, when asked a question intended to determine whether the participant endorses moral realism with respect to a particular moral issue, such as lying, participants instead interpret the question about whether they think lying is morally right or wrong. Pölzler (2018b) drew attention to this methodological concern in a critique of research on folk metaethics that focused primarily on the disagreement paradigm. According to Pölzler:

> Another problem with many studies on folk moral realism is that they have not sufficiently accounted for moral realism and anti-realism's moral neutrality. In particular, researchers have attempted to measure subjects' intuitions about these views by using scenarios, questions and answer choices that may also have prompted strong first-order moral intuitions. Subjects' responses in these studies may consequently partly be explained by these intuitions, rather than by their views about moral realism and anti-realism alone. (p. 657)

Early versions of the disagreement paradigm are especially susceptible to this concern. For instance, after asking participants about their own first-order moral stance towards a moral issue, then telling them that another person disagreed with them, Goodwin and Darley (2008) asked participants to select one of the following response options:

*(1) The other person is surely mistaken.*

*(2) It is possible that neither you nor the other person is mistaken.*

*(3) It could be that you are mistaken, and the other person is correct.*

*(4) Other.*

Participants may have interpreted this as a question about their moral stance towards the issue in question, *not* their metaethical stance. Suppose, for instance, you are a moral relativist, and your moral

claims reflect your subjective moral standards. When asked whether someone who holds contrary moral standards is *mistaken*, you could say *yes* as a way to express your normative moral stance that you reject contrary moral standards, *not* to convey the view that people with contrary moral views fail to endorse the stance-independently correct set of moral facts. I refer to these unintended interpretations *normative conflations*. Normative conflations occur whenever someone interprets a question or statement about metanormative (second-order) concepts to instead be a question about or statement about normative (first-order) concepts. In this case, there is no need to speculate. In **Supplement 4**, thematic analysis of Goodwin and Darley's data revealed that normative responses (rather than metaethical responses) were the most common theme for unintended interpretations in both of the studies for which their data was available (19% of responses in Study 1 and 17.8% in Study 2). In other words, I have already provided empirical evidence that normative conflations are (in at least some cases) very common.

One indirect line of evidence supporting the ubiquity of normative conflations is the consistent correlation between measures of the strength of people's first-order moral positions and their metaethical positions towards the same moral issue (Pölzler, 2018b). In other words, the more strongly people reported that a given issue was morally right or wrong, the more likely they were to also endorse the realist response option towards that same issue (e.g., Beebe & Sackris, 2016; Beebe et al., 2015, Goodwin & Darley, 2008). Overall, a majority of studies have found a correlation between attitude strength and realism (though Wright et al., 2013 found no significant correlation between attitude strength and realism). Such findings provide some circumstantial support for the claim that normative conflations are common. It's possible that greater attitude strength predicts greater propensity for normative/metaethical conflation. For instance, one possibility is that attitude strength is associated with a greater affective response, which could result in an emotional bias that influences how participants respond to what are intended to be questions about their metaethical positions.

Consider, for instance, two questions about a person's metaethical stance: whether they're a realist about the moral status of (a) speeding on the highway and (b) kidnapping children to harvest their organs. A person may have a more sober reaction to (a) than to (b), regarding the latter as far more evocative and disturbing. As a result, when asked about their metaethical stance towards these issues, if they're presented with an opportunity to express that someone who disagrees with them, and thinks speeding on the highway and kidnapping children to harvest their organs are "not wrong," they may be more inclined to judge that this person is mistaken in the latter case, not because they're a realist about the issue, but because they're more motivated to express their *normative* opposition to (b).

While this explanation would characterize normative conflations as a form of bias, normative conflations could also be a byproduct of a rational tendency to prioritize managing one's reputation rather than responding to questionnaires as researchers intend. If we think about our own experiences, we can probably recall times where people made an explicit point, or go out of their way, to express their condemnation for a particular action, or to disavow some person or deed. Such instances are often ones in which a person's reputation is on the line, and they deem it important to distance themselves from the act in question. People may internalize a disposition to manage their reputation in this way and, as a result, be more inclined to opt to express a normative stance towards highly egregious moral transgressions most people would oppose, and less disposed to do so when the moral issue in question is controversial or not severe.[95]

Far from indicating a bias or error, such responses would instead indicate an emphasis on ensuring one does not say things that could harm their reputation over a literal but socially flat-footed style of response where we ignore the potential reputational harms that could come to us by answering questions in particular ways. This sociofunctional explanation may account for normative conflations

---

[95] Drawing on the same example, expressing opposition to kidnapping and harvesting organs may be more important for managing your reputation than condemning speeding, since giving the impression that you don't find the former repugnant would be far more counternormative and reputationally costly than a failure to express opposition to speeding.

not by attributing them to a performance error, but to a practical approach to responding to questions that prioritizes maintaining a positive reputation over responding in the most literal, direct, or philosophically sophisticated way possible, regardless of the consequences.[96]

One line of evidence consistent with this hypothesis is a reliable and strong correlation between realism and perceived consensus (Goodwin & Darley, 2012). In fact, when Goodwin and Darley experimentally manipulated the alleged consensus on a moral issue, this caused an increase in realist responses. One explanation for these findings is that people use consensus as *evidence* of realism, i.e., if most people agree about a particular moral issue, people may take this to be evidence that there is a stance-independent fact about that issue, or at least that there is a single correct answer (Ayars & Nichols, 2020). However, it is also possible that perceived consensus tracks perceived social stakes: the more there is a general consensus about the moral status of the action, the greater the cost of expressing antirealism towards that stance, since doing so could fail to convey a strong and intolerant stance towards people with contrary moral views.

To illustrate, consider how people might react to a person who expresses relativism towards a low consensus moral issue, such as euthanasia. A person who expresses relativism about this issue may be perceived as having a tolerant and permissive attitude towards those with contrary views. Whatever we think of such a permissive attitude towards euthanasia, consider instead a person who expresses relativism towards genocide. If this is perceived as indicating a tolerant or permissive attitude towards genocide, others may recoil at such an attitude, and respond with disgust and outrage. At least, they may do so to a comparatively greater extent than they would towards someone who holds a contrary view towards euthanasia. This illustrates that character judgments about a person expressing

---

[96] Such responses need not even be motivated by any awareness of such a prioritization. One's interpretation of such stimuli could in principle occur in a way that is introspectively inaccessible and precedes conscious reflection over one's interpretation of the question.

a particular metaethical position towards a given moral issue may vary in accordance with perceived consensus towards that issue (and, for that matter, perceived seriousness).

Regardless of *why* normative conflations may occur, a correlation between attitude strength and metaethical stance only alludes to a potential conflation, but does not directly establish it. My findings, on the other hand, *do* strongly indicate that normative conflations are common, at least for certain metaethics paradigms. While **Chapter 4** focuses on intended and unintended interpretation rates, the accompanying **Supplement 4** employs thematic analysis to categorize responses in accordance with a variety of ways in which participants interpreted questions about metaethics. The *normative* theme captures those instances in which participants appeared to express a normative moral stance rather than a metaethical position. The *normative* theme was one of the most prominent and recurring themes across a majority of studies, appearing in the top five most common themes in at least one condition in studies 1, 2, 4, 5, and 6 (and only failing to appear in the top five in studies 3 and 7). I'll provide a handful of examples to illustrate instances in which participants offer responses that appear to express a normative rather than metaethical rationale for their answers. In study 4B, participants were asked:

> *In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is objective?*

Here are a handful of participant responses classified as *normative*:

> *To say that denying someone else's equally valid claim to life as that of the murderer would be morally wrong.*

> *that murder is a bad thing*

> *Murder IS morally wrong, it's taking the life of someone who has more life to live. It's taking their right to live their full life.*

When directly asked what it means for morality to be "objective," these participants expressed what appear to be normative claims about the moral status of murder, *not* claims about the stance-independent of murder. A similar response pattern emerged in most other studies that appeared in

**Chapter 4**, with the *normative* theme often being the first or second most common theme, including (as previously noted) Study 1A (19% of responses) and 1B (17.8% of responses), as well as 4B (21.1% in the realism condition, 25.6% in the relativism condition), 5A (11.5%), and 5B (15.4%). People consistently expressed a normative moral stance in response to a variety of questions, including what they thought the source of disagreement was between them and a previous participant (Studies 1A and 1B), asking the participant what someone who expressed a realist or antirealist stance meant (Studies 2A and 2B), directly asking participants what they think it means to say that "abortion is morally wrong" or "murder is morally wrong" is objective or relative (Study 4B), and asking participants to explain their level of agreement with items that appear on metaethics scales (Studies 5A, 5B, 6A, and 6B).

While the normative theme did not always emerge as one of the most common themes, it frequently appeared in a variety of paradigms. This illustrates both the ubiquity and generality of the tendency for people to appeal to normative moral considerations in studies intended to assess their metaethical stances. Such findings provide considerable support for the conclusion that normative conflations are quite common. Indeed, in subsequent work, Pölzler (2018a) and Pölzler and Wright (2020b) repeated their concerns with the potential for studies to prompt conflations between normative and metaethics, concerns which culminated in efforts to partially circumvent the problem through the use of training paradigms (i.e., paradigms that provide extensive instructions or exercises intended to familiarize participants with the relevant metaethical terms and concepts). In spite of optimism about the prospects for more rigorous methods of evaluating folk metaethical stances and commitments, Pölzler (2018b) concedes that "Avoiding first-order moral intuitions in studies on folk moral realism altogether may be methodologically infeasible" (p. 658).

I am sympathetic to Pölzler's concern. It may not be possible to completely separate metaethical and normative considerations when evaluating how ordinary people think about

metaethical questions.[97] However, we can address normative conflations and the potential associations between folk conceptions of metaethics and normative ethics in at least two ways:

(1) to the extent that such conflations occur, we can attempt to understand *why*, and (2) we can attempt to mitigate the influence of normative conflations by devising alternative paradigms that are less likely to prompt normative conflations.

One notable feature of all previous metaethics paradigms, is that they attempt to directly evaluate the *participant's* metaethical stance, and often do so while evaluating their normative stances as well, often simultaneously or in near proximity to one another. For instance, participants in a typical study employing the disagreement paradigm will be asked to express their level of agreement or disagreement with a moral claim. Then, they will be asked whether someone who disagrees with them is mistaken or is also correct. The use of concrete moral issues may exacerbate the degree to which people are disposed towards normative conflations. In addition, by putting people in situations in which they are asked to consider *normative* moral disagreements with others, people may retain a desire to express their opposition to people with conflicting normative moral standards. Consider an ordinary person asked to consider a dispute about the moral status of abortion. When confronted with a person who holds a contrary position, the importance of signaling one's opposition to a contrary position on abortion may be more important than expressing a position on the metaethical status of moral claims about abortion. In other words, such scenarios are especially prone to amplifying the social incentives participants have to signal their normative moral stances. For ordinary people, everyday moral judgment isn't merely a matter of expressing the correct or most logically consistent

---

[97] It is also unclear whether normative conflations are an artifact of limitations in experimental design, or reflect substantive features of the way ordinary people think. In other words, it could be that ordinary people can and do distinguish metaethical and normative considerations, but the way studies are conducted fail to properly disentangle and independently elicit people's metaethical and normative judgments. Alternatively, it could be that ordinary moral psychology is structured in such a way that the two are inextricably linked (for at least some people, in some populations). If so, normative conflations may be a feature of ordinary moral psychology, *not* a methodological artifact. While philosophers may draw a conceptual distinction between metaethics and normative ethics, this is no guarantee that ordinary people distinguish them from one another, at least not without training.

judgment, it is a matter of maintaining one's status both within their society at large and within one's ingroup. In the former case, this is achieved by expressing condemnation of antisocial or universally condemned behaviors, and in the latter it is achieved by demonstrating convincing display of loyalty and commitment to the group's, and a rejection of rival moral standards.

Even when participants are asked in abstract terms about their metaethical standards, and they are not also asked about their normative moral standards, as in e.g., Collier Spruel et al.'s (2019) moral relativism scale, which does not include questions about specific moral issues, normative conflations may *still* be present. This is because metaethical claims, even when they are expressed in abstract terms, can prompt normative concerns, insofar as a particular metaethical stance may be taken to signal or imply a normative moral stance, or to more generally imply aspects of a person's character that are not strictly entailed by a metaethical stance. In particular, the *expression*, if not a genuine commitment to or endorsement of, a particular metaethical position may serve one or more sociofunctional purposes, e.g., expressing relativism may signal one's tolerance for others, while expressing realism may signal one's commitment to their values. Conversely, expressing a relativist could signal a lackadaisical attitude towards morality that could make one seem noncommittal and wishy-washy, while expressing a realist stance could lead others to infer that you are dogmatic, rigid, or inflexible.

Previous research is consistent with these suggestions. Collier-Spruel et al. (2019) found a strong significant correlation between responses to their metaethical relativism and tolerance scales ($r$ = .56). This indicates that the stronger people endorse relativism, the more they express tolerance for people with contrary moral views. A slate of other studies likewise reveal that greater endorsement of relativism is associated with greater tolerance for people with opposing moral standards (Wright, Cullum, & Schwab, 2008; Wright, McWhite, & Grandjean, 214; Wright & Pölzler, 2022) and openness to experience (Feltz & Cokely, 2008). Moreover, Goodwin and Darley (2012) found that realist responses were associated with reduced willingness to reconsider one's moral beliefs, greater

discomfort with people with different moral values, and a tendency to regard people with conflicting moral standards as more immoral. Taken together, these findings suggest that, at the very least, relativist responses are associated with tolerance, while realist responses are associated with discomfort, negative moral evaluation, and close-mindedness.[98]

These findings suggest that there may be a genuine tendency for relativists to be more tolerant than realists. Of course, they don't directly establish that people perceive others who express relativism or realism towards a moral issue to be more tolerant or intolerant. Yet the fact that such an association exists points to the possibility that people anticipate how others would perceive their response to questions about metaethics, and that this in turn influences what metaethical position they express or adopt. In other words, it could be that, to the extent that people anticipate, consciously or otherwise, how others might perceive the expression of a particular metaethical stance, they may strategically eschew or employ such language in accordance with their social goals. If so, "realist" and "relativist" responses could be caused by one's normative moral attitudes, including their tolerance for opposing moral views, but it could also be the result of the strategic adoption of locally appropriate language for the purposes of managing one's reputation or persuading others (Mercier & Sperber, 2011). In sum, there may be a variety of factors prompting normative conflations: ordinary people may interpret questions about metaethics as questions about their normative standards, they may wish to signal a genuine commitment to particular normative moral standards, or they may be motivated to express

---

[98] Of course, previous chapters have raised substantial doubts about the validity of the measures used to identify people as realists or relativists. Such concerns are less of an issue here because my primary concern is with whether people *perceive* an association between metaethical views and moral attitudes. It could be that people who are more disposed towards tolerance are more likely to favor relativist-sounding responses, while those less disposed towards tolerance favor realist-sounding responses, not because relativists are more tolerant than realists, but because more tolerant people favor more relativist-sounding responses, *even if* they don't interpret them as intended. That is one advantage of the present set of studies: at the very least, they can establish whether people's *responses* to questions about metaethics are associated with their normative moral standards, even if we remain skeptical about the validity of those measures. Indeed, to the extent that responses to questions ostensibly about metaethics are associated with a variety of normative moral attitudes, character traits, and so on, this could be leveraged to corroborate doubts about the validity of metaethics measures. But even if such concerns are set aside these results would still be informative.

normative moral standards rather than their metaethical views to achieve various practical social goals, such as managing their reputation or persuading others.

Whatever factors drive normative conflations, it may be possible to minimize the risk of these conflations by asking participants about what they believe *other* people's metaethical standards are. This approach could mitigate normative conflations by providing participants with the opportunity to assess what metaethical standards they believe typify the way ordinary people speak or think without entangling these judgments with their personal moral standards and commitments. This could reduce both the risk of unintended interpretations and the risk that participants would believe their responses to questions about metaethics could signal additional information about their normative moral standards, character, or other traits.

The first goal of this chapter is to develop just such a paradigm. I implement a *third person paradigm* that the participant's judgments about the metaethical standards of a typical person in their society rather than their own metaethical views. The third person paradigm also exhibits a number of other advantages. Part of the goal of traditional philosophical approaches to descriptive metaethics is to assess the public meaning of moral claims. Philosophers are not generally interested in each individual's proprietary ways of thinking and speaking, but in the best externally adequate account of ordinary moral discourse (see e.g., Sinnott-Armstrong, 2009). A third person paradigm may provide more direct evidence of such an account than other paradigms (e.g., the disagreement paradigm), thereby yielding results with greater relevance to philosophers. Such an outcome could be of more immediate interest to psychologists as well, since it's possible this approach could minimize the risk of spontaneous theorizing (i.e., when participating in a study causes a participant to express or develop a view they did not hold prior to participation). Participants encouraged to reflect on their own views may be prompted to reflect in ways that result in novel or spontaneous judgments that don't genuinely reflect how they or others tend to think outside the experimental context in which the study is

conducted. In contrast, judgments about how a typical person thinks or speaks about metaethics may be a better reflection of typical usage. If so, third person paradigms may have greater external validity than standard first-person paradigms.

It's also possible that the third person paradigm could find a similar pattern of endorsement for realism and antirealism as existing measures. If so, this could reveal that many of the concerns expressed here are unfounded, which could partially vindicate existing paradigms. If, on the other hand, the overall pattern differs, this would suggest either a shortcoming with the third person paradigm or further substantiate concerns with the validity of existing measures. I test for this by presenting participants with first-person versions of the same questions, which allows for a direct comparison of first-person and third-person responses. Studies 1 and 2 address these questions in the moral domain by assessing what metaethical inferences people make about a typical person in their society, either in the absence of any information about that person at all, or given that the person has made a moral assertion that an unspecified action is morally right or wrong (Study 1), or that a particular concrete moral issue is morally wrong (Study 2).

Finally, previous studies employing the disagreement paradigm have also used questions in nonmoral domains, including taste/aesthetics, social conventional, and factual claims about e.g., science and history (Goodwin & Darley, 2008; 2012; Wright, Grandjean, & McWhite). I employ a third person paradigm for these same nonmoral domains in Study 3. Since normative conflations are less plausible in these domains, such findings may also reveal a similar pattern of results in these domains when compared to first person responses but a different pattern of results when comparing third and first person moral judgments, a disparity that could raise questions about the degree to which certain methodological problems may be distinctive to moral questions, further bolstering the case for the role normative conflations may play in threatening the validity of first-person metaethics paradigms.

If nothing else, this approach provides a novel way to assess how ordinary people think about metaethics that could yield insights inaccessible to existing paradigms. This is especially important, since attempts to assess realism and antirealism in nonmoral domains have received comparatively little attention. The attention that been directed at them has also typically been perfunctory and superficial, since these domains are treated as control groups or as ancillary to questions about metaethics, and often rely on crude adaptations of measures intended for metaethics, rather than measures carefully designed specifically to assess realism and antirealism in that particular domain. Of all the nonmoral domains, folk realism and antirealism about aesthetics has received the most attention (e.g., Bonard, Cova, & Humbert-Droz Cova & Pain, 2012; Cova, Garcia, & Liao, 2015; Cova et al., 2019; Rabb et al., 2020). Unfortunately, as Moss and I argue in Moss and Bush (2020), the measures that appear in Rabb et al. (2020) suffer especially poor validity, in that their instructions fail to adequately distinguish realism and antirealism. Other problems are also common, including misleading or biasing instructions and the outright misoperationalization of the relevant constructs, both of which threaten the validity of at least some studies on folk metaaesthetics in ways that are even more serious than the threats to folk metaethics. This further highlights the need for novel paradigms that do not suffer from as many methodological deficiencies.

Given the poor validity of existing measures, and the need to devise new measures in the absence of substantive and reliable evidence about folk metaethics, my goals are largely exploratory. Nevertheless, consistent with the most rigorous and recent studies (in particular, Pölzler & Wright, 2020a; 2020b), I expect people to judge a typical person to lean towards relativism, and to be less disposed to endorse realism and universalism.[99] I have no meaningful expectations about

---

[99] This expectation is at least partially based on more recent and more comprehensive studies on the psychological of metaethics, which tend to find that a majority of participants favor antirealist positions (Davis, 2021; Pölzler, Zijlstra, & Dijkstra, 2022; Pölzler & Wright, 2020a; 2020b). In addition, my studies draw on a participant pool that is disproportionately likely to be nonreligious (Levay, Freese, & Druckman, 2016), a trait negatively associated with endorsing

noncognitivism, though it's worth noting that when Beebe (2015) introduced a noncognitivist response option to the disagreement paradigm it was frequently selected, while Davis (2021) found noncognitivism to be the modal response using a modified version of the disagreement paradigm (cf. Pölzler & Wright, 2020a).

I am skeptical of such results, however, since judging that moral claims can be neither right nor wrong could serve as an epistemic hedge that avoids committing one to a *normative* stance on a given issue; as such, I suspect people may not be interpreting this response option as intended, or may not be selecting it as a genuine reflection of their metaethical stance. In addition, I introduce distinct measures of both realism and universalism. Some researchers have noted that realism and universalism are distinct positions, and have opted to measure the latter instead of the former (Ayars & Nichols, 2020; Rose & Nichols, 2019). My goal in assessing both is largely exploratory. Findings may indicate that people make little or no distinction between the two, or people's responses to realism and universalism could turn out to vary.. I also introduce measures of character to assess how participants perceive the character of people who make particular normative moral claims.

Although studies 1-3 assess whether participants think a typical person in their society is a realist or antirealist given a normative claim (or at least a first-order claim, in the case of the factual domain), Study 4 departs from the third person paradigm in a critical way: participants are asked to judge the character of a person who expresses a metaethical stance, effectively reversing the approach taken in studies 1-3. In other words, the third person paradigm asks people to infer a person's metaethical stance or commitment (either in the absence of any information, or given the target of their judgment's normative moral stance), while Study 4 examines the participant's normative moral evaluation of a person given that person's metaethical stance.

---

moral realism (Goodwin & Darley, 2008) and positively associated with endorsing moral relativism (Collier Spruel et al., 2019).

Study 4 has a different aim. It tests a specific hypothesis about why some people may be disinclined to express antirealism towards some moral issues: namely, that doing so may pragmatically imply that they are less opposed to serious moral transgression and lack an appropriate emotional attitude towards them (i.e., disgust or outrage). If participants perceive a person who expresses an antirealist stance to have similar or equal moral character to a person who expresses a realist response only when they *cancel* whatever implicature or pragmatic inference may be associated with expressing such a stance, but not when they don't, this would support the possibility that participants perceive expressing antirealism as potentially signaling undesirable character traits, a factor that could motivate participants to avoid expressing antirealism due to its anticipated reputational consequences.

## 6.1 Study 1: Third person paradigm (pilot)

Study 1 introduces the third person paradigm. In this initial version of the paradigm, I devised three conditions: a condition in which participants were asked to judge the metaethical beliefs of a typical person in theirs society (the *no statement* condition), and two conditions in which they were asked to judge the metaethical beliefs of a person who either asserted that an unspecified action was morally right (*abstract right* condition) or morally wrong (*abstract wrong* condition).

Although my primary goals were largely descriptive and exploratory, I predicted that judgments about whether another person was a realist or antirealist would vary significantly across conditions and, in particular, that people would be least likely to judge another person to be a realist in the no statement condition, most likely to judge others to be realists in the abstract wrong condition, and that the abstract right condition would be intermediate between the two. This is because a person who has made no statements at all provides little or no information about how they think about morality, so participants must infer what such a person is like based on their prior knowledge about what people tend to be like. However, the moment a person makes a normative moral assertion that a given action is right or wrong, this provides *some* information that they take a stance on moral issues,

which may increase or render salient the notion that they also think there are objective or universal facts about moral issues. Counterfactually, one might reason that a person who wouldn't state that an action is morally right or wrong would be less likely to think an action is *objectively* or *universally* right or wrong, given that they've provided no indication that they think anything is morally right or wrong at all (objectively, universally, or otherwise). Someone who asserts that an action is morally wrong may also be seen as especially judgmental, since someone who asserts that a given action is right could, in comparison, appear more morally permissive.

Aside from these predictions, another goal was to assess how various metaethics measures introduced in this study performed in the context of the paradigm, since these same measures were initially devised for a conventional self-report scale. I also wanted to see whether responses to the third person and first person versions of these measures would differ, though I didn't have any specific predictions. For instance, I wanted to assess whether mean scores on the various subscales (realism, universalism, relativism, and noncognitivism) roughly approximate how people responded when asked to report their own metaethical standards, or whether there would instead be meaningful differences between third person and first person measures. Finally, I introduced both realism and universalism as separate measures, in order to assess whether or not, and to what extent, people might distinguish the two.

**Methods**

***Participants.*** I aimed to recruit 400 participants. 402 participants began the study.[100] However, 12 participants did not complete the study and were excluded from further analysis. This left a total of 390 participants. Participants consisted of 390 adult US residents on Amazon's Mechanical Turk (185 females, 203 males, 1 other, 1 unreported, $M_{\text{age}}$ = 39.8, $SD_{\text{age}}$ = 11.7, age range = 21-73).

---

[100] The collection of participants beyond the intended sample size is a quirk of the recruitment process on Amazon's Mechanical Turk, which were recruited through CloudResearch.

**Procedure.** All participants were given the same initial set of instructions, which asked them to imagine a typical person in their society. Then they were randomly assigned to one of three conditions: (1) a *no statement* condition, (2) an *abstract "wrong"* condition, or (3) an *abstract "right"* condition. In the no statement condition, participants were given the following additional instructions:

> *Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

In the *abstract "wrong"* and *abstract "right"* conditions participants were instead asked to consider a scenario in which a person was having a discussion about moral issues and, during the course of this discussion stated, "That is morally wrong," or "That is morally right," respectively. After reading this scenario, participants were asked to:

> *Think about the person in this scenario. Please rate how likely or unlikely it is that they hold the following beliefs.*

After this, participants were asked a set of twelve questions about the metaethical standards they believe this person would have. After completing these measures, participants were then told that they'd be asked what they thought about the statements they were just asked about. They were then asked to express their level of agreement with each of the same twelve items and to answer demographic questions.

**Measures.** All participants judged the typical person in their society's metaethical views by expressing how likely it was that the person endorsed realism, universalism, relativism, and noncognitivism 1 (Very unlikely) to 7 (Very likely).

Realism reflects the view that there are stance-independent moral facts about what is right or wrong. This means that moral facts are not made true by people's attitudes or stances. For example, one realism item states, "They believe moral truth is independent of cultural standards and personal beliefs."

Universalism is the view that there is a single correct moral standard (Ayars & Nichols, 2018).[101] Universalism is conceptually distinct from realism. At the risk of oversimplification, universalism concerns *how many* correct sets of moral standards there are (just one), while realism concerns whether moral truths are stance-independent, i.e. *what makes them true* (*not* people's stances).[102] Items representing universalism were intended to capture this distinction, e.g.: "They believe there is a single set of moral standards for all people and cultures."

Relativism holds that moral claims are only true or false relative to the standards of different individuals or groups. One such item stated that "They believe that things are only morally right or wrong according to different points of view."

Finally, cognitivism is the view that moral claims *propositions* (sentences that can be true or false), while noncognitivism holds that they are neither true nor false because they aren't propositions. Instead, noncognitivists maintain that moral claims express nonpropositional content, such as emotions or commands (e.g., "don't murder!"). This distinction was represented by three items, one reflecting noncognitivism and two that were reverse coded to reflect cognitivism. One such item is, "They believe that judgments about whether an action was morally right or wrong can be correct or incorrect."

All participants were then asked to express how much they agreed or disagreed with each of these items 1 (Strongly disagree) to 7 (Strongly agree). All items were reworded to remove the third-personal attribution at the start of the sentence (e.g., "They believed that…"), resulting in otherwise

---

[101] The term "universalism" could be used in other ways as well. For instance, it could refer to *who moral standards apply to*, i.e., the *scope* of moral concerns (Goodwin & Darley, 2008). What matters is the content of the construct (in this case, the notion that there is a single moral standard), and not the term we use to refer to it.

[102] Some theories allow for both stance-dependence and universalism, including ideal observer theory, which holds that moral facts are those facts that an ideally rational and fully informed person would endorse (Joyce, 2021). There may be a single universal moral standard, but it would still be made true by a stance, even if it is the stance of a hypothetical agent. More generally, one could in principle hold that all moral stances depend on a single agent's stance. This would make this account a form of moral antirealism (since moral facts are stance-dependent) but universal (since there is only one correct moral standard).

identical items, such as "Judgments about whether an action was morally right or wrong can be correct or incorrect." All items are available in **Appendix D**.

**Results**

Cronbach's alpha was calculated for the *third person* measure for each of the subscales. Following convention, Cronbach's alpha was poor for the 3-item realism subscale (α = .49) and good for the 3-item universalism subscale (α = .86) and the 3-item relativism subscale (α = .81). However, Cronbach's alpha for the 3-item noncognitivism scale was exceptionally low (α = .19). This pattern changed somewhat when assessing participant's self-reported metaethical positions (i.e., *first person* judgments). Cronbach's alpha was adequate for the realism subscale (α = .78), very high (and possibly somewhat redundant; see Tavakol & Dennick, 2011) for the universalism subscale (α = .92), good for the relativism subscale (α = .82), and again extremely poor for the noncognitivism condition (α = .42). There are considerable limitations with relying on Cronbach's alpha as a measure of internal consistency or unidimensionality (Sijtsma, 2009).[103] As such, the quality of each subscale was also assessed by conducting exploratory factor analysis for both the third person and first person scales. Results can be seen in **Supplement 6**, section ***S6.1.1.*** Put briefly, these findings suggest that while items on the relativism subscale loaded reasonably well onto their own factor in both the self and other conditions, realism and universalism items all loaded onto the same factor in both conditions, suggesting that participants may make little to no distinction between them. Critically, in both the self and other conditions, none of the three items in the noncognitivism subscale loaded onto the same

---

[103] Sijtsma puts it bluntly:

"Alpha is not a measure of internal consistency. Neither is it a measure of the degree of unidimensionality [...] Alpha has been shown to correlate with many other statistics and much as these results are interesting, they are also confusing in the sense that without additional information, both very low and very high alpha values can go either with unidimensionality or multidimensionality of the data. But given that one needs the additional information to know what alpha stands for, alpha itself cannot be interpreted as a measure of internal consistency. (2009, p. 119)

factor, and most loadings did not exceed 0.3 on any factor. As a result, the cognitivism subscale was dropped from all subsequent analyses.

For the third person measures of metaethical beliefs, the mean for the realism subscale was above the midpoint ($M = 5.26$, $SD = 0.97$), as were scores for the universalism subscale ($M = 5.46$, $SD = 1.15$), while the mean score for relativism subscale was below the midpoint ($M = 3.43$, $SD = 1.43$). Overall, these findings suggest that people on Amazon's Mechanical Turk in the US tend to think, at least in abstract cases where no concrete moral issue is specified, that a typical person in their society tends to endorse realism and universalism, and to reject relativism. For first person measures of metaethical beliefs, the mean score for the realism subscale was marginally above the midpoint ($M = 4.19$, $SD = 1.51$), and the mean score for universalism was, surprisingly *exactly* at the midpoint ($M = 4.00$, $SD = 1.80$), while the mean score for relativism was slightly above the midpoint ($M = 4.46$, $SD = 1.54$). These findings indicate that participants exhibit no strong tendency towards realism or antirealism, and may slightly favor relativism on average. Descriptive statistics for all measures appear in **Table 6.1**.

**Table 6.1**

*Descriptive statistics for Study 1: Third person paradigm (pilot)*

| condition | | n | mean | SD |
|---|---|---|---|---|
| realism | - abstract | 130 | 4.96 | 0.990 |
| | - right | 132 | 5.31 | 0.928 |
| | - wrong | 128 | 5.52 | 0.902 |
| universalism | - abstract | 130 | 5.01 | 1.264 |
| | - right | 132 | 5.54 | 1.064 |
| | - wrong | 128 | 5.83 | 0.946 |
| relativism | - abstract | 130 | 3.86 | 1.385 |
| | - right | 132 | 3.27 | 1.364 |
| | - wrong | 128 | 3.16 | 1.449 |

Averaging scores within each subscale, there was a small positive correlation between third person and first person responses for realism, $r(388) = .13$, $p = 0.01$, universalism $r(388) = .11$, $p = 0.03$, and relativism $r(388) = .16$, $p = 0.002$.

A one-way ANOVA and pairwise comparisons across conditions (*no statement*, *abstract right*, and *abstract wrong*) were conducted for each of the three subscales, in both the third person and first person measures.[104] A one-way ANOVA was performed to compare the effect of condition on third person realism subscale. Consistent with predictions, results showed that there was a significant main effect of condition on realism scores $F(2, 258) = 11.54$, $p < .001$. Games-Howell post-hoc tests were conducted for all pairwise comparisons. As predicted, the mean realism score for the no statement condition ($M = 4.96$, $SD = .99$) was significantly different than the mean realism score for the abstract right condition ($M = 5.31$, $SD = .93$), $t(258) = -2.94$, $p = .01$, and it was also significantly different than the mean realism score for the abstract wrong condition ($M = 5.52$, $SD = .90$), $t(254) = -4.79$, $p < .001$. Contrary to predictions, there was no significant difference between the abstract right and wrong conditions, $t(258) = -1.90$, $p = .142$.

Consistent with predictions, a one-way ANOVA revealed a significant main effect of condition on third person universalism scores $F(2, 255) = 17.30$, $p < .001$. Pairwise comparisons indicate that the mean universalism score for the no statement condition ($M = 5.01$, $SD = 1.26$) was significantly different than the abstract right condition ($M = 5.54$, $SD = 1.06$), $t(251) = -3.67$, $p < .001$, and the abstract wrong condition ($M = 5.83$, $SD = .95$), $t(239) = -5.89$, $p < .001$. The mean universalism score for the abstract right condition ($M = 5.54$, $SD = 1.06$). Contrary to predictions,

---

[104] I employed Welch's ANOVAs throughout. A Shapiro-Wilk test of normality revealed that none of the six conditions appeared to be normally distributed. More generally, Welch's tests are likely to perform better under a variety of conditions common to social psychological research, and are less likely to inflate Type I error rates compared to classical (Fisher's) ANOVA since the latter relies on assumptions about the data that many datasets don't meet (Delacre et al., 2019).

there was no significant difference between the abstract right and wrong conditions ($M = 5.83$, $SD =$ .95), $t(256) = -2.31$, $p = .057$.

Consistent with predictions, a one-way ANOVA revealed a significant main effect of condition on third person relativism scores $F(2, 258) = 9.22$, $p < .001$. Pairwise comparisons indicate that the mean relativism score for the no statement condition ($M = 3.86$, $SD = 1.39$) was significantly different than the abstract right condition ($M = 3.27$, $SD = 1.36$), $t(260) = 3.47$ $p = .002$, and the abstract wrong condition ($M = 3.16$, $SD = 1.45$), $t(255) = 3.92$, $p < .001$. Contrary to predictions, there was no significant difference between abstract right and wrong conditions $t(256) = 0.59$, $p = .824$.

Although there was no specific prediction, one-way ANOVAs revealed no main effect of condition for first person realism, universalism, or relativism scores.[105] Summary tables and graphs for the third person tests are available in **Supplement 6**, section **S6.1.2**.

**Discussion**

Overall, findings were consistent with predictions with respect to the difference between the no statement and abstract right and wrong statements, but not with respect to the difference between the abstract right and wrong statements. Consistent with expectations, participants perceived other people to endorse realism and universalism and to reject relativism the most when those people stated that an action was wrong, and the least when they provided no statement. Findings were less consistent with demonstrating a consistently intermediate pattern for those who stated an action was right. While there was a significant difference between the no statement and right conditions for all three subscales, the difference between the abstract right and wrong statement conditions was only significant for universalism. Although it may have approached significance for the realism condition, it fell very far from significance when comparing relativism scores between the two conditions. This indicates that

---

[105] A one-way ANOVA was conducted for each of the three subscales for the first person condition. A one-way ANOVA found no significant effect for realism $F(2, 258) = 0.66$, $p = .518$, universalism $F(2, 258) = 1.27$, $p = .282$, or relativism, $F(2, 258) = 2.42$, $p = .091$.

while the difference between the no statement and abstract statement conditions may be robust, there may be little or no difference in perception of metaethical beliefs between those who state an action is right and those who state that it is wrong, or, at the very least, this may not be true for perception of relativism in particular.

While these results were the only predictions that were made explicit, they are of secondary interest to what I take to be the more interesting implications of these findings. Among other considerations, the results of Study 1 suggest that MTurk participants in the United States tend to view typical people in their society as exhibiting a slight tendency towards realism and universalism, and a slight tendency against relativism. This was especially true in the abstract right and wrong statement conditions, compared to the no statement condition, which was closer to the midpoint. In the absence of any additional information, people from this population may be hesitant to draw strong conclusions about other people's metaethical standards.

There was also a slight tendency towards judging a typical person to be especially likely to endorse realism and universalism, and reject relativism, when that person stated that an action was morally wrong rather than morally right. However, this tendency was not significant. This may indicate a very slight tendency to judge people who have a more restrictive moral attitude to be realists/universalists, and to reject relativism, an effect which could perhaps be detected in a sample with greater power. However, even if there were such an effect, it would appear to be weak. This is surprising, since realism (though it's less clear the same holds for universalism) is associated with a more intolerant and close-minded attitude towards people with contrary moral beliefs. It seems plausible in light of this that judging an action to be morally wrong would be a stronger cue towards a less permissive and more closed perspective. It could simply be that the difference between conditions is too insignificant, given how the conditions are, overall, extraordinarily impoverished in

terms of the information given. It may be that in real-world circumstances, people are sensitive to a variety of cues to infer other people's metaethical views that were absent from these conditions.

Yet the most important finding is that people's judgments about a typical person's metaethical standards were only very weakly correlated with their own first person metaethical judgments. This conflicts with traditional philosophical presumptions. Philosophers have historically presumed that all competent members of a particular linguistic community share a uniform and determinate set of metaethical presuppositions implicit in the way they speak and think (Gill, 2009). Yet participants do not appear to assume the typical person shares their own metaethical presuppositions. Of course, philosophers could maintain both that people do share the same metaethical presuppositions, but that ordinary people's judgments about how other people's metaethical stances and commitments may reflect a kind of second-order incompetence or performance error (Martí, 2009).

This could be true, but if so, why the disparity between first and third person judgments? Are we to consider people accurate judges of their own metaethical standards, but incompetent with respect to judging a typical person's standards? And that people have made the further mistake of presuming other people have different metaethical standards than they do, even though *ex hypothesi* this is false (or even impossible) since all competent members of a given community must share the same uniform metaethical presuppositions (i.e., *everyone* must either be a realist about *all* moral claims or *none*)? Perhaps so, but we're not entitled to presume this is the case. Philosophers are in no position to dismiss empirical data merely on the basis of it conflicting with their idealistic conceptions of people *ought* to speak and think. The matter calls for empirical resolution.

Taken at face value, these findings suggest that ordinary people have little trouble supposing that other people *don't* share their metaethical presuppositions. And this raises important methodological questions with respect to existing metaethics paradigms. First, suppose that participants are a representative sample of the population about whom they are judging in the third

person measures. If so, it is possible that participants have an inaccurate picture of how other people think: people mistakenly think other people are more inclined towards realism and universalism, and less inclined towards relativism, than themselves. This would be an important discovery all on its own, and one well worth exploring, given that one's perception of other people's metaethical beliefs may have important psychological and behavioral implications. It's also possible people's third personal judgments provide a more accurate picture of folk metaethics. People may be more subject to normative conflations when asked about their own metaethical standards. If so, the disparity between third person and first person judgments may be due in part to the minimization of normative conflations, though present findings cannot resolve the matter.

These findings suggest another possibility, however. It may be that participants *don't* represent the population they have in mind when judging third person items. If participants are judging a typical person in their society, i.e., someone in the United States, but participants are skewed towards particular demographics dimensions relevant to their own metaethical standards, then differences between first and third person judgments could both be accurate. If so, there may be another explanation for these results: they may reflect an accurate assessment of the typical American, relative to participants. Although previous studies have found no association between political conservatism and realism when judging concrete moral issues (Goodwin & Darley, 2008, Pölzler & Wright, 2020b), they *have* found an association between conservatism and realism for *abstract* measures (Pölzler & Wright, 2020b). Collier-Spruel et al. (2019) also found that scores on their Moral Relativism Scale (MRS) were negatively associated with a conservative political orientation, right wing authoritarianism, but positively associated with progressive values.[106] In addition, participants who choose realist responses have consistently reported greater religiosity while low religiosity has been consistently

---

[106] Progressive values were calculated by subtracting the three foundations emphasized more by American conservatives—purity, ingroup loyalty, and respect for authority—from their average score on the remaining two foundations, harm and fairness, which tend to be emphasized more by American liberals (Collier-Spruel et al., 2019, p. 12; Graham et al., 2009).

linked to antirealist responses (Collier-Spruel et al., 2019; Goodwin & Darley, 2008, Yilmaz & Bahçekapili).

If participants are less religious and conservative than whoever they have in mind when responding to third person items, this could account for the difference between third and first person judgments. Participants in the present sample were on average more inclined towards liberalism than conservatism ($M = 3.58$ $SD = 1.80$) and were extremely low in religiosity ($M = 2.99$, $SD = 2.16$), with 43.8% of participants selecting 1 out of 7 for "not religious at all." In line with research assessing demographic differences between MTurk participants and the general population, these results show that the participants in this sample are both less conservative than the general population and *much* less religious. As one recent study found, "a strikingly substantial percentage of the MTurk sample—over 40%—identifies as agnostic or atheist," while a more representative sample of the general population finds only 10% of the population identifies as agnostic or atheist (Levay, Freese, & Druckman, 2016, p. 4).

This suggests a potentially major advantage of third person paradigms over first person paradigms: first person paradigms may be more susceptible to poor generalizability, insofar as the samples researchers tend to draw on are skewed towards particular demographic variables that render them less representative of the samples they are drawn from in demonstrably relevant ways. Third person paradigms may serve as a buffer against unrepresentative sampling, and provide a more accurate picture of the metaethical views of the populations they are drawn from. If so, present findings would suggest that researchers may be underestimating the degree to which people in the United States endorse realism, insofar as present sampling methods are skewed towards less religious and less conservative participants.

There are a number of important observations about these findings that are relevant both to subsequent studies reported here and to future research on folk metaethics. One intriguing takeaway

is the lack of distinction between realism and universalism. Psychologists interested in mapping latent variables in a bottom-up way, without the theoretical baggage imported from philosophy, may shrug and suggest that these subscales should be collapsed into a single realism-universalism dimension. Philosophers may balk at such a notion. They may point out that realism and universalism *are not conceptually identical*, and that at least under ideal circumstances a reflective person ought not to conflate them, any more than one ought to treat a square and a rectangle as the same shape. Whatever affinities the two may have, there is a difference, and it would be quite a concession to make to accept that ordinary people do not conceptually distinguish realism and universalism.

I cannot adjudicate this dispute here, but it is worth noting that both sides have points in their favor; on the one hand, philosophers are not entitled to insist that folk psychology must comport with the conceptual distinctions that animate contemporary analytic philosophy. On the other hand, psychologists may be too quick to treat as distinct or indistinct, as the case may be, features of ordinary thought, rather than consider the possibility that the patterns observed in their data are a methodological artifact: it could be that the lack of distinction between realism and universalism is due to a pattern of unintended interpretations that yield a converging response pattern but does not reflect a genuine psychological construct (i.e., some shared folk notion that combines realism and universalism). Psychologists should be cautious in identifying new psychological constructs when results could be due to the poor validity of their measures, especially when there is some theoretical rationale for suspecting such constructs would entail attributing conceptually muddled notions to ordinary human thought. For the purposes of these studies, realism and universalism will be treated as distinct. Assessment of the degree to which people distinguish or fail to distinguish realism and universalism will be left to future research.

It's also unclear why Cronbach's alpha was so low for the realism subscale in the third person condition, given that it was much higher in the first person condition. It may be that people are less

certain about the metaethical beliefs of a typical person than themselves, especially when they are given little or no information about people in their society. However, this doesn't explain why Cronbach's alpha was high in the third person versions of the universalism and relativism subscales.

Examination of the frequency of participants who selected each of the possible responses on the Likert scale suggests participants are fairly evenly distributed across all response options, with roughly half above and below the midpoint (a finding which holds true even without collapsing the no statement, right, and wrong conditions). At face value, this suggests people have highly variable views towards moral realism, though the even distribution of responses could suggest that participants had trouble interpreting the items and chose randomly. One might wonder why acquiescence didn't favor overall average agreement, though this could have been offset by the participants in the sample exhibiting a tendency towards lower realism and universalism and greater relativism. The lack of main effects in the self conditions may also be due in part to the greater standard deviations for responses to the first person versions of the subscales, compared to the third person versions of those scales. First person scales may require larger sample sizes to detect significant differences. While this may be a cost to employing first person measures, it may be necessary insofar as it captures genuinely greater individual differences in self-reported metaethical views, compared to third person judgments.

Previous studies have shown differences in the proportion of realist and antirealist responses when considering morality in the abstract versus considering concrete moral issues (Pölzler & Wright, 2020a; 2020b). Most importantly they've shown high levels of variation between different concrete moral issues, and in line with previous research, results often vary in accordance with concrete moral issues (e.g., Beebe, 2015; Beebe & Sackris, 2016; Goodwin & Darley, 2008; 2012; Pölzler & Wright 2020b, Wright, Grandjean, & McWhite, 2013). In the next study, I sought to expand on the present findings by exploring third personal judgments in response to a set of concrete moral issues.

**6.2 Study 2: Third person paradigm with concrete moral issues**

Study one introduced the third person paradigm. This involved assessing how participants evaluated other people's metaethical beliefs either when asked about a typical person in their society, or when presented with a person who made a moral assertion that an unspecified action was morally right or wrong.

This study expands on Study 1 by introducing four concrete moral statements in which the participant is told that someone asserted that a particular action was morally wrong. This study also employs a within-subjects design, with all participants responding to both the no statement condition in Study 1 and each of the four concrete moral issues. I also introduced a handful of items that measure how participants evaluate the character of the person who made the statement. These measures were introduced primarily for exploratory purposes, so there were no predictions related to them. Finally, the sample size was increased. These changes were intended to both assess the results of the third person paradigm for concrete moral issues and to do so under conditions with substantially greater power. Once again, since these findings represent the introduction of a novel paradigm for measuring folk metaethical stances and commitments, my goals were largely exploratory. Due to the exploratory nature of these findings, I have relegated detailed description of the methods and results of Study 2 to **Supplement 6**, section *S6.2*.

**6.3 Study 3: Third person paradigm (nonmoral domains)**

The goal of Study 3 is to expand the third person paradigm to nonmoral domains. One of the most fascinating and important elements of moral judgments is the way they differ from nonmoral judgments, such as matters of taste and claims about science and history. Although people may turn out to have more diverse and nuanced views than researchers imagine, one might generally suppose that people lean towards realism about factual matters in math, science, history, and so on, but lean more towards antirealist views such as subjectivism when thinking about matters of personal taste.

Matters of social convention may likewise be thought of not as stance-independent facts, but as culturally constructed norms and institutions that exist by social consensus. Previous research on folk metaethics has occasionally sought to compare judgments in each of these domains: morality, taste (and aesthetics), social convention, and facts (e.g., Beebe, 2015; Beebe et al., 2015, Goodwin & Darley, 2008, Wright, Grandjean, & McWhite, 2013). These findings generally comport with what researchers might expect: in descending order, realist ratings are highest for the factual domain, followed by morality, social conventions, and taste.

This general pattern is complicated by extremely high domain variation *within* each domain, meaning that many participants choose realist responses for some issues and antirealist responses for others within each domain. For Instance, Beebe (2015) found substantial variation in responses towards factual issues (such as scientific claims), with a consensus in favor of realism towards some and antirealism towards others, while Goodwin and Darley (2008) found that their participants were almost evenly divided about matters of social convention (Goodwin & Darley, 2008), prompting researchers to conclude that people may be *pluralists* about these topics (Pölzler & Wright, 2020b; Wright, Grandjean, & McWhite, 2013).

The reasons *why* people are (or at least appear to be) pluralists are complex and have yet to be fully explicated. At least one reason people may only appear to be pluralists about moral issues could be due to normative conflations, though, as I have argued at length in preceding chapters, a variety of other factors likely influence variable response patterns as well. Yet the factors contributing to apparent pluralism in the moral domain may differ in relative importance for nonmoral domains, or may differ altogether. For instance, epistemic conflations may play an important role when considering claims in non-normative domains, e.g., descriptive statements about science, history, or math. Beebe (2015) found that when participants were asked to judge disagreements about uncontroversial factual issues, realist response rates were very high, a finding consistent with previous research. However,

when they were asked to judge disagreements about factual disputes issues for which there is greater uncertainty, or no practical way to know what the correct answer is, antirealist response rates shot up dramatically, approaching and in some cases exceeding 50%. For instance, about 45% of participants judged that if two people disagreed about whether Julius Caesar drank wine on his 21ˢᵗ birthday, they could both be correct. We *could* interpret this as evidence that around half of Beebe's participants believe there is no objective fact about certain historical events, but it is far more plausible that these participants did not interpret this question as intended. Given that the greater the *uncertainty* about the factual dispute in question, the greater the "antirealist" response rate, the best candidate for these findings are epistemic conflations.

Critically, Beebe's findings also suggest that participants may interpret analogous stimuli in different domains in different ways, even when virtually all other aspects of the design are held constant. For instance, if participants are asked to judge an otherwise identically-phrased disagreement between two people, disagreements about moral issues and disagreements about scientific or historical disputes may be interpreted differently. This possibility threatens inter-domain comparisons, since it suggests that there may not only be substantial methodological shortcomings with questions about realism and antirealism in general, but that there may also be an additional layer of methodological shortcomings whenever one seeks to compare judgments about realism and antirealism in different domains: people may interpret the "same" question about whether, e.g., if two people disagree, both can be correct, differently based entirely on the domain of disagreement.

If so, responses to what is intended to be the same question across domains would effectively function as responses to different questions, undermining our ability to make inter-domain comparisons. In short, it is important to devise paradigms with cross-domain interpretative consistency…yet, notably, most research making inter-domain comparisons so far has relied

exclusively on the disagreement paradigm with all its attendant methodological shortcomings.[107] This

makes the rationale for studying these domains using a novel paradigm especially worthwhile, but it

comes at the cost of abandoning the primary rationale for the third person paradigm in the first place:

to minimize normative conflations. Such conflations seem far less likely in nonmoral domains. Even

so, this is offset by the potential advantages of introducing a novel method for assessing realism,

universalism, and relativism in nonmoral domains.

It would also be helpful to establish just what these nonmoral domains are supposed to reflect.

The *taste* domain encompasses aesthetic judgments about art, music, and other matters of creative

expression, as well as personal taste in e.g., food and drink. For example, one item from the realist

subscale states:

> *They believe there are facts about which foods taste good or bad that are true even if some people or societies*
> *think otherwise.*

*Social conventions* refer to nonmoral norms regarding appropriate social conduct, such as rules about

how to dress, how to refer to others, how to greet other people (bowing, shaking hands, and so on),

use of body language, how much personal space to accord others, and so on. While many Western

populations may reliably distinguish paradigmatic moral norms from social conventions, the boundary

between these domains often blurs, and people vary both within and between populations in the extent

to which they regard a given issue as a matter of morality, social convention, or neither (see **Appendix

E**). Nevertheless, researchers may still draw *a priori* distinctions when designing measures such that, if

participants do not themselves conceive of moral and social conventions as distinct, this may or may

not be reflected in their pattern of responses, and need not jeopardize the validity of the measures.

Finally, the factual domain refers to *descriptive* (rather than normative) claims that purport to

tell us what the world is like. This domain encompasses claims about science, history, math, and other

---

[107] I address *why* this has occurred in **Supplement 6**, section *S6.3.1*.

related claims. The following is perhaps the most illustrative example of an item in the factual domain (from the realist subscale):

> *They believe that claims that describe what the world is like, such as the claims made in geography, history, and physics, are correct regardless of the preferences or values of different individuals or cultures.*

This is arguably the least conceptually well-defined category. Roughly, the factual domain refers to claims ostensibly intended to reflect what the world around us is like, and more generally issues about which one might imagine people are generally inclined to think are discovered, rather than created or a matter of individual preference or social consensus. This is an admittedly somewhat unsatisfying way of defining the domain since, after all, the whole point of distinguishing these domains is to determine whether people endorse realism, universalism, and relativism about the domain in question. It seems a bit strange, given this, to define the domain as presupposing e.g., realism by definition. This is why it may be more helpful to distinguish this domain from others by noting that the claims in question are not normative, but instead consist of *descriptive* claims, claims intended to describe some state of affairs. Perhaps it would be more helpful to describe the *descriptive* domain.[108]

My general expectation is that we would find roughly the same pattern for the third person versions of these measures: participants would generally favor realism and universalism but reject relativism in the factual domain, and this pattern would decline and perhaps reverse as one moved in descending order to the moral domain, to social conventions, and finally to matters of taste. Consistent with the results of Study 1 and Study 2, I expected participants to continue to judge a typical person to lean towards realism and universalism and to reject relativism. I expected this pattern to reverse for matters of taste, with third person judgments indicating that participants thought a typical person rejected realism and universalism about such matters, and was more inclined towards relativism. It is less clear what to expect in the case of social conventions. They tend to fall somewhere in the middle, with

---

[108] See **Supplement 6** section *S6.3.2* for additional commentary.

Goodwin and Darley (2008)'s findings placing them closer to facts and morality, but Wright et al.'s (2013) findings suggesting they perform more similarly to matters of taste. Nevertheless, I expected social conventions to perform closer to matters of taste.

**Methods**

***Participants.*** I aimed to recruit 800 participants. 801 participants began the study. However, one participant did not complete the study and was excluded from further analysis. This left a total of 800 participants. Participants consisted of 800 adult US residents on Amazon's Mechanical Turk (430 females, 362 males, 8 other, $M_{age}$ = 41.9, $SD_{age}$ = 12.3, age range = 19-78).

***Procedure.*** The procedure employed in Study 3 was almost identical to Study 1. All participants were randomly assigned to one of four conditions: moral, aesthetic, social convention, or facts. All conditions were identical to the no statement condition in Study 1, in that participants were asked to consider the beliefs of a typical person in their society, without being provided any additional information. The only difference between conditions were the measures. The moral condition used the same measures as Study 1, with the exclusion of the three-item noncognitivism subscale, resulting in a total of nine items: a three-item realism subscale, a three-item universalism subscale, and a three-item relativism subscale. Order of all nine items was randomized within condition. The measures in each of the nonmoral conditions also consisted of three-item subscales for realism, universalism, and relativism, each adapted to reflect the respective realist, universalism, and relativist notions of truth in the relevant domain and, where possible, to serve as close analogs to their counterparts in the morality subscales.[109] After answering the third-person versions of each question, participants were presented with first person versions of the same items, and asked to judge how much they agreed or disagreed with them. Finally, participants were asked to report their age and gender.

---

[109] I could describe these measures as metanormative positions for the moral, aesthetic, and social conventional domains, but it is less clear whether the measures used in the factual condition could be appropriately described as normative. Nevertheless items in the factual domain still reflect realism, universalism, and relativism about the relevant claims.

***Measures.*** Measures used in the third and first person moral condition were identical to the measures used for the realism, universalism, and relativism subscales in Study 1. Three-item subscales for realism, universalism, and relativism were introduced for each of the nonmoral domains (taste, social conventions, and facts), and were likewise adapted for use as both third and first person measures. All scale items are available in **Appendix D**.

**Results**

The mean, standard deviation, and Cronbach's alpha for the realism, universalism, and relativism subscales for each of the four domains are available on **Table 6.2**.

My primary goal was to evaluate differences in the degree to which participants would judge a typical person in their society to endorse realism, universalism, and relativism with respect to each of the four domains: morality, taste, convention, and facts. First, I conducted a Welch's one-way ANOVA to compare the effect of domain on realism. There was a significant main effect of condition on realism $F(3, 438) = 50.2, p < .001$. All pairwise comparisons are available on **Table 6.2**, and means with 95% CIs are featured in **Figure 6.1**.

**Table 6.2**

*Descriptive statistics for Study 3: Moral and nonmoral domains*

| 3rd/1st person | metaethical stance | condition | mean | SD | α |
|---|---|---|---|---|---|
| third person | realism | moral | 5.09 | 0.98 | 0.576 |
| | | taste | 4.09 | 1.25 | 0.686 |
| | | convention | 4.94 | 1.20 | 0.787 |
| | | facts | 5.44 | 0.93 | 0.635 |
| | universalism | moral | 4.88 | 1.42 | 0.913 |
| | | taste | 3.53 | 1.35 | 0.776 |
| | | convention | 4.37 | 1.48 | 0.838 |
| | | facts | 5.03 | 1.12 | 0.635 |
| | relativism | moral | 3.45 | 1.31 | 0.801 |
| | | taste | 5.06 | 1.20 | 0.808 |
| | | convention | 4.53 | 1.32 | 0.854 |
| | | facts | 3.72 | 1.14 | 0.636 |
| first person | realism | moral | 4.36 | 1.42 | 0.748 |
| | | taste[110] | 3.98* | 1.72* | 0.629* |
| | | convention | 3.37 | 1.53 | 0.822 |
| | | facts | 5.33 | 1.20 | 0.641 |
| | universalism | moral | 4.01 | 1.86 | 0.93 |
| | | taste | 2.48 | 1.39 | 0.941 |
| | | convention | 2.54 | 1.58 | 0.89 |
| | | facts | 5.06 | 1.35 | 0.646 |
| | relativism | moral | 4.10 | 1.57 | 0.822 |
| | | taste | 5.78 | 1.05 | 0.987 |
| | | convention | 5.62 | 1.22 | 0.793 |
| | | facts | 3.85 | 1.33 | 0.658 |

*Note. The mean, standard deviation, and alpha reported for the first person realism taste subscale are based on items #2 and items #3. Item #1 was dropped from analysis due to an error.*

---

[110] Item #1 in the first person realism taste subscale was dropped from analysis due to an error. Specifically. I accidentally included item #3 from the first person taste relativism subscale in place of the correct item. Unfortunately, this means that only two items are included in the first person realism taste subscale. In addition, item #1 on the first person realism convention subscale and item #1 on the first person realism facts subscale both mistakenly include the word "that." In particular, item #1 for the first person realism convention subscale states, "When two people disagree about cultural practices, such as appropriate work attire, that [sic] at least one of them must be incorrect." Item #1 for the first person realism facts subscale states, "When two people disagree about matters of science or history that [sic] at least one of them must be incorrect." Corrected versions of all three items appear in **Appendix D**.

**Table 6.3**

*Post Hoc Comparisons - realism*

| comparison | | mean difference | df | t | p-value |
|---|---|---|---|---|---|
| condition | condition | | | | |
| moral | - taste | 0.997 | 370 | 8.82 | < .001 |
| | - convention | 0.145 | 380 | 1.31 | 0.554 |
| | - facts | -0.352 | 402 | -3.70 | 0.001 |
| taste | convention | -0.853 | 391 | -6.89 | < .001 |
| | facts | -1.349 | 360 | -12.19 | < .001 |
| convention | facts | -0.497 | 371 | -4.61 | < .001 |

**Figure 6.1**

*Study 3: Comparison of means (with 95% CIs) across domains for third person|realism condition*
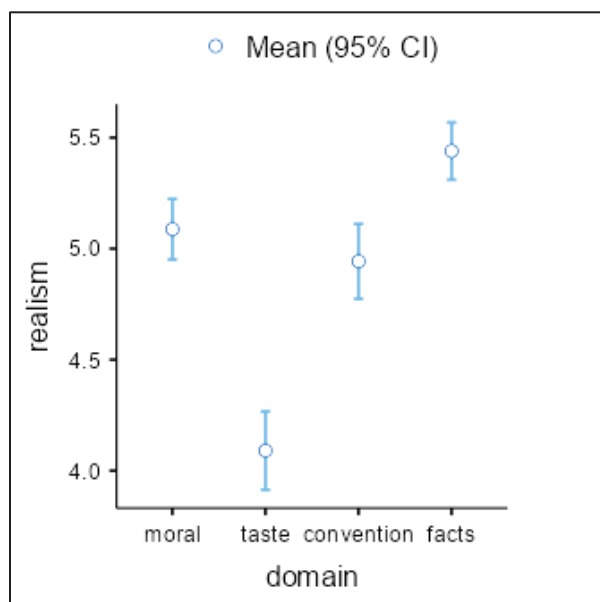


The next metaethical domain I assessed was universalism. There was a significant main effect of domain on universalism $F(3, 438) = 54.1, p < .001$. All pairwise comparisons are available in **Table 6.3**, and means with 95% CIs are featured in **Figure 6.2**.

**Table 6.4**

*Post Hoc Comparisons – universalism*

| comparison | | mean difference | df | t | p-value |
|---|---|---|---|---|---|
| condition | condition | | | | |
| moral | - taste | 1.350 | 396 | 9.70 | < .001 |
| | - convention | 0.508 | 397 | 3.50 | 0.003 |
| | - facts | -0.153 | 381 | -1.21 | 0.623 |
| taste | convention | -0.841 | 390 | -5.89 | < .001 |
| | facts | -1.502 | 378 | -12.09 | < .001 |
| convention | facts | -0.661 | 366 | -5.04 | < .001 |

**Figure 6.2**

*Study 3: Comparison of means (with 95% CIs) across domains for third person|universalism condition*
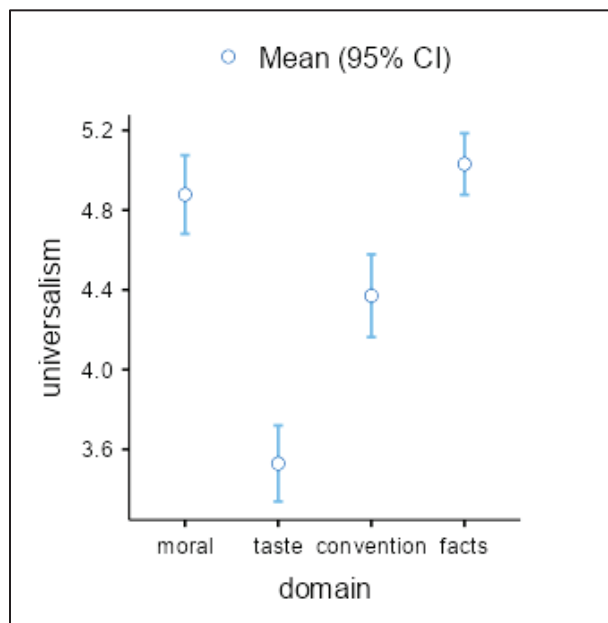


Finally, I assessed domain differences with respect to relativism. There was a significant main effect of domain on relativism $F(3, 438) = 54.1$, $p < .001$. All pairwise comparisons are available in **Table 6.4,** and means with 95% CIs are featured in **Figure 6.3**.

**Table 6.5**

*Post Hoc Comparisons - relativism*

| comparison | | mean difference | df | t | p-value |
|---|---|---|---|---|---|
| condition | condition | | | | |
| moral | - taste | -1.610 | 395 | -12.80 | < .001 |
| | - convention | -1.078 | 398 | -8.21 | < .001 |
| | - facts | -0.269 | 395 | -2.21 | 0.123 |
| taste | convention | 0.536 | 389 | 4.22 | < .001 |
| | facts | 1.345 | 394 | 11.49 | < .001 |
| convention | facts | 0.809 | 387 | 6.58 | < .001 |

**Figure 6.3**

*Study 3: Comparison of means (with 95% CIs) for third person|relativism condition*
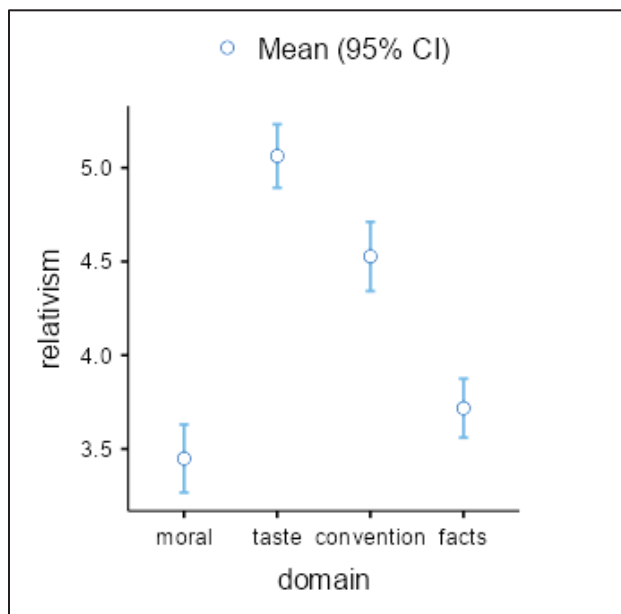


Almost all results were consistent with expectations. I expected participants to judge that a typical person in their society exhibited comparatively higher levels of realism and universalism for facts, followed by morality, social convention, and finally taste. Two of these differences were not

statistically significant. First, with respect to realism, the difference between morality ($M$ = 5.09, $SD$ = 0.98), and convention ($M$ = 4.94, $SD$ = 1.20) was not significant, $t$(380) = 1.31, $p$ = 0.554. Second, with respect to universalism, the difference between morality ($M$ = 4.88, $SD$ = 1.42), and facts ($M$ = 5.03, $SD$ = 1.12) was not significant, $t$(381) = -1.21 $p$ = .623. However, all differences were in the expected order, with point estimates falling exactly in line with expectations for both universalism and relativism. This suggests that the expected pattern may generally hold, albeit in some cases the size of the difference may be small, and, while it is possible that the order of one or more domains may flip in a handful of cases, such deviations from the expected pattern would likely be very small and merely reflect an approximate lack of difference that still mostly maintains the expected ordering of realism and universalism judgments across domains.

With respect to relativism, the general pattern was also mostly in line with expectations. Relativism scores were highest for the taste domain, followed by social conventions. However, contrary to expectations, there was no significant difference in relativism scores between morality ($M$ = 3.45, $SD$ = 1.41) and facts ($M$ = 3.72, $SD$ = 1.14), $t$(395) = -2.21 $p$ = .123. Critically, this difference was *not* in the expected direction, with relativism scores somewhat lower for morality than for facts. Of course, this difference wasn't statistically significant, but it still reflects a departure from expectations about the overall ordering of relativism judgments across domains. It is possible that this difference is due to the particular stimuli I used. The items used for the factual domain were first introduced in this study, and their general psychometric properties have yet to be fully explored. As such, future studies could vindicate the hypothesized pattern. Of course, it is also possible that people genuinely do tend to judge a typical person in their society to be approximately similar with respect to relativism for both moral and factual issues, or to even regard morality as less relativistic than factual claims. Given serious concerns about adapting paradigms designed to assess moral disagreements to factual disagreements (see Beebe, 2015), it may be that the factual domain is subject to distinctive

methodological concerns. In particular, participants may conflate epistemic considerations with metaphysical ones, resulting in implausibly high rates of apparent "relativism" for factual disputes. As such, I'm skeptical of this inconsistency, though of course such skepticism should be qualified by the clear motivation to not want to believe results contrary to my expectations.

One result that did unambiguously fly in the face of expectations was the overall performance of social conventions. I expected social conventions to either be significantly less than or not significantly different from the midpoint for realism and universalism, while I expected social conventions to be significantly above the midpoint for relativism, or at least not significantly different from it. In other words, I expected people to tend to regard social conventions as relative, but not stance-independently true or universal. To test for these hypotheses, I conducted right-tailed one-sample t-tests to compare the mean score for the convention domain against the midpoint (i.e., $H_a$ $\mu$ > 4 ) for realism, universalism, and relativism. Contrary to predictions, the mean score for realism ($M$ = 4.94, $SD$ = 1.20) was significantly *greater* than the midpoint $t(197)$ = 11.02, p < .001, $d$ = 0.783. Also inconsistent with predictions, the mean score for universalism ($M$ = 4.37, $SD$ = 1.48) significantly greater than the midpoint, $t(197)$ = 3.52, p < .001, $d$ = 0.250. However, consistent with predictions, the mean relativism score for conventions ($M$ = 4.53, $SD$ = 1.32) was significantly greater than the midpoint, $t(197)$ = 5.63 p < .001, $d$ = 0.400.

**Discussion**

Overall, the results of Study 3 mostly corroborated expectations. Consistent with what I expected, participants tended to regard a typical person in their society as strongly inclined towards realism and universalism for factual disputes, and disinclined towards relativism. This same pattern held, in descending order of support for realism and universalism and rejection of relativism, for their judgments about morality, social conventions, and taste.

However, I expected the factual and moral domains to remain above the midpoint for realism and universalism, and below the midpoint for relativism, but for this pattern to flip for both social conventions and matters of taste. Instead I found that while social conventions appeared in the expected order (i.e., between morality and taste) for all three metaethical dimensions, participants' judgments of realism and universalism were both above the midpoint, which unexpectedly indicates that participants thought of a typical person as more inclined on average towards realism and universalism. In other words, if participants interpreted items in the conventional domain as intended, this would mean that they believe a typical person in their society is more inclined to think of social conventions such as how to dress and how to greet one another as stance-independent normative facts that are universally applicable to everyone, including people *outside* that person's culture.

Perhaps participants do think this way, but if so, this would be a surprising result. To add to the puzzling nature of these results, if participants were generally inclined towards unintended interpretations of the conventional domain in a way that inclined them towards thinking of it in realist and universalist terms, one might also expect participants' first person judgments to also be significantly above the midpoint. This is not what I found. Left-tailed one-sample t-tests revealed, just as I would expect, that the mean realism ($M = 3.37$, $SD = 1.53$) score for first person judgments was significantly *below* the midpoint $t(197) = -5.82$ p $< .001$, $d = -0.414$, and the mean score for universalism ($M = 2.54$, $SD = 1.58$) was likewise significantly below the midpoint, $t(197) = -13.00$ p $< .001$, $d = -0.924$, while a right-tailed test found, again consistent with expectations, that the mean relativism score ($M = 5.62$, $SD = 1.22$) was above the midpoint, $t(197) = 18.78$, p $< .001$, $d = 1.335$.

This is exactly the pattern I would expect. The disparity between first and third person is thus at least somewhat more likely to reflect a genuine difference in people's conception of a typical person relative to themselves, and counts somewhat against dismissing such findings as a methodological fluke. I am not sure what to make of these particular findings. While I expect they may still reflect a

non-obvious pattern of unintended interpretations, evaluating the validity of these measures and assessing how participants are interpreting items within each of these domains is a task for future research.

**6.4 Study 4: Normative entanglement**

Study 4 diverges from the third person paradigm in order to assess a proposed phenomenon that could play a role in accounting for why people are on occasion disposed to express moral realism, even if they don't actually endorse it, or endorse it based on mistaken inferences about its implications: *normative entanglement*. Normative entanglement occurs whenever a metaethical claim is embedded in a more complicated assertion that also includes normative content, such that whoever asserts or endorses the metaethical element of the claim pragmatically implies information about their normative moral stance and/or their attitude towards a given normative moral issue, *even though such implications are not logically entailed by the expression of a metaethical stance*. The best way to illustrate how normative entanglement arises is to point to a hypothetical exchange between a moral realist and a moral antirealist:

Realist: *"Do you think it's objectively wrong to torture babies for fun?"*

Antirealist: *"No, I don't think it's objectively wrong to torture babies for fun."*

Realist: *"Wow! So you think it's totally okay to torture babies for fun? You're a disgusting monster!*

One problem with the realist's response is a mistake with respect to the negation of a normative claim. To deny that an action is wrong does not entail that one thinks the action is "totally okay," or, to put it in less colloquial terms, it does not entail that the action in question is permissible (or, for that matter, recommended, required, supererogatory, and so on). All denying that an action is wrong would

entail is the denial that it is "wrong"; it doesn't *necessarily* entail any positive normative stance towards the issue in question.[111]

However, this concern is moot, because the antirealist's response does not logically entail *any normative position at all*. Denying that an action is objectively wrong *is not* a denial that it is wrong. Merely because someone does not think that an action is *objectively* wrong, it does not follow that they don't think it's wrong in some non-objective way. For comparison, you could think that your favorite flavor of ice cream is good, but this does not mean that you must believe it is *objectively* good, and that anyone who prefers another flavor of ice cream is mistaken (Loeb, 2003). Having a normative moral stance doesn't *require* objectivity. *Some* moral antirealist positions may also entail normative implications. A thoroughgoing nihilist could reject the notion there is any respect at all in which anything could be morally right or wrong, and may also deny not only that anything is intrinsically valuable, but that anything can have meaning or value in any respect at all. On account of this, they may reject all notions of compassion or concern for the welfare of others. Such an account *is not* a logical entailment of denying moral realism. Contrary to how the rejection of realism is sometimes framed, our options aren't restricted to either realism or a deeply nihilistic existential malaise that rejects any and all meaningful conceptions of goodness and value. A variety of moral antirealist positions are consistent with having substantive first-order moral stances about the rightness and wrongness of moral actions, including cultural relativism, individual subjectivism, constructivism, and ideal observer theory. Some expressivist accounts likewise allow antirealists to sincerely assert that actions are morally right or wrong without error or contradiction (Blackburn, 1991; Joyce, 2021). In short, antirealism only involves the rejection of stance-independent moral facts. It does not entail the rejection of normative moral standards, i.e., beliefs about what is morally right or wrong.[112]

---

[111] See **Supplement 6**, section *6.3.3* for additional commentary.
[112] See **Supplement 6**, section *6.3.4* for additional commentary.

Rejecting the notion that a given action is objectively wrong likewise does not logically entail anything about one's attitudes or emotional responses towards the action in question. Someone who denies that it's objectively wrong to torture babies can still oppose baby torture. Antirealism does not require moral indifference or a muted emotional response to moral transgressions. Denying that something is objectively wrong is completely consistent with finding the action in question repugnant, viewing those who commit such actions with contempt, supporting laws against the action, wanting to see perpetrators punished, and so on.

For comparison, denying that chocolate cake is *objectively* tasty doesn't entail that you believe chocolate cake doesn't taste good, or that you find it disgusting, or that you are as equally willing to eat chocolate cake as a bucket of dirt (Loeb, 2003). A person who denies realism about gastronomic facts could have equally strong food preferences, and be just as disposed to judge foods as "good" or "bad" as a realist. The heart of the problem, then, is an equivocation between *good* and *objectively good*, where one mistakenly infers that to deny that something is objectively good is to deny that it is good, full stop. This is a mistake because our normative and evaluative judgments about whether things are right or wrong, and good or bad, respectively, need not *require* realism. Metaethical and normative considerations are conceptually distinct.

Of course, such rhetorical ploys likely occur primarily in academic disputes, since ordinary people are less disposed to engage in philosophical arguments about moral realism and antirealism. Explicit metaethical language may appear in everyday discourse on occasion (though efforts could be made to evaluate how often and under what circumstances). Yet the specific example given here merely reflects an explicit form of the phenomenon. Normative entanglement could appear in far subtler ways in everyday discourse. More importantly, it reflects a distinct way in which pragmatic considerations, such conversational implicature, can influence the inferences people make about what others say (Grice, 1975; Levinson, 2000). Unfortunately, both philosophers and psychologists often

ignore, or at least fail to adequately consider, the potential role pragmatics could play in influencing how participants respond to stimuli. In some cases, pragmatics may fully drive participant responses in ways that create the illusion of substantive psychological phenomenon (Adams & Steadman, 2004a; 2004b; Lindauer and Southwood, 2021; cf. Sytsma, Bishop, & Schwenkler, 2022).

This shortcoming highlights yet another methodological vulnerability to research on metaethics: participants may make inferences about the implications of selecting a realist or antirealist response that cause them to select responses intended to express a particular metaethical position for reasons unrelated to endorsing the actual content of the metaethical position itself. In practice, this may involve participants inferring that a response option intended to measure antirealism carries additional implications, and these implications are what drive them to not choose the antirealist response, *not* the fact that the statement expresses antirealism. More generally, if those inferences are *good*, if, e.g., participants perceive endorsing an antirealist stance to imply a desirable normative moral stance, or to have socially desirable implications, such as signaling tolerance or open-mindedness, participants may favor those responses, either because they mistakenly believe the position entails those implications, or because, even if they do not, they judge (quite possibly *accurately*) that selecting that response would signal positive character traits, even though the response does not accurately reflect their metaethical stance.

Conversely, if selecting a particular metaethical response would signal undesirable attributes, such as a normative moral stance the participant rejects, or an attitude that could harm their reputation, they may shy away from such responses, again, not because such responses accurately reflect their metaethical stances, but because they want to avoid signaling undesirable information about themselves.

Here, I will focus exclusively on the latter possibility. For instance, expressing an antirealist stance towards a serious moral transgression could signal inadequate opposition or insufficient

repugnance, not merely in arcane academic disputes, but in real-world social contexts. Imagine Alex is a politician who has just launched her first political campaign. The public doesn't know who she is. She has studied metaethics, and privately endorses moral antirealism. Catching wind of this, her political rival Sam directly asks her in front of a crowd:

*"Alex, do you think it would be objectively good to prevent school shootings? Yes or no?"*

Technically speaking, the honest answer would be "no." And yet it seems obvious to me at least that to respond with a "no" would instantly end Alex's political career. To say "no," would not lead her audience to express curiosity about the subtleties of her antirealist position. They would not pause to consider the possibility that she could be equally or more opposed to school shootings than Sam, is just as horrified by shootings as Sam, and wants just as strongly to prevent them. Instead, they would respond with immediate outrage, disgust, and horror. At best, people would infer that Alex isn't especially concerned with school shootings. At worst, they may conclude that Alex is a raving psychopath who is utterly unconcerned with the lives of children.

This example highlights a far more important reason for studying normative entanglement than its methodological implications: if normative entanglement occurs in lab settings, it would not only illustrate people's sensitivity to pragmatics in an experimental context, it would point to a potential *social* and *reputational* role metaethical discourse could have in everyday thought and discourse. And it highlights the possibility that pragmatic concerns like these could generalize to moral psychology as a whole, and indeed, *all* of human psychology.

Researchers should be doubly concerned, and doubly attentive to two concerns: first, inattention to the role pragmatics can play in how participants interpret and think about experimental stimuli could reflect a ubiquitous, subtle, and serious threat to the validity of existing research. Second, much of how people think and speak in everyday interactions is plausibly driven not by a blind adherence to strictly stating what seems most logically consistent and philosophically astute, but to

comporting one's thought, language, and action in accordance with one's *goals*, many of which are deeply embedded in our social interests, e.g., maintaining a positive reputation, ensuring our success and the success of our friends and family, outcompeting our rivals, improving our status, and identifying enemies and allies. Given that ordinary people's interests are far more aligned with practical rather than philosophical goals we should, if anything, expect the ordinary person to be *more* sensitive to normative entanglement and related concerns than philosophers. Given the frequency of such ploys even in academic disputes, the failure of researchers to grapple with the reputational and social costs of endorsing or rejecting a metaethical position, with all the attendant implications (real or imagined) associated with doing so, is a serious shortcoming with research on metaethics.

Of course, all this pulpiteering is irrelevant if it turns out participants aren't sensitive to normative entanglement. Study 4 assesses this question by testing whether comparing perceptions of three different responses to a direct question about whether a particular action is morally wrong. Participants were presented with a hypothetical exchange where one person, Alex, asks another person, Sam, whether an action is objectively morally wrong. In the realist condition Sam responds by stating that yes, they do think the action is objectively wrong. Other participants were assigned to one of two antirealist conditions: antirealism without cancellation, and antirealism with cancellation. In the condition without cancellation, Sam responds by simply stating that no, they do not think the action in question is objectively wrong. In the cancellation condition, Sam likewise states that no, they do not think the action is objectively wrong, but they then attempt to negate any inferences others might make about what this means by adding that they do think the action is wrong, they just don't think anything is *objectively* wrong. They then add that they are opposed to the action and find it terrible. These additional remarks attempt to *cancel* any implication that the antirealist lacks an appropriate normative moral stance, and fails to exhibit an appropriately negative attitude towards the action in question.

I predicted that participants will successfully perceive Sam to endorse realism (and universalism) and to reject relativism in the realism condition, and to likewise successfully perceive Sam to reject realism and universalism and endorse realism in both antirealist conditions. My hypothesis relies on this outcome, since it is especially critical that participants continue to perceive Sam as endorsing antirealism in the cancellation condition. In addition, participants were asked to judge Sam's moral character, trustworthiness, and empathy, how judgmental Sam is, how committed Sam is to their values, how desirable Sam would be as a social partner, and to assess how religious and politically liberal or conservative Sam is. I predicted participants would perceive Sam in the realist condition to be more morally good, more trustworthy, more empathic, more judgmental, more committed to their values, more desirable as a social partner, more politically conservative, and more religious compared to the antirealism without cancellation condition, but for the antirealist condition with cancellation to mitigate or completely negate several of these differences. In particular, I expected antirealism with cancellation to prompt participants to perceive Sam as having similar moral character, trustworthiness, and empathy, and to be similarly desirable as a social partner as Sam in the realist condition, but to have comparatively less of an effect on perceptions of how judgmental Sam is, how committed they are to their values, and how conservative and religious they are. In other words, I expected cancellation to buffer antirealists against the perception that they have worse moral character, are untrustworthy, and lack empathy, but not necessarily to influence judgments about other features of their beliefs and attitudes that are more plausibly associated with antirealism as a metaethical position. However, this does *not* mean that I expected cancellation to fully eliminate any negative character judgments, such that participants would perceive Sam's character identically in the realism and cancellation conditions. While this is possible, I predicted *either* that there will be no difference between these conditions, or, at the very least, that the cancellation condition will be intermediate between the realist and no cancellation conditions.

I am less confident about the mechanism that would drive these effects (if they are real). Much of the preceding account relies on the presumption that pragmatics, and in particular conversational implicature, could play a central role in how participants think about and respond to such questions. While these factors could be driving responses, there is considerable room for the results to emerge even in the absence of pragmatics per se. It could be that, at least in the experimental context of this particular study, that nonphilosophers who lack adequate knowledge and training in certain technical distinctions are being asked to assess the meaning of an exchange when they simply lack adequate background knowledge and context to competently do so. If so, results could be driven by task-specific performance error.

Even so, pragmatics could very well be driving this or similar events in everyday interactions. Here's why. Suppose one person asks another if an action is objectively wrong. Ordinarily, when one's normative moral standards are made salient, we might expect it to be important for a person to confirm what those standards actually are. In other words, when asked whether e.g., murder is objectively wrong, we might expect that, whatever else that person says, they should at least state that they think murder is wrong. Failure to do so could conversationally imply that they don't think murder is wrong. Since it would be strange to avoid an implicature that would have serious reputational costs, people may expect someone who is asked about their metaethical stance and their normative stance simultaneously to respond to both, *even if* the question is framed in such a way that the question appears to be only explicitly solicit a response about the respondent's metaethical position. Suppose Alex asks Sam:

Alex: *"Do you think your spouse is one of the fifty most attractive people in the world?"*[113].

We might raise an eyebrow if Sam were to simply respond, "No, I do not." Such a response may be a direct response to a literal interpretation of the question, but it could readily be interpreted to imply

---

[113] David Moss proposed an example of this kind to illustrate the point. I have adapted it for present purposes.

that Sam doesn't find their spouse to be especially attractive (even though they could think their spouse is the 51st most attractive person in the world, or at least very attractive). We expect Sam to make some remark about how attractive their spouse is, either by stating that, yes, they do think their spouse is in the top fifty (or is the most attractive of all), perhaps in a lighthearted tone that conveys that they are not interpreting the question literally nor responding literally, or they may not directly answer the question at all and merely state that their spouse is attractive, or they may reject the presumption that you can rank people in this way and maintain that beauty is in the eye of the beholder, perhaps affirming that they consider their spouse the most attractive person in the world (though they should be careful in doing so not to insinuate that their appeal to subjectivism isn't a dodge to avoid implying that their partner is unattractive), or, finally, they may say "No," but qualify this by affirming their spouse's attractiveness. As these many possibilities illustrate, the way we respond to such questions often involves a sensitivity to social context, and people are motivated to navigate such questions with a motivation not simply to respond in the most flat-footed and literal way possible, but with a sensitivity to what would be implied by what they say and, often just as importantly, by what they *don't* say.

I suspect a similar implication by omission could drive negative perceptions of antirealist responses, discouraging participants from selecting responses intended to convey antirealism in studies, and discouraging both philosophers and ordinary people from endorsing antirealist positions in academic and everyday contexts. Critically, research on the psychology of metaethics rarely presents participants with the ability to qualify their endorsement of antirealism by allowing them to simultaneously affirm their normative moral stance or their attitude towards the action in question[114].

---

[114] Even when participants are asked both about their normative stance and their metaethical stance towards an issue, these questions are asked separately and independently of one another. This means their response to a metaethical question is still straightforward and without qualification, absent contextualization with the rest of their responses. As such, even these studies are still vulnerable to lack of qualification.

As such, selecting an antirealist response option typically leaves participants with no choice but to risk signaling undesirable character traits. If my predictions are born out, this would raise new questions about the validity of existing paradigms.

**Methods**

*Participants.* I aimed to recruit 900 participants. 901 participants began the study. However, one participant reported being 3 years old and was excluded from further analysis. This left a total of 900 participants. Participants consisted of 900 adult US residents on Amazon's Mechanical Turk (405 females, 490 males, 4 other, 1 unreported, $M_{age}$ = 41.8, $SD_{age}$ = 12.4, age range = 19-85, 1 unreported age).

*Procedure.* All participants were randomly assigned to one of three response conditions: (1) realism, (2) antirealism without cancellation, or (3) antirealism with cancellation. In all conditions, participants were asked to consider a conversation between two people named Alex and Sam. In all conditions, Alex asks Sam whether an action is objectively morally right or wrong:

> *Alex: "Sam, do you believe it is objectively wrong [to torture and kill a baby just for fun]?"*

In each condition, Sam's response was displayed in bold immediately following the question. In the realism condition, Sam said "Yes," and affirmed that the action is "objectively wrong." In the antirealism condition Sam said "No," and denied the action is objectively wrong. Finally, in the antirealism with cancellation condition, Sam said "No" and denied the action is objectively wrong but then followed this with additional remarks stating that, although Sam doesn't find the action objectively wrong, Sam does think the action is morally wrong, but doesn't believe in objective morality. Sam then added that they consider the action terrible and are opposed to it. The exact wording for Sam's response in each condition is as follows:

Realism
**Sam: "Yes, I believe it is objectively wrong [statement]."**

Antirealism without cancellation

**Sam: "No, I don't believe it is objectively wrong [statement]."**

Antirealism with cancellation

**Sam: "No, I don't believe it is objectively wrong [statement]. But I do think it is morally wrong. I just don't believe morality is objective. I still find it to be terrible and I am deeply opposed to it."**

In addition to being randomly assigned to one of the three conditions, participants within each condition were randomly assigned to one of three distinct moral issues. Two issues were identical to those used in Study 2:

*It is morally wrong for a person to go to a funeral to mock the deceased person in front of their family.*

*It is morally wrong for a woman who knows she is pregnant to drink alcohol.*

Normative entanglement is most likely to occur when the moral transgression in question is perceived as serious. Of the four items, these appeared to be the most serious transgressions. However, I devised an especially serious transgression specifically for this study:

*It is morally wrong to torture and kill a baby just for fun.*

Next, participants were told to "Think about what Sam said in this scenario," and were then asked to judge what Sam's metaethical beliefs are using the same measures used in Study 1 (except for the noncognitivism items) and to assess Sam's character, attitudes, beliefs, and desirability as a social partner. Finally, participants were asked to provide their age and gender.

**Measures.** Participants assessed Sam's metaethical beliefs using the same three-item subscales used in Study 1 for realism, universalism, and relativism (but not noncognitivism). They were also asked to assess Sam's moral character (1 = Very morally bad, 7 = Very morally good), trustworthiness (1 = Very untrustworthy, 7 = Very trustworthy), empathy (1 = Not empathy at all, 7 = Very empathic), how judgmental Sam is (1 = Not judgmental at all, 7 = Very judgmental), how committed Sam is to their values (1 = Not committed at all, 7 = Very committed), how liberal or conservative Sam is (1 =

Very liberal, 7 = Very conservative), how religious Sam is (1 = Not religious at all, 7 = Very religious), and how desirable Sam is as a social partner (1 = Not at all, 7 = Very good).

**Results**

Although I employed three distinct moral issues, most of the differences in the dependent variables were not statistically significant. I collapsed all measures across the three moral issues and analyzed results only across the realism, antirealism without cancellation, and antirealism with cancellation conditions.

First, I assessed whether participants attributed metaethical stances in each condition consistent with what the speaker stated. If the manipulations were successful, this would mean that participants would judge the speaker in the realism condition to endorse realism and reject relativism (I also expected participants to attribute universalism to them, given its consistent, close association with realism). Conversely, participants should judge the speaker in the antirealism without cancellation condition to endorse relativism, but to deny realism (and universalism). Critically, participants should also judge the speaker in the antirealism with cancellation condition to endorse relativism and to deny realism (and universalism). Cancellation is only intended to cancel the normative implications of rejecting realism, i.e., that the person in question may be indifferent to or fail to exhibit appropriate opposition to the moral transgression. It is not intended to override the perception that the person in question nevertheless endorses an antirealist metaethical stance, and rejects a realist stance. If it did, this would mean that the manipulation failed, since the cancellation condition failed to disentangle attributions of metaethical and normative stances.

I conducted a one-way ANOVA to compare the effect of condition on realism. Consistent with predictions, there was a significant main effect of condition on attribution of realism $F(2, 576) = 75$, $p < .001$. As predicted, the mean realism score for the realist condition ($M = 5.38$, $SD = 1.10$) was significantly different than the mean realism score for both the antirealism without cancellation

condition ($M = 4.39$, $SD = 1.51$), $t(542) = 9.18$, $p < .001$ and the antirealism with cancellation conditions ($M = 4.19$, $SD = 1.59$), $t(523) = 10.70$, $p < .001$. Also consistent with expectations, there was *no* significant difference in realism attributions when comparing the antirealist conditions, $t(590) = 1.59$, $p = .251$.

Next I conducted a one-way ANOVA to compare the effect of condition on universalism. Consistent with predictions, there was a significant main effect of condition on attribution of universalism $F(2, 565) = 182$, $p < .001$. As predicted, the mean universalism score for the universalism condition ($M = 5.40$, $SD = 1.21$) was significantly different than the mean universalism score for both the antirealism without cancellation condition ($M = 3.28$ $SD = 1.77$), $t(524) = 17.1$, $p < .001$ and the antirealism with cancellation conditions ($M = 3.67$, $SD = 1.94$), $t(492) = 13.06$, $p < .001$. Contrary to predictions, there was also a significant difference between antirealist conditions $t(586) = -2.57$, $p = .028$. However, consistent with expectations, the difference in means was fairly small. For the no cancellation condition, the 95% CI centered on the mean of 3.28 was [3.08, 3.48] while for the cancellation condition the 95% CI was [3.45, 3.90]. The upper bound of the no cancellation condition just barely passes the lower bound of the cancellation condition. Thus, while the difference between conditions may have been statistically significant, and may indicate that participants do differ in the degree to which they attribute universalism to the antirealist in the cancellation and no cancellation conditions, this difference is likely *much* smaller than the difference between these conditions and the realist condition, and still suggests that participants perceive the antirealist in both conditions far more similarly to one another than to the realist. This is apparent by examining **Figure 6.4**. Examination of the differences between the conditions merely suggests cancellation was not *perfect*, and that's not something strictly required (or even expected) of my hypothesis.

**Figure 6.4**

*Study 4: Comparison of means (with 95% CIs) for universalism*



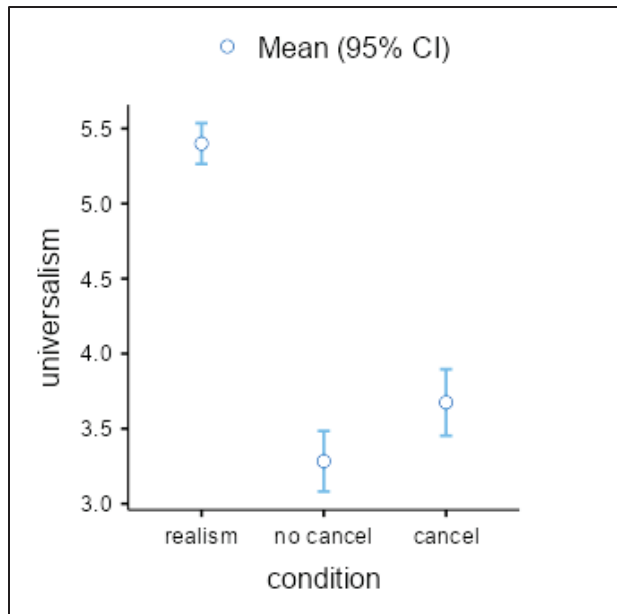For the final test to assess proper metaethical attribution I conducted a one-way ANOVA to compare the effect of condition on relativism. Consistent with predictions, there was a significant main effect of condition on attribution of relativism $F(2, 597) = 85.1, p < .001$. As predicted, the mean relativism score for the realist condition ($M = 3.24, SD = 1.45$) was significantly different than the mean relativism score for both the antirealism without cancellation condition ($M = 4.53, SD = 1.47$), $t(601) = -10.90, p < .001$ and the antirealism with cancellation conditions ($M = 4.65, SD = 1.54$), $t(595) = -11.53, p < .001$. Also consistent with expectations, there was *no* significant difference in relativism attributions when comparing the antirealist conditions, $t(590) = -0.96, p = .604$.

These results demonstrate a consistent tendency for participants to attribute metaethical stances across conditions as expected. Next, I assessed how participants judged Sam's character and non-metaethical beliefs by condition. First I conducted a one-way ANOVA to compare the effect of condition on judgments of morally good or bad Sam is. Consistent with predictions, there was a significant main effect of condition on judgment of moral character $F(2, 582) = 236, p < .001$. As

predicted, the mean score for judgment of moral character was significantly different in the realist condition ($M = 5.54$, $SD = 1.14$) than the mean score for judgment of moral character in the antirealism without cancellation condition ($M = 3.02$, $SD = 1.66$), $t(524) = 21.7$, $p < .001$. Although I made no decisive prediction either way, the mean moral character score was also significantly different than the antirealism with cancellation condition ($M = 4.86$, $SD = 1.36$), $t(575) = 6.66$, $p < .001$. Also consistent with expectations, there was a significant difference in mean judgments of moral character when comparing the cancellation and no cancellation conditions, $t(571) = -14.79$, $p < .001$. Notably, all of these significant differences were in the expected direction: as expected, participants judged the realist to be the most morally good, followed by the antirealist who canceled their rejection of moral realism, and then finally the antirealist who did not cancel their rejection of moral realism. While I was uncertain whether cancellation would *completely* cancel comparatively worse character judgments, it came very close. As can be seen in **Figure 6.5**, cancellation appears to have substantially mitigated negative character evaluation, even if it did not completely eliminate it.

Judgments about trustworthiness exhibited a similar pattern. Once again I conducted a one-way ANOVA to assess the effect of condition on judgments of trustworthiness. Consistent with predictions, there was a significant main effect of condition on judgment of trustworthiness $F(2, 575) = 151$, $p < .001$. As predicted, the mean score for trustworthiness was significantly different in the realist condition ($M = 5.43$, $SD = 1.16$) than the mean score for judgment of moral character in the no cancellation condition ($M = 3.32$, $SD = 1.77$), $t(510) = 17.4$, $p < .001$. Although I made no decisive prediction either way, the mean trustworthiness score was also significantly different than the cancellation condition ($M = 4.86$, $SD = 1.51$), $t(553) = 5.18$, $p < .001$. Also consistent with expectations, there was a significant difference in judgments of trustworthiness between the cancellation and no cancellation conditions $t(579) = -11.47$, $p < .001$. Once again, all differences were in the expected direction, with participants judging Sam to be most trustworthy in the realist condition,

somewhat less trustworthy in the cancellation condition, and much less trustworthy in the no cancellation condition.

**Figure 6.5**

*Study 4: Comparison of means (with 95% CIs) for moral character judgments*



The same pattern also held for perceived empathy. Consistent with predictions, there was a significant main effect of condition on perception of empathy $F(2, 588) = 173, p < .001$. As predicted, the mean empathy score was significantly different in the realist condition ($M = 5.12, SD = 1.33$) than the no cancellation condition ($M = 2.78, SD = 1.76$), $t(552) = 18.4, p < .001$. Once again, although I made no decisive prediction, the mean empathy score was also significantly different than in the cancellation condition ($M = 4.60, SD = 1.50$), $t(586) = 4.49, p < .001$. Also consistent with expectations, there was a significant difference in judgments of empathy between the cancellation and no cancellation conditions $t(579) = -13.53, p < .001$. As before, all differences were in the expected direction, with participants judging Sam to be most empathic in the realist condition, somewhat less empathic in the cancellation condition, and much less empathic in the no cancellation condition.

Next, I assessed the degree to which participants perceived Sam to be judgmental across conditions. Consistent with predictions, there was a significant main effect of condition on perceived judgmentalness $F(2, 588) = 23.2$, $p < .001$. As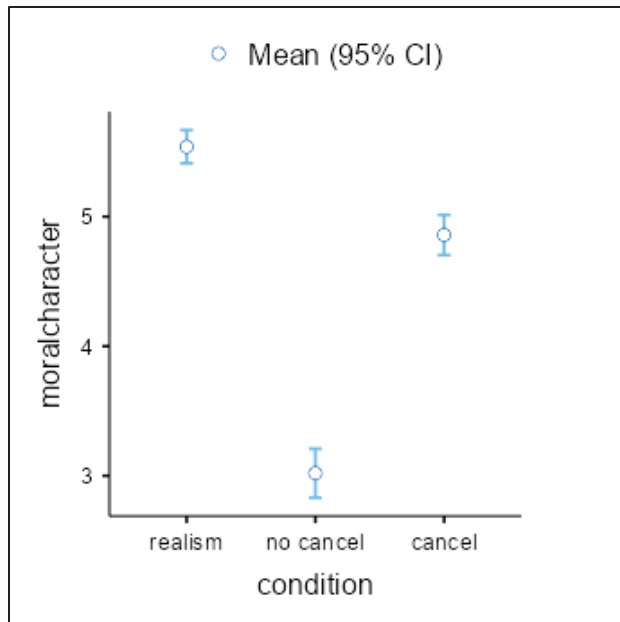 predicted, the mean judgmentalness score was significantly different in the realist condition ($M = 4.42$, $SD = 1.37$) than the no cancellation condition ($M = 3.66$, $SD = 1.78$), $t(558) = 5.89$, $p < .001$, and the cancellation condition ($M = 3.76$, $SD = 1.61$), $t(579) = 5.46$, $p < .001$. I expected cancellation to *not* fully eliminate differences between the realism and cancellation conditions. This is because I anticipated that stating that an action is objectively wrong would plausibly be seen as judgmental, but stating that it isn't objectively wrong, and that in fact one does not think *anything* is objectively wrong, but that one still opposes it would be perceived as *less* judgmental. This was born out in the results. However, it was less clear whether there would be any significant difference between the cancellation condition and the no cancellation condition. As such, I expected the cancellation condition to at least be intermediate between the realist and no cancellation conditions. Thus, my expectations were consistent both with the possibility that the cancellation and no cancellation difference were significantly different or not significantly different. So long as both appeared to have lower mean scores than the realist condition, whether they differed from one another is incidental to my main hypotheses. It turned out that there was no significant difference between the cancellation and no cancellation conditions $t(587) = -0.69$, $p = .770$. This suggests that cancellation (or at least the precise cancellation wording used in this study, for the particular moral issues it was used for) has little or no impact on how judgmental a person is perceived when they qualify an expression of antirealism with additional remarks about their opposition to the issue in question.

I next performed a one-way ANOVA to assess perceived commitment to their moral values across conditions. Consistent with expectations, there was a significant main effect of condition on perceived commitment to moral values $F(2, 578) = 68.7$, $p < .001$. As predicted, the mean perceived

commitment score was significantly different in the realist condition ($M = 5.67$, $SD = 1.11$) than the no cancellation condition ($M = 4.33$, $SD = 1.67$), $t(516) = 11.6$, $p < .001$, and the cancellation condition ($M = 5.10$, $SD = 1.40$), $t(562) = 5.56$, $p < .001$. Once again, I expected the cancellation condition to be at least intermediate between the realist and no cancelation conditions for perceived commitment (i.e., significantly less than realism, and significantly more than no cancellation), but I did not predict whether it would fail to significantly differ from the no cancellation condition. Consistent with expectations, the cancellation condition did significantly differ from the no cancellation condition in the expected direction, $t(576) = -0.606$, $p < .001$. Once again, the mean score for the cancellation condition was intermediate between the realist and no cancellation condition, in line with expectations.

Next, I assessed whether participants judged Sam to be a good person to have as a social partner, such as a coworker or a friend. Consistent with predictions, there was a significant main effect of condition on partner preference $F(2, 577) = 148$, $p < .001$. As predicted, partner preference was significantly different in the realist condition ($M = 5.14$, $SD = 1.20$) than both the no cancellation condition ($M = 3.01$, $SD = 1.79$), $t(516) = 17.2$, $p < .001$, and the cancellation condition ($M = 4.56$, $SD = 1.56$), $t(553) = 5.14$, $p < .001$. I expected the cancellation condition to be at least intermediate between the realist and no cancelation conditions for partner preference, such that Sam would have the highest mean partner preference score in the realist condition, and the lowest in the no cancellation condition, with the cancellation condition either being intermediate or not significantly different than the realist condition. Consistent with this expectation, partner preference was significantly different in the cancellation and no cancellation conditions in the expected direction, $t(582) = -11.29$, $p < .001$. Once again, the mean score for the cancellation condition was intermediate between the realist and no cancellation condition.

After this, I assessed differences in judgments about whether Sam was more politically liberal or conservative across conditions. Consistent with expectations, there was a significant main effect of

condition on perceived political orientation $F(2, 587) = 22.3$, $p < .001$. As predicted, Sam's perceived political orientation was significantly different in the realist condition ($M = 4.22$, $SD = 1.26$) than both the no cancellation condition ($M = 3.47$, $SD = 1.61$), $t(563) = 6.29$, $p < .001$, and the cancellation condition ($M = 3.70$, $SD = 1.57$), $t(566) = 4.47$, $p < .001$. Consistent with expectations, there was no significant difference in perceived political orientation for the cancellation and no cancellation conditions $t(592) = -1.71$, $p = .203$. Overall, people perceived Sam in the realist condition to be more inclined towards political conservatism than in either antirealist condition, while the antirealist conditions did not significantly differ from one another. This suggests that cancellation did not substantially change how participants perceived Sam's political orientation and that, as expected, those who endorse moral realism (at least for the moral issues used in this study) are perceived as more disposed towards political conservatism.

Lastly, I assessed differences in perceived religiosity across conditions. Consistent with expectations, there was a significant main effect of condition on perceived religiosity $F(2, 585) = 107$, $p < .001$. As predicted, perceived religiosity was significantly different in the realist condition ($M = 4.31$, $SD = 1.28$) than both the no cancellation condition ($M = 2.56$, $SD = 1.65$), $t(559) = 14.50$, $p < .001$, and the cancellation condition ($M = 3.47$, $SD = 1.63$), $t(559) = 7.04$, $p < .001$. Contrary to expectations, there was a significant difference in perceived religiosity between the cancellation and no cancellation conditions $t(592) = -6.74$, $p < .001$. However, consistent with expectations, perceived religiosity for the cancellation condition was intermediate between the realism and no cancellation conditions. Overall, I didn't expect cancellation to substantially increase perceived religiosity, though it appears that it did. Nevertheless, perceived religiosity fell midway between the realism and no cancellation conditions, suggesting that while the additional cancellation remarks may have increased perceived religiosity, they did not so as much as expressing realism does. What would have been genuinely shocking and contrary to my expectations would be if the cancellation condition was not

significantly different from the realism condition, or was significantly larger. This did not occur, so while the overall findings are not precisely what was anticipated, they are still in line with the generally expected pattern, with the cancellation being intermediate between realism and no cancellation.

**Discussion**

Overall, the results of this study were highly consistent with the proposed phenomenon of *normative entanglement*. When asked whether a serious moral transgression is objectively wrong, people who express a realist stance, compared to those who bluntly express antirealism without qualification, are perceived as more morally good, more trustworthy, more empathic, more committed to their values, more judgmental, better social partners, and as more politically conservative and religious. Yet precisely those effects that were expected to be mitigated or eliminated when the antirealist attempts to cancel their rejection of moral realism with an expression of their opposition to the transgression were, attenuated or eliminated, including moral character, trustworthiness, and empathy, as well as quality as a social partner, yet a person who expressed antirealism with cancellation was still perceived as endorsing relativism and rejecting realism and universalism, and was still regarded as less committed to their moral values than the realist. These findings suggest that cancellation can function to attenuate the negative effects of endorsing antirealism without qualification, while leaving other judgments about the person intact.

Discovering *why* cancellation has an ameliorative effect on perceptions of the moral character of antirealists is a task for future research. While I suspect this effect may be driven by conversational implicature, it could also be driven by simple performance errors or unintended interpretations of the stimuli. If so, it could be an artifact of experimental design. However, insofar as similar remarks are made in ordinary circumstances, such errors may likewise occur in the real world. As such, even this latter possibility doesn't rule out a potentially important discovery.

Finally, these results should be qualified by another observation. I have been careful to frame judgments about the antirealist who does not cancel their rejection of realism as inferences that are not *logically entailed* by an endorsement of antirealism. This isn't merely because they aren't entailed, but could be *mistakenly* perceived to be implied by expressions of antirealism. Rather, to reject moral realism in response to a question that involves normative entanglement technically provides *some* evidence, all else being equal, that the person in question *really does* endorse both the metaethical question (whether the act in question is *objectively* wrong) *and* the normative claim (that the act in question is *wrong* simpliciter). This is one of the most insidious and unfortunate aspects of normative entanglement. Normative entanglement amounts to a kind of complex question, where a person is explicitly asked a question about their metaethical stance towards a moral act, but, technically, they are *also* asked about their *normative* stance…*conditional* on whether they are a realist about the moral issue in question. If asked whether torturing babies just for fun is objectively wrong, if Sam says "Yes," Sam is in fact expressing *both* a metaethical position *and* a normative moral position:

(1) Torturing babies just for fun is morally wrong (normative)

(2) This normative fact is objectively true (metaethical)

Yet if Sam says "No," this is technically only a denial of (2), not (1). There is, therefore, an asymmetry between saying "Yes" and "No": saying yes *explicitly* addresses both questions, while saying "No" only explicitly addresses one of the two questions. This asymmetry creates a kind of pragmatic vacuum that could be filled by implicatures about the speaker's normative stance. Yet setting aside what may be conversationally implied by a "No," observe that if someone says "No," this provides, in the absence of cancellation, some genuine evidence about their normative moral stance: relative to a person who says yes, and thereby explicitly expresses a normative moral stance (namely, that the action is wrong), they are more likely not to endorse a contrary moral stance (at the very least, that the action is *not* wrong). For comparison, imagine Alex asks Sam:

*Do you like pepperoni pizza?*

If Sam says "Yes," this indicates both that Sam likes pepperoni pizza *and that Sam likes pizza*. But suppose Sam says "No." While this would indicate that Sam does not like pepperoni pizza, it does not follow that Sam dislikes pizza *without* pepperoni. Consider the epistemic asymmetry between the two conditions: We know that if Sam says "Yes," that Sam likes Pizza, but if Sam says "No," we still don't know whether Sam likes or dislikes pizza (without pepperoni). Yet we do know one thing: if Sam says "No," Sam is *less likely* to like pizza than if Sam says "Yes." If we think like a good Bayesian, to say "No" to this question should cause us to reduce our confidence that Sam likes pizza. A similar epistemic asymmetry is present in cases of normative entanglement. To oversimplify for demonstrative purposes, there are four positions a person could have with respect to the metaethical status, and the normative status, of an act:

| | Morally wrong | Not morally wrong |
|---|---|---|
| Realism | The action is objectively wrong | The action is objectively not wrong |
| Antirealism | The action is wrong, but is not objectively wrong | The action is not wrong and is not objectively wrong |

A person who responds with a "Yes" puts themselves in the top left category, while a person who says "No" could endorse *any* of the other three positions. However, only the two positions in the right column indicate that one does not hold the normative moral stance that the action is wrong. Thus, even if a person endorses the bottom left position, saying "No," does not clearly indicate this in the way saying "Yes" indicates the top left position. Since a "Yes" allows us to exclude endorsement of either position on the right, but saying "No," doesn't, to say "No," provides at least *some* evidence, relative to the counterfactual in which a person says "Yes," that the person does *not* hold the normative

moral stance that the action in question is wrong. Whether or not, and to what extent ordinary people are responsive to such evidence is a matter of future empirical inquiry.

**6.5 General discussion**

Given the somewhat disparate nature of the three main studies reported here, most of the relevant discussion appears in the discussion sections for each of those studies. Nevertheless, there are a few central findings worth emphasizing. First, Study 1 demonstrated that judgments of realism, universalism, and relativism are not the same when participants are judging the metaethical stances of other people compared to when they report their own metaethical stance. There is at best only a weak correlation between third and first person judgments, suggesting that people's judgments about other people's metaethical standards are not simply a projection of their own metaethical stances. The disparity between first and third person judgments is an important finding, since it vindicates the importance of comparing first and third person judgments, and raises questions about relying exclusively on first person judgments.

While I favor normative conflations as a proposed mechanism behind at least some of this disparity, I did not directly test for the effects of normative conflations. However, this wasn't the goal of the present set of studies. Instead, the goal was to devise a set of measures that could first point to a disparity in need of explanation. Having established such a disparity, future efforts will be directed at explaining its cause. The pattern of results I found is consistent with normative conflation accounting for at least some of the disparity between third and first person judgments. Study 1 revealed that participants judged other people to exhibit greater support for realism and universalism, and less support for relativism, than themselves. This *could be* driven, in part, by normative conflations. Participants may be disinclined to endorse realism and universalism, and more inclined to endorse relativism, because doing so could signal an undesirable normative moral stance, or undesirable character traits, such as rigidity, intolerance, and close-mindedness.

However, I suspect an additional factor is in play. The participants in these studies were less politically conservative, and much less religious, than the population of the United States as a whole. Since each of these factors is linked with greater support for realism and rejection of relativism, it seems possible that the participants that were sampled were not representative of the population they were drawn from (i.e., adults in the United States). If so, participants may have both accurately reported their own metaethical stances *and* accurately judged a typical person in their society to be more inclined towards realism and universalism, and less inclined towards relativism. After all, the typical person is more politically conservative, and much more religious, than participants themselves. In imagining what other people think, participants may be responsive to these differences, and correctly recognize that their own metaethical stance may differ from the general population.

One way these findings corroborate previous research is that the same pattern of realism and relativism emerged across moral and nonmoral domains as prior research using first person measures (e.g., Goodwin & Darley, 2008; Wright, Grandjean, & McWhite, 2013). Participants judged others to be most inclined to support realism and reject relativism for facts, followed by morals, social conventions, and matters of taste. While this corroborates the *order* of judgments about realism and relativism with respect to these domains, it remains an open question whether or not, and to what extent, any given population is more inclined towards realism or antirealism about morality in particular. The factual and taste domains serve largely to anchor the assessment of different domains, given that most studies find the former to exhibit high levels of realism and the latter to exhibit low levels of realism. The findings reported here are consistent with that pattern. Social conventions stand out as an anomaly, however. Previous research has likewise failed to pin down how people think about the metanormative status of social conventions. Social conventions have historically straddled a middle ground between morality and taste.

The current findings are consistent with this pattern, yet it was surprising that I also found that people judged others in their society to be more inclined towards realism and universalism, and less inclined towards relativism for social conventions than themselves. I am unsure why this is the case. On the one hand, the unsatisfyingly equivocal results of previous research, coupled with these findings, point to the potential for a rich and interesting possibility for research on metanormative psychology outside the moral domain. On the other hand, I suspect these findings may be due to systemic patterns of unintended interpretations when questions about metanormativity are directed at matters of social convention.

Finally, Study 4 points to a potentially significant phenomenon: normative entanglement. Current findings suggest that when a person responds to a question about whether a serious moral transgression is "objectively wrong," by stating that it is, people have a fairly positive appraisal of such a person, especially compared to a person who denies that the action in question is objectively wrong. Yet current findings indicate that, at least among the participants that were sampled, when someone who initially responded by simply denying that the act in question was objectively wrong, but then qualified that remark by also expressing their opposition and negative attitude towards the act in question, most of these negative character judgments were nearly eliminated, while participants continued to regard the person who made such a remark as a moral antirealist. These findings suggest that people who bluntly express moral antirealism may be perceived extremely negatively, at least in contexts where questions about their metaethical standards are entangled with normative questions about serious moral transgressions.

When asked about their own metaethical stances, participants are rarely given the opportunity to qualify those remarks in this way. As a result, participants may be disinclined to endorse antirealist responses when they judge that doing so could carry reputational costs or when they judge those responses to also carry implications about their normative moral stance, and their attitudes and

emotional responsiveness to a given moral transgression. If so, studies that rely on concrete measures may underestimate support for antirealism, or may provide misleading evidence of "pluralism" (i.e., participants endorsing realism for some moral issues but not others) that is actually driven in part by variation in the degree to which participants judge an "antirealist" response to carry undesirable implications. More generally, while the observed effect may be limited to the stimuli and sample reported here, these findings point to the more general possibility that pragmatic factors may influence how participants respond to stimuli, potentially threatening the validity of other measures and providing the impetus for new research on the role pragmatics may play in how participants interpret psychological stimuli.

However, a more intriguing possibility, given these findings, is that the consequences of entangling metaethical and normative/social implications may also emerge outside the lab. Reasoning has traditionally been thought of as a tool for acquiring accurate beliefs about what the world is like. Yet Mercier and Sperber (2011; 2017) marshal an impressive array of arguments and evidence that reasoning may have evolved to construct and evaluate arguments and, more generally, to *persuade*. Indeed, it is also possible that, as Haidt (2001) argues, many moral judgments are not generated so much by reasoning as such, but are instead the output of automatic and intuitive processes, with reasoning largely to provide post hoc justification for one's moral conclusions.

If these findings are on the right track, they may explain why normative entanglement occasionally emerges in academic contexts and disputes among those without formal training in philosophy: such examples may illustrate a more *general* tendency for people to exploit the rhetorical impact of normative entanglement in interactions outside academic contexts. For instance, people may exploit realist-sounding language in a moral dispute in an attempt to smear a rival by implying that their rival fails to adequately oppose atrocities. If so, there is little reason to think such rhetoric would be limited specifically to normative entanglement. Rather, it could be that people strategically

express realist-sounding and anti-realist sounding language, or attribute realism or antirealism to others, in ways that serve their social and argumentative goals. Such rhetorical ploys could be highly effective even if the speaker has little or no understanding of the relevant metaethical positions as they are understood by academic philosophers. In short, to the extent that metaethical terms and concepts, or at least language that superficially resembles their use within academic philosophy, appears in everyday discourse, it may appear there for reasons unrelated to the genuine adoption of metaethical stances among nonphilosophers, or the robust comprehension of distinct metaethical positions.

If so, this could partially account for the poor validity of existing research on folk metaethics. It may even go some way in explaining why philosophers and psychologists have been slow to recognize that ordinary people have no determinate metaethical positions. Terms like "objective" and "relative" have colloquial analogs to their philosophical counterparts, unlike many of the technical terms used in e.g., physics or mathematics. Scholars studying metaethics may have developed the misperception that ordinary people have substantive, determinate metaethical stances due in part to the fact that ordinary people use similar terms to those used among academics. Such usage may not even have evolved independently; it's possible notions devised in the academy leak into public discourse. If so, there may be a degree of terminological and conceptual cross-pollination between academic and everyday contexts that bolsters the illusion that ordinary people are "realists" and "antirealists." Instead, it may be that ordinary people have a superficial or muddled understanding of the meaning of the relevant terms and concepts, and are employing them for largely argumentative purposes, *not* to express sober stances towards abstract philosophical issues about the metaphysics of morality. Of course, these considerations are speculative, and gathering support for them is a task for future research.

There are likely a number of limitations with these studies that have gone unnoticed, or that others will point out in due time. One significant limitation with these studies is how little they tell us. Thus, in describing future directions I am simultaneously addressing the study's central shortcomings.

First, future versions of the third person paradigm should develop more robust measures by expanding on the number of items that appear in each subscale, and make a more concerted effort to assess their psychometric quality by e.g., analyzing their factor structure and recruiting experts familiar with the relevant concepts to judge their validity. Assessing how ordinary people interpret these questions using the methods introduced in earlier chapters would likewise serve a valuable role in refining our measures. It would also be worthwhile to make a second attempt at devising valid measures of noncognitivism, though this may ultimately prove infeasible. Even if efforts to devise adequate measures for noncognitivism, future research could expand on the range of metanormative concepts by devising new subscales. This could include efforts to distinguish cultural relativism from individual subjectivism, naturalist from non-naturalist conceptions of moral realism, and, though it may prove difficult, less well-studied and potentially more complicated metaethical positions, such as error theory, constructivism, and ideal observer theory. Such efforts will be challenging, if not futile, at least using ordinary survey items. Such concepts may require more sophisticated paradigms. Researchers could also directly assess whether ordinary people explicitly endorse metaethical pluralism or indeterminacy, by directly or indirectly asking them about each.

Future research could also refine or expand on the third-personal aspect of the paradigm. One shortcoming with the current set of studies is that they provide little information about *who* participants had in mind when thinking about a typical person or judging the metaethical beliefs of a person with no information other than that they made a particular moral claim. Researchers could incorporate additional measures that assess what kind of person participants have in mind by asking them to describe or make additional judgments about the imagined person. Future studies could also provide

richer and more detailed information about the target participants are asked to judge. This could include demographic information, such as age, gender, or political leanings, or via richer statements or vignettes that provide more context.

There were also only a limited number of statement types: I only asked people to judge scenarios in which a person stated that an action was "right" or "wrong." Moral vocabulary is far more expansive than this. Future efforts could assess a wider range of moral statements, including statements that an action is "obligatory," "prohibited," "permissible," "required," "acceptable," and so on. Nuance could be added to such claims by including qualifiers, e.g., that an action is *always* prohibited, or *sometimes* permissible. Such qualifiers may mislead participants, but could also highlight ways in which participants interpret items in unintended ways, e.g., people may judge that a person who says that an action is *sometimes permissible* is more likely to endorse moral relativism, even though such a claim is logically consistent with moral realism and universalism.

It may also be of interest to ask participants to assess what metaethical standards people *outside* their culture endorse, or even to ask participants to speculate about the metaethical beliefs of nonhumans, e.g., extraterrestrials or artificial intelligences. While people would not, of course, have any way of knowing the answer to these questions, such findings could reveal features of how people think about the nature of morality that might otherwise be difficult to assess.

In addition to alternating features of the target of judgment in third person paradigms, it would also be helpful to recruit participants from more diverse backgrounds. The present studies only recruited MTurk participants, but results showed that, consistent with research on the demographic characteristics of workers on MTurk, these participants were less politically conservative and much less religious than the general population of the United States. Since these variables play a significant role in how people think about metaethics, the present studies do not generalize well to people in the United States. It would also be helpful to focus on members of particular religions (or even atheists

or agnostics), or to assess demographically distinct subpopulations, e.g., members of distinct religious communities such as Mormons, though it would be especially interesting to survey non-WEIRD populations (Henrich et al., 2010).

Future studies could also incorporate additional nonmoral domains, such as epistemic and prudential norms. It could also turn out that existing domains have not been conceptualized or operationalized adequately. It may be, for instance, that people do not treat aesthetic judgments about art and beauty the same way as food preferences. Some normative domains may also exhibit intradomain variation. There may be systematic differences in how people judge moral issues that reflect Haidt's moral foundations. Davis (2021) has already addressed this topic using an adapted version of the disagreement paradigm, though given its methodological shortcomings, a look with a new paradigm would be worthwhile. Other domains may likewise exhibit subdomains worthy of targeted inquiry, e.g., people may think differently about scientific claims versus other sorts of "factual" claims, such as math or religion.

It would also be helpful to expand on the number of concrete items in the moral domain. Study 2 only employed 4, yet future research could incorporate a broader and more representative sampling of distinct moral issues. In addition, there were *no* concrete measures for nonmoral domains. These domains are typically neglected relative to the moral domain, yet this could be addressed by providing sufficient attention to testing a well-designed set of concrete issues for these domains as well. It would also be worthwhile to explore differences in judgments about nonmoral domains using a within-subjects design.

Much of the present research focuses on identifying whether or not there is a disparity between third and prison judgments, and speculating about what might cause such a disparity. Yet the present findings are incapable of directly identifying what mechanisms could be driving this difference, or explaining why participants endorse various metaethical positions in general. While I will abstain from

sketching any particular research programs for identifying the factors that explain people's responses, I believe exploring these factors is one of the most critical pieces of the broader puzzle in the psychology of metaethics.

One way to contribute to this task, and that could be of considerable independent interest, would be to explore the relationship between responses to the third person paradigm and a variety of psychological variables that could be related, e.g., Big Five personality traits or other philosophical beliefs, such as belief in free will.

Finally, there are a number of avenues for future research that are distinct to Study 4. However, I should first acknowledge one limitation with this study. My main test of whether normative entanglement occurred consisted of a variety of measures assessing the character of Sam, the target who made the statement. While this is one element of normative entanglement, and the most important one from a social psychological perspective, I did not directly assess judgments about Sam's normative moral position. That is, I did not ask whether Sam thinks the action in question is morally right or wrong. While it may serve as little more than an attention check for the realism and cancellation conditions, since Sam explicitly states that they think the action is wrong, it is important to confirm that the proportion of participants who judge that Sam does *not* think the action is morally wrong is lower in the no cancellation condition compared to the other two conditions. I intend to rectify this lacuna in subsequent research. In addition, I plan to include a variety of additional measures designed to bolster the case for normative entanglement. In particular, I plan to ask participants whether Sam would be more or less likely to engage in the action in question, how Sam would react to someone committing the action, whether Sam would favor laws against the action, and whether Sam would want to punish people who commit the action. In short, I want to evaluate whether participants go beyond merely judging Sam's character negatively in the no cancellation condition, but

whether they also differ in their judgments about Sam's normative moral beliefs and Sam's attitudes and behavioral disposition.

Future research could also assess whether phenomena similar to normative entanglement occur in everyday circumstances. In particular, such studies could assess how people judge arguments that employ various types of metaethical language, in both direct ways and in ways designed to serve rhetorical purposes (e.g., implying that someone who does not endorse moral realism is evil). Do they find such arguments persuasive? Would they employ such arguments themselves? Do they employ these arguments when given the opportunity? Can we identify instances of their use among e.g., politicians, legal contexts, or academic work? Once we reorient our perspective on the functions moral language, judgment, and norms play in everyday life, whole new avenues of research open up. For instance, Sperber and Baumard (2012) maintain that one of the central functions of moral behavior is to develop a positive reputation in order to facilitate cooperation with others. They acknowledge that a genuine motivation to do what is morally good (or at least considered morally good in one's community) may be one of the surest ways to maintain a positive reputation. However, they argue that if, from an adaptive perspective, the target of such behavior is our reputation (rather than e.g., "doing the right thing"), it may be too costly to leave the fortunes of our reputation to our conscience alone. They propose that we may have an evolved predisposition to be *directly* concerned with managing reputation. If so, much of our motivation in moral disputes may be instrumental, with an eye towards navigating social hierarchies, and avoiding loss in power, status, and cooperative opportunities.

Such proposals synergize well with the argumentative theory of reasoning, and with the potential argumentative role of metaethical terms and concepts. In short, the claim that metaethical discourse may play a largely argumentative and rhetorical role has not emerged within a theoretical vacuum, but dovetails with emerging theories about the social function of moral reasoning and

behavior. Future studies can capitalize on this by directly assessing what role (if any) metaethical terms and concepts play in argumentative contexts and other contexts where people's reputation is at stake.

## 6.6 Conclusion

Previous chapters offered a bleak picture of the state of research on folk metaethics. This chapter offers a more optimistic outlook. I suspect the most constructive changes will emerge from a reorientation in our general approach to studying how ordinary people think about traditionally philosophical topics like moral realism and antirealism (and their nonmoral analogs). Rather than identifying the distinctions of interest to academic philosophers, directly operationalizing these into a set of measures, and then throwing them at participants with little or no context, then expecting people's responses to be informative, those studying the psychology of metaethics should appreciate that morality is a deeply social and practical enterprise. Real moral judgment and behavior doesn't occur in the imaginary worlds dreamed up by philosophers, where space pirates threaten annihilation and people must contemplate hurling their grandma into a volcano to save a dozen doctors. Real moral judgments involve messy situations that are contextually rich, and often feature people we know and care about. Our reputations are often on the line, and we stand to gain or lose a great deal if we say or do the wrong thing. Stripped of this context, much existing research on moral psychology is ecologically impoverished in ways researchers are only beginning to recognize and detect (see e.g., Navarro-Plaza et al., 2020).

The third person paradigm goes some way in slightly ameliorating these concerns by focusing judgments on others rather than asking for self-reports. Yet the real potential of the third person paradigm is that it can be expanded in the many ways highlighted here: we can add richer and more detailed context than is typical of much existing research in moral psychology. In addition, by driving a wedge between third and first person judgments, the disagreement paradigm could serve as the start of a research project designed to assess the role that normative conflations and other methodological

concerns may play in undermining the validity of previous research. Finally, Study 4 assesses the possibility that implications about a person's character may influence how participants respond to metaethical stimuli.

Once we internalize the fact that participants may be unable to simply turn off concerns about how others might perceive them based on their response to stimuli, we can begin to appreciate that participants may not be responding to metaethics stimuli like a philosopher, trained (with a questionable degree of success) to set aside their emotions, concerns about their reputation, and fears about the implications of facing the truth, to render a sober, hard-nosed judgment about the fundamental nature of reality. That just isn't how *ordinary* people think. And yet research on the psychology of metaethics has treated ordinary people as if they were philosophical automata capable of readily distinguishing abstract and unfamiliar philosophical concerns about, of all things, the metaphysical status of truth claims from far more familiar, mundane questions about whether they think an action is morally right or wrong. And we expect them to do that while holding in abeyance millions of years of evolution and decades of cultural conditioning that impel them to prioritize their reputation when considering moral issues.

This is, to put it mildly, absurd. Progress in the psychology of metaethics, and in moral psychology more generally, will be best served by recognizing that morality, as a psychological phenomenon, has not been accurately represented by focusing so much on the rarefied ratiocinations of an insular coterie of academics whose thought processes have been transmogrified by their shared induction into a canon of idiosyncratic thinkers. No, morality, *actual* morality, occurs in the real world, with its rich social context, and with all the attendant social motivations, reputational stakes, fear of shame and punishment, identifying cooperative opportunities, signaling of our commitment to our ingroups, distancing ourselves from outgroups, and coming out on top in disputes by winning arguments, that occur in that world.

## CHAPTER 7:

### Conclusion

Many readers will recognize that the title of this dissertation is a reference to Schrödinger's cat. Schrödinger's cat is a thought experiment devised to illustrate the bizarre and seemingly paradoxical nature of quantum mechanics (Schrödinger; 1935; Villars, 1986). Erwin Schrödinger imagined a scenario in which a macroscopic event—the fate of a cat—could be entangled with the outcome of a seemingly indeterminate quantum event. Wineland (2013) provides a summary of the scenario:

> Erwin Schrödinger [...] realized that, in principle, quantum mechanics should apply to a macroscopic system in a more complex way, which could then lead to bizarre consequences. In his specific example, the system is composed of a single radioactive particle and a cat placed together with a mechanism such that if the particle decays, poison is released, which kills the cat. Quantum mechanically we represent the quantum states of the radioactive particle as undecayed $=|\uparrow\rangle$ or decayed $=|\downarrow\rangle$, and live and dead states of the cat as $|L\rangle$ and $|D\rangle$. If the system is initialized in the state represented by the wave function $|\uparrow\rangle|L\rangle$, then after a duration equal to the half life of the particle, quantum mechanics says the system evolves to a superposition state where the cat is alive and dead simultaneously, expressed by the superposition wave function:

$$\Psi = \frac{1}{\sqrt{2}}[|\uparrow\rangle|L\rangle + |\downarrow\rangle|D\rangle].$$

> Schrödinger dubbed this an entangled state because the state of the particle is correlated with the state of the cat. That is, upon measurement, if the particle is observed to be undecayed, one can say with certainty that the cat is alive, and visa versa. But before measurement, the particle and cat exist in both states. (p. 1103)

For our purposes, the critical element of this thought experiment is that the cat exists in a state of superposition, simultaneously both dead and alive, and this state of indeterminacy collapses only when the wave function itself does. Prior to the collapse, *there is no determinate fact of the matter* about whether the cat is alive or dead.

I didn't draw on an analogy between this scenario and metaethical indeterminacy just because I found it amusing, or because I really wanted to include a bad pun in the title.[115] I drew on it for a different reason: I believe that it's time for philosophers to begin a similar paradigm shift. Physicists have had a century to grapple with the unsettling implications of quantum mechanics, and while they have yet to reach a consensus on how (or even whether) to interpret quantum mechanics, enough time has passed for them to have become comfortable with the idea that the world, as described by fundamental physics, is radically at odds with our ordinary ways of thinking and speaking. Yet contemporary analytic philosophy has yet to make a comparable shift. Philosophers have not adequately grappled with the possibility that the way they think about the world is also radically at odds with our ordinary ways of thinking and speaking.

I suspect that this is due to an entrenched methodological focus on the analysis of terms and concepts. Since the "linguistic turn," analytic philosophy developed a pathological obsession with language, and metaethics was swept up in this endeavor. MacIntyre (1957), one of central figures of 20th century metaethics, put it bluntly: "The central task to which contemporary moral philosophers have addressed themselves is that of listing the distinctive characteristics of moral utterances" (p. 325). Philosophers in the 20th century appeared to believe they could solve fundamental questions in ethics by studying how ordinary people talk, and the inertia behind this approach has continued into the present.

Stranger still, discovering the "distinctive characteristics of moral utterances" looks like an empirical task. Yet philosophers did not get out the clipboards and pens and start surveying people. Instead, they were content to address how ordinary people think and speak from the armchair, on the assumption that their education endowed them with unique insight into the nature of ordinary thought and language. I've never encountered any good reason to think armchair methods are adequate to

---

[115] Though I admit these motivations played some role.

determine how ordinary people speak or think about morality specifically, and the only way to decisively resolve such disputes is with empirical data, e.g., from linguistics, psychology, and anthropology.

I'm not alone in holding such a view. Experimental philosophy emerged in the 21st century as a proposed remedy to exclusively armchair analyses. Rather than waiting for psychologists to pick up the slack, experimental philosophers began conducting their own studies. Social scientists quickly followed suit, conducting their own studies on folk philosophy. Yet regardless of their disciplinary backgrounds, all of these researchers made the same fundamental mistake: rather than addressing the study of folk philosophy with a bottom-up approach, they came equipped with an arsenal of philosophical terms, concepts, and distinctions from a tradition that had never engaged in *any* meaningful empirical research on folk morality.

As a result, researchers studying folk philosophy have taken their cues from contemporary analytic philosophy: they have co-opted the terms, concepts, and distinctions devised by academic philosophers, and have adapted these into a host of scales and paradigms designed to capture ordinary thought, as though these philosophical concepts could be converted wholesale into psychological constructs. This whole enterprise rests on the dubious notion that the categories and distinctions that characterize contemporary analytic philosophy are merely the domesticated forms of philosophical views that could be found in the "wild" in the words, thoughts, and deeds of ordinary people. In short, researchers have simply presumed that their accounts map onto psychological constructs that are present in ordinary thought with virtually no theoretical justification, and no substantive empirical evidence beyond appeals to anecdotes and personal experience.

I have argued for an alternative possibility: that this simply isn't the case. Perhaps philosophical notions of realism and antirealism aren't refined accounts of what ordinary people are saying and doing when they reason that murder is wrong. Instead, these philosophical positions are academic

inventions that don't reflect what people are actually saying and doing at all when they engage in moral judgment, reasoning, and behavior.

The goal of this dissertation has been to argue that this hypothesis, *folk metaethical indeterminacy*, offers a plausible explanation of existing patterns of data in folk metaethics research. In chapters two and three (and the accompanying supplements), I argued that existing methods are not valid, and have thus not provided us with good evidence of metaethical pluralism or uniformism. In chapters four and five, I presented qualitative and quantitative evidence that ordinary people do not interpret questions about metaethics as intended. At the end of chapter five, I showed that even when we have good reason to believe that a given set of concepts do not plausibly figure into ordinary thought and language, ordinary people can still be induced to respond in a nonrandom way. Were researchers so inclined, studies like this could be used as evidence of "folk quantum mechanics." Yet I have provided two better explanations for these results. In some cases, these findings may be due to *spontaneous theorizing*, i.e., participants developing a philosophical stance in the course of the study. In other situations, results may fail to reflect what participants actually think, since they cannot express their views due studies employing forced choice designs. Either way, we have no good reason to think ordinary people *outside* of the study are "implicit Copenhagians" or "implicit Many Worlders." The observed patterns of data are better understood as empirical artifacts.

I propose that the same is true of folk metaethics. First, we have the unusual pattern of data obtained when such research is conducted. Our best explanation for these results is not that ordinary people are metaethical pluralists, or that they are uniformists making systematic errors. The best explanation is that most participants are not interpreting stimuli in the way researchers intend.

Our next question is *why* people are not interpreting questions as intended. Some of this is undoubtedly due to researcher error, but it's not plausible that it's *entirely* due to inadequate study design. There is no reason to suppose researchers studying folk metaethics are uniquely bad at

designing studies. On the contrary, researchers studying folk metaethics have been highly receptive to the methodological shortcomings of earlier research, and have done far more than is typical in other areas of research to minimize conflations and enhance the proportion of intended interpretations. This has proven so difficult that researchers have resorted to giving participants extensive instructions, training exercises, and extremely complicated response options. This may appear to be a solution to the problem, but it isn't. It simply trades one problem for another. It would appear that the only way to ensure that ordinary people interpret questions about metaethics is to give them an impromptu course in philosophy. Either such efforts fail, in which case we'd need to employ even more robust training protocols, or they succeed, in which case participants are no longer ordinary people, but junior philosophers.

This reveals an insurmountable paradox to folk metaethics research: our goal is to study how people without philosophical training think about a philosophical topic. It turns out that they don't seem to interpret what we're asking them as intended, and the only way we may be able to ensure that they do interpret what we're asking as intended is to train them to do philosophy, exactly as Kauppinen (2007) proposed. Yet once we do this, they are no longer people without philosophical training. This is why I drew on the metaphor of Schrödinger's cat. Once we open the box, we'll find a cat that is either dead or alive, but we won't be able to say that the cat was dead or alive *all along*. Prior to opening the box, there was no determinate fact of the matter about whether the cat was alive or dead. I propose that the same is true for folk metaethics: prior to participating in a study, there is no fact of the matter about whether ordinary people are moral realists or antirealists.

There is little reason to think such indeterminacy is unique to folk metaethics. Instead, I suspect indeterminacy may apply to many classical philosophical concepts and distinctions. Prior to engaging with philosophy, people may have an underlying disposition to think one way or another, but in many cases they have no position on the matter, nor do they speak or think in ways that commit

them to a particular philosophical position. With respect to philosophical stances, they exist in a state of philosophical superposition that resolves into a determinate position only by becoming sufficiently familiar with the philosophical issue in question to form a position on it. Likewise, with respect to philosophical commitments, people's linguistic practices need not conform to any particular philosophical position unless, and only unless, such a commitment actually played a role in regulating their linguistic practices. And we have no justification for presuming that it must play such a role in the absence of any evidence or an adequate theoretical rationale.

Maybe there are good reasons to believe people are implicitly committed to realism or antirealism, or hold substantive stances towards these positions. If so, it will be difficult to account for the existing research on folk metaethics. If people really have positions on these topics, it's unclear why it's so difficult to figure out what they are. And if they have implicit commitments, a century of philosophy and nearly two decades of empirical research have yet to reveal anything about them. Perhaps future studies will succeed where existing ones have failed, but metaethical indeterminacy offers a simple and straightforward alternative explanation: all of this research mistakenly assumed that there was something to measure in the first place. Why should we have ever supposed that ordinary people are moral realists or antirealists, any more than we imagine that they endorse or are implicitly committed to a particular decision theory, e.g., causal decision theory (CDT) or evidential decision theory (EDT) (McCarthy, 2016; Munier, 1988)? There may be no reason to think that the philosophical distinctions philosophers care about capture *any* stable features of ordinary thought and language. The same applies to folk philosophy more broadly. I expect future research in folk philosophy to reveal that researchers grossly underestimated the chasm between ordinary moral thought and philosophical thought, and further, that they incorrectly presume the former mirrors (even if in some nascent way) the latter.

Even if people did have determinate philosophical stances, such research may still be methodologically misguided. I find it profoundly strange that researchers seem not to appreciate just how tall an order it would be to imagine that we could readily solicit responses to complicated philosophical questions using extremely sparse stimuli, stimuli that often include little or no explanation of the relevant concepts, and with almost no efforts made to disambiguate the mountain of potential conflations that would impede people's interpretations of them. After all, one of the central tasks, if not *the* central task of many academic philosophers is to *clarify* the philosophical positions that they are putting forward. Yes, philosophers present arguments. But the bulk of many academic works consists in painstakingly disentangling a concept from a host of potential confusions and misunderstandings, or carefully proposing and distinguishing terminological distinctions with the conceptual equivalent of a surgeon's practiced hand. But after the final satisfying keystroke, one can rest assured that if you say anything interesting at all, some pedant from Neverheardofit University will spring up out of nowhere with an article explaining why you got it all wrong.

And they might have a point. Even professionals struggle to fully grasp the nuances of the topics they write about, and often err in their understanding of concepts they've studied for years. The content of philosophical accounts are highly refined structures. They often bear more resemblance to a Gothic cathedral than the ordinary person's couch. And yet, for some reason, philosophers imagine the equivalent of *Basilica Sancti Francisci Assisiensis* is concealed in everyday thought and speech. Perhaps the reason people struggle so much with questions about metaethics (or questions about philosophy in general) is because there's far less philosophical content implicit in the way people speak and think than researchers imagine. We should consider the sobering possibility that there's no flying buttresses and elaborate frescoes lurking beneath the couch cushions of ordinary thought and language, only stale popcorn and loose change.

This may seem like an extraordinarily bleak conclusion. After all, I'm suggesting that much of the folk philosophical research produced in the past two decades has been fundamentally misguided, delivering results that tell us little or nothing about how ordinary people actually think or speak. It's true that I'm a pessimist about much of this research, and that I'm highly critical of the current state of philosophy and psychology, but my pessimism isn't born out of bitter naysaying and a desire to tear down the establishment. Folk philosophical research started off on the wrong foot. But there is ample time to chart a new course. Such course-correction may be a little humbling and more than a little difficult, but I find the prospects of venturing into uncharted seas exhilarating. Far better to lay new foundations than to lay bricks in a crumbling edifice.

What should we expect to find in charting this new course? I don't know, and that's precisely why it's thrilling to consider a fundamental shift in our methods. While it's too soon to say what we'll find, I would like to end by speculating on how we might get there. The central flaw I've identified in contemporary folk philosophical research has been its emphasis on *top-down*, *a prioristic* methods. Naturally, then, I believe the way forward will require a *bottom-up* approach that emphasizes descriptive and observational research. This suggestion is precisely in line with the observations Rozin (2001) made in commenting on what he perceived to be social psychology's premature attempt to present itself as a mature science before proceeding through the necessary stages for any science to mature. As Rozin puts it,

> I believe that social psychology, modeling itself in the mid-20th century primarily on the natural sciences and on sensory psychology, has concentrated on the advancement of a formal, precise, and experimental science. However, unlike the successful work in the natural sciences and sensory psychology, the work in social psychology has not been preceded by an extensive examination and collection of relevant phenomena and the description of universal or contingent invariances. In the more advanced sciences that social psychology would like to emulate, there is much more emphasis on phenomena and "description" than there is in social psychology, and there is less reliance on experiment. Such sciences, particularly the life sciences, also pay less attention to models and hypotheses and more attention to evidence as opposed to proof or "definitive" studies. (p. 3)

I believe the same is true, to an even greater extent, for research on folk philosophy, and folk metaethics in particular. Unfortunately, most contemporary folk philosophical research hasn't ventured far enough from armchair musings. Instead, it largely consists of soliciting the armchair musings of untrained survey respondents, as though this were an adequate substitute for studying actual moral thought and behavior. Such research may play an important role in our understanding of folk morality, but how is an almost exclusive reliance on such data supposed to yield insights into what people are *actually* thinking and doing in real world instances? Relying on such data is the methodological equivalent of ecologists studying nature by analyzing Bob Ross paintings.[116]

If we really want to understand folk morality, we should study moral thought, language, and behavior *as it occurs in the real world*. The grist for the philosophical mill must come from deliberate engagement with morality in the environments in which it occurs. In the case of folk moral philosophy, this should begin with a *bottom-up* approach that relies on gathering vast amounts of descriptive and observational data that exhibits a deep appreciation for cultural, linguistic, and sociodemographic variations in human populations and the recognition that morality is a *social* phenomenon that should be studied in the ecological contexts in which it occurs. Such research should deliberately eschew the imposition of philosophical concepts and categories onto the data in an attempt to categorize all responses in accordance with the categories and distinctions in analytic philosophy. If we begin empirical inquiry with the presumption that folk thought *must* conform to a predetermined set of categories and distinctions, we can overlook genuine alternatives within those categories. At the same time, we also risk failing to observe features of folk thought that have little or nothing to do with these categories. It may turn out that there are many features of the way ordinary people think about (e.g.,

---

[116] Granted, enough observational research and engagement with empirical evidence can facilitate clever hypotheses from the armchair. Galileo's thought experiment disproving Aristotelian conceptions of falling bodies exemplifies this (Gruszczyński, 2020). However, philosophers rarely engage in rigorous, systematic observational research of ordinary moral thought and practice.

the nature of morality) that aren't reducible to whether they endorse realism or antirealism, cognitivism or noncognitivism, and so on. We'll never know if we only design studies that attempt to sort people according to traditional philosophical distinctions. I am not suggesting we never conduct experiments, or engage in theorizing, or attempt to categorize participants in usefully discrete ways. I am simply claiming that such efforts are the task of a more mature science that has a rich dataset to work with, not the anecdotes and armchair speculation of a handful of unrepresentative academics. Here are a few elements that would be critical to the success of this research:

1. *Extensive pretesting to assess interpretation rates for traditionally philosophical topics.*

   Researchers should not presume people understand the terms and concepts used in folk philosophical research. Instead, they should engage in extensive pretesting to assess how people interpret stimuli in order to determine whether people are interpreting stimuli as intended. This will require doing far more than including a single, trivially easy comprehension check, then declaring that participants understood what was asked because most answered this question correctly. It may involve using open response questions, cognitive interviewing, focus groups, and a variety of qualitative and quantitative methods to assess how well people understand questions. Note that evidence of adequate interpretation rates in one population doesn't justify the inference that *other* populations would interpret items as intended. Variables such as age, language, culture, and socioeconomic status could influence how different populations interpret stimuli.

2. *Account for interpretative variation.*

   Researchers should be mindful of the possibility that participants interpret the same stimuli differently from one another, and factor this into their designs. Interpretative variation may contribute an unavoidable degree of noise to studies, and it may be substantial in cases where items are vague or ambiguous.

3. *More interdisciplinary collaboration*

   Psychologists studying folk metaethics have often made serious errors in how they've operationalized metaethical concepts. This could have been avoided if they'd worked more closely with trained philosophers. Philosophers, in turn, often employ poorly designed studies due to an insensitivity to considerations that psychologists are keenly aware of. Researchers from a variety of backgrounds should work together to complement one another's skillsets.

4. *Expand methods beyond traditional surveys*

Researchers rely far too much on traditional surveys. Researchers should instead conduct more field studies and engage in more observational research, drawing on or engaging in ethnographic research by e.g., consulting content from the HRAF, or engaging in corpus analysis (for a recent example of the latter, see Stojanovic & McNally, 2022).

5. *Conduct more descriptive research*

Researchers often attempt to manipulate philosophical positions *as if* they were psychological constructs, without first establishing that they are, and that these attributes function in the way psychological constructs do. This is a mistake. You can't manipulate something that isn't there. While there is nothing wrong with manipulating folk philosophical beliefs and attitudes that *are* present, you have to first demonstrate that they're there before you can manipulate them.

6. *Don't assume one-to-one mapping between philosophical concepts and psychological constructs*

Researchers should not presume that a particular philosophical concept automatically corresponds to an analogous psychological construct, such that there's a one-to-one correspondence between philosophical positions and ordinary psychology. We should not assume, for instance, that since philosophers tend to endorse compatibilism, incompatibilism, or libertarian free will, that these are reflected in ordinary human psychology as three distinct, corresponding psychological profiles. Nor can we presume that because philosophers use a particular term to refer to a particular concept, that belief or disbelief in this concept is represented by a single, continuous variable in the population. To again draw on the concept of free will, it makes no sense to think that *all* ordinary people can be placed on a single spectrum of greater or lesser belief in "free will," since such measures presume that there is a shared concept that people exhibit varying degrees of belief or disbelief in.

7. *Don't assume associations that philosophers endorse are present in folk philosophy*

Philosophers often believe that particular philosophical positions or concepts are closely associated with other positions and concepts, or that philosophical concepts have entailments or implications for other philosophical positions. But they should not assume that such associations are present in ordinary thought. For instance, philosophers may perceive a very close association between academic disputes about free will and moral responsibility, so much so that one could almost define academic notions of free will as those features of agency necessary for certain forms of moral praise and blame. Yet there is little evidence ordinary people psychologically link free will and moral responsibility in the same way as philosophers.

8. *Avoid forced choice paradigms*

Researchers should avoid forced choice paradigms that require participants to rigidly fit into a narrow subset of categories. Such studies can force people to respond to questions in ways that don't represent what they think, or exacerbate spontaneous theorizing.

9. *Develop the concept of spontaneous theorizing and test for it*

The notion of spontaneous theorizing proposed here emerged over the course of my research. There may be similar notions already present in the literature that I'm not aware of. If so, perhaps they are highly developed. But if they are absent or underdeveloped, greater effort should be made to expand on the concept of spontaneous theorizing, and to devise ways of testing for the role that it plays in any given study. It may be that a much wider range of folk philosophical research is capturing conclusions people reached by participating in the study, rather than stances or commitments they held prior to doing so. If so, this could represent a significant threat to a great deal of research.

10. *Conduct more cross-cultural research*

One of the more obvious shortcomings with existing research is the lack of cross-cultural research. Growing evidence suggests WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations are one of the least psychologically representative populations in the world. Findings that would generalize to humanity as a whole already require more diverse sampling on these grounds alone. However, the need for more culturally diverse samples is especially important for folk philosophy, since most studies draw on categories and distinctions derived from an extremely narrow subset of WEIRD populations.

11. *Study broader array of philosophical traditions, and don't ignore less popular philosophical views*

Folk philosophical research often opts to simplify studies by exclusively assessing folk stances and commitments towards the most popular or simplified versions of a philosophical position. This is why, for instance, researchers have focused almost exclusively on realism and antirealism, or in some cases realism and cultural relativism or individual subjectivism. It took several years before researchers began to include noncognitivism. Yet even this doesn't cover the variety of views that are actually out there. Noncognitivism has been exclusively portrayed as a crude form of emotivism, when contemporary expressivist theories are richer and more sophisticated. And so far, no studies have incorporated measures of constructivism. There is little justification for the exclusion of constructivism. The 2020 PhilPapers survey showed that 20.8% of philosophers endorse or lean towards constructivism, which was greater than expressivism (10.6%) and error theory (5.3%) (Bourget & Chalmers, ms). Why, then, do the

most recent studies include response options for expressivism and error theory, but not constructivism?

12. *Draw on psychological theories and approaches to inform hypotheses*

Lastly, researchers can and should draw on the wealth of psychological approaches, theories, and phenomena to inform folk philosophical hypotheses, rather than on philosophical tradition alone.

I want to end on this last suggestion by pointing to one example of how this could be done. One of the most obvious shortcomings to my argument for metaethical indeterminacy is that it consists solely of a negative thesis. I'm claiming that ordinary people have no determinate stance towards realism and antirealism. This entails that when ordinary people engage in everyday moral judgment, reasoning, and behavior, they're *not* making claims that commit them to realism, and it further entails that views about realism and antirealism don't typically figure into their judgments and reasoning at all. Yet this doesn't tell us what people *are* doing.

The traditional philosophical view presumes when people make moral claims and engage in moral judgment, that they are implicitly appealing to a principled and logically consistent set of moral standards, or are at least *attempting* to do so. As a result, the primary purpose of moral judgment is to assert what's *true*, and the primary purpose of moral reasoning is to enhance moral knowledge by identifying and resolving inconsistencies and errors. While this may be the purpose of moral *philosophy*, researchers seem to imagine that ordinary moral psychology functions in the same way, and that to the extent that ordinary moral thought is biased or inconsistent, this is an unfortunate byproduct of faulty and unrefined thinking.

There is good reason to question these assumptions. Perhaps the apparent biases and inconsistencies in ordinary moral thought are a feature, and not a bug. Moral reasoning is a specific instance of our more general capacity for reasoning. And while researchers have traditionally supposed that the purpose of reasoning is to enable us to have more accurate beliefs, this may not be true. As

Sperber and Mercier (2011; 2017) point out, reasoning is subject to a variety of biases and distortions that seem inconsistent with a psychological system optimized for delivering true beliefs. They propose that reason functions primarily to enable us to persuade others, and is thus best suited to constructing and evaluating arguments. They marshal an impressive body of reasons and evidence for taking this proposal seriously, which they summarize as follows:

> Reasoning so conceived is adaptive given the exceptional dependence of humans on communication and their vulnerability to misinformation. A wide range of evidence in the psychology of reasoning and decision making can be reinterpreted and better explained in the light of this hypothesis. Poor performance in standard reasoning tasks is explained by the lack of argumentative context. When the same problems are placed in a proper argumentative setting, people turn out to be skilled arguers. Skilled arguers, however, are not after the truth but after arguments supporting their views. This explains the notorious confirmation bias. This bias is apparent not only when people are actually arguing, but also when they are reasoning proactively from the perspective of having to defend their opinions. (2011, p. 57)

I suspect that incorporating this account of reasoning into the way we analyze ordinary moral thought could go a long way towards better understanding what people are actually doing in everyday moral discourse. Philosophers have observed that people use moral language in a variety of ways. It sometimes appears to reference stance independent moral facts, while at other times it serves to make claims about our standards or the standards of our culture. In still other cases, it serves to express our emotional states or to issue commands. Philosophers look at this mess and imagine that one of these uses is the proper function of moral language, while the rest are aberrant or parasitic on its primary and central function (Gill, 2008). This presumes that ordinary moral language must conform to a particular metaethical thesis, such that the role of moral claims is implicitly driven by a principled commitment to logical consistency and the goal of accurately expressing one's moral beliefs. This is why philosophers have either never entertained pluralism, or objected to it when it's been proposed (Johansson & Olson, 2015). Indeed, Colebrook (2021) argues that if pluralism is true, this would entail that folk metaethical judgments are *irrational.*

The argumentative theory provides an alternative explanation. To the extent that ordinary moral language appears in some contexts to involve reference to stance-independent facts, in others to appeal to the standards of people or groups, and in others to serve a primarily expressive function, these are not genuine reflections of an underlying commitment to an inconsistent or irrational set of metaethical standards. Instead, such language is superficial, and functions primarily to serve *argumentative* and *rhetorical* purposes. In everyday contexts, moral judgments have consequences. Our moral attitudes towards others can lead to praise and blame, loss of friends, increases or decreases in our status, and all manner of practically relevant consequences. In many cases, the outcome of a moral judgment is a literal matter of life and death. Why, then, should we imagine that people are primarily driven to say what's *true*? Why not, instead, imagine that the primary function of moral judgment, reasoning, and discourse is to facilitate our sociofunctional goals? Humans are a social species. Our reproductive fitness is typically enhanced not by doing philosophy well, but by saying and doing things that will allow us to enhance our status and prestige, to form fruitful relationships with others, to insulate ourselves from threats to our health and success, and to avoid punishment, blame, and loss of reputation. *This* is the context in which moral thought and discourse occurs. Why assume that none of this is relevant to its function?

Once we reorient our perspective on the role of moral thought and language in everyday life, we can see why the apparent inconsistency of moral language is nothing of the sort. While ordinary moral language may not consistently fit with traditional philosophical accounts, this is because such accounts appeal to *semantic* consistency. Yet the argumentative theory centers consistency in the *functional role* of moral language: if using realist-sounding language would serve one's argumentative goal in one context, while using antirealist-sounding language would serve those same goals in a different context, people would be using such language consistently *with respect to the goal of persuading others*. If so, ordinary moral language may appear to be used in ways that fit with one or another

metaethical analysis, not because such language reflects a genuine commitment to a bizarre, highly variable form of moral pluralism, but because people vary how they talk about morality to serve different argumentative ends in different contexts. If so, apparent commitments to metaethical positions aren't so much *built into* the semantics of moral language, so much as they are external semantic drifters capable of fusing with moral language in particular conversational contexts to serve various social goals, without being incorporated into the deeper commitments that characterize moral language *itself*.

This suggestion is little more than a speculative hypothesis that may or may not be vindicated by future research. I don't bring up the argumentative theory of reasoning to defend its application to folk metaethics. I present it as an example of the kinds of empirically grounded paradigm shifts that can fundamentally alter the course of folk philosophical research. My criticisms of folk metaethics may give the impression that I believe there is little work to be done. This example also serves to highlight that nothing could be further from the truth. My critique of folk metaethics does not show that studying how ordinary people think about philosophical topics is a waste of time. On the contrary, it highlights how ordinary thought is deeply embedded in the practical and purpose-driven nature of everyday life. Philosophy was once seen as a method for reflecting, not just on abstract and esoteric matters with little practical relevance, but as a tool for assessing how to live well and pursue a life of purpose, meaning, and value. Far from the most insular and irrelevant of disciplines, it served as the foundation not only for science and academia, but for the ordinary person. If I've demonstrated anything, it's that philosophy has strayed far from this path. While I'm advocating substantial reform and innovation, I am, in a somewhat paradoxical way, also suggesting that philosophy return to its roots.

# REFERENCES

Abshire, J. (2011). *The history of Singapore*. Santa Barbara, CA: ABC-CLIO.

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Is it time for studying real-life debiasing?. Evaluation of the effectiveness of an analogical intervention technique. *Frontiers in Psychology*, *6*, 1120. https://doi.org/10.3389/fpsyg.2015.01120Adams, F., & Steadman, A. (2004). Intentional action and moral considerations: Still pragmatic. *Analysis*, *64*(3), 268-276.

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2012). Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology*, *121*(2), 484-497.

Aikins, H. A. (1902). *The principles of logic*. New York: Henry Holt and Company.

Alexander, J. (2012). *Experimental philosophy: An introduction*. Cambridge, UK: Polity Press.

Alexander, J., Mallon, R., & Weinberg, J. M. (2010). Accentuate the negative. *Review of Philosophy and Psychology*, *1*(2), 297-314.

Alexander, S. (2009, April 28). Generalizing from one example. *LessWrong*. https://www.lesswrong.com/posts/baTWMegR42PAsH9qJ/generalizing-from-one-example

Alexander, S. (2013, April 12). Noisy poll results and reptilian muslim climatologists from Mars. *LessWrong*. https://www.lesswrong.com/posts/CsKrQdQJJCFPjfKjF/noisy-poll-results-and-reptilian-muslim-climatologists-from-mars/

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: what counts?. *Philosophical Psychology, 26*(5), 673-701.

Allen, K., Quinlan, P., Andow, J., & Fischer, E. (2021). What is it like to be colour-blind? A case study in experimental philosophy of experience. *Mind & Language*, 1-26. https://doi.org/10.1111/mila.12370

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment, 38*(1), 1-5.

Allidina, S., & Cunningham, W. A. (2018). Moral cues from ordinary behaviour. *Behavioral and Brain Sciences, 43*. https://doi.org/10.1017/S0140525X18000018

Altman, D. G., & Bland, J. M. (1996). Absence of evidence is not evidence of absence. *British veterinary journal, 74*(4), 311-311.

Amira, K., Cooper, C. A., Knotts, H. G., & Wofford, C. (2018). The southern accent as a heuristic in American campaigns and elections. *American Politics Research*, *46*(6), 1065-1093.

Andow, J. (2016). Qualitative tools and experimental philosophy. *Philosophical Psychology*, *29*(8), 1128-1141.

Apter, M. J. (1982). The experience of motivation: The theory of psychological reversals. London, UK: Academic Press.

Ardila-Rey, A., & Killen, M. (2001). Middle class Colombian children's evaluations of personal, moral, and social–conventional interactions in the classroom. *International Journal of Behavioral Development, 25*(3), 246-255.

Arnett, J. J. (2009). The neglected 95%, a challenge to psychology's philosophy of science. *American Psychologist, 64*(6), 571-574.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PloS One*, *9*(9), e106361. https://doi.org/10.1371/journal.pone.0106361

Asch, S. E. (1987). *Social psychology*. New York, NY: Oxford University Press. (Original work published 1952)

Ash, K., Johnson, W., Lagadinos, G., Simon, S., Thomas, J., Wright, E., & Gainous, J. (2020). Southern accents and partisan stereotypes: Evaluating political candidates. *Social Science Quarterly*, *101*(5), 1951-1968.

Ayars, A., & Nichols, S. (2020). Rational learners and metaethics: Universalism, relativism, and evidence from consensus. *Mind & Language*, *35*(1), 67-89.

Babushkina, D. (2019). Bradley's "my station and its duties" and its moral (in) significance. *Zeitschrift für Ethik und Moralphilosophie*, *2*(2), 195-211.

Bagnoli, C. (2021). Constructivism in metaethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). https://plato.stanford.edu/archives/spr2021/entries/constructivism-metaethics

Baguley, T. (2012, June 21). *The stimuli-as-fixed-effect fallacy*. R-bloggers. https://www.r-bloggers.com/2012/06/the-stimuli-as-a-fixed-effect-fallacy/

Ballester, J., Abdi, H., Langlois, J., Peyron, D., & Valentin, D. (2009). The odor of colors: Can wine experts and novices distinguish the odors of white, red, and rosé wines?. *Chemosensory Perception*, *2*(4), 203-213.

Barnard, R., & Ulatowski, J. (2013). Truth, correspondence, and gender. *Review of Philosophy and Psychology*, *4*(4), 621-638.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536-554.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323-370.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260-268.

Baz, A. (2012). *When words are called for.* Cambridge, MA: Harvard University Press.

Baz, A. (2015). Questioning the method of cases fundamentally—Reply to Deutsch. *Inquiry*, *58*(7-8), 895-907.

Baz, A. (2017). *The crisis of method in contemporary analytic philosophy.* Oxford, UK: Oxford University Press.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Becklloyd, D. (2021). *Is common sense neither compatibilist nor incompatibilist?* [Master's thesis, Open Access Te Herenga Waka-Victoria University of Wellington]. https://doi.org/10.26686/wgtn.14927007.v1

Beebe, J. R. (2012). Experimental epistemology. In A. Cullison (Ed.), *Continuum companion to epistemology* (pp. 248-269). London, UK: Continuum.

Beebe, J. R. (2014). How different kinds of disagreement impact folk metaethical judgments. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 167-187). London, UK: Bloomsbury.

Beebe, J. R. (2015). The empirical study of folk metaethics. *Etyka*, *50*, 11-28.

Beebe, J. R. (2020). The empirical case for folk indexical moral relativism. In T Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 4, pp. 81-111). Oxford, UK: Oxford University Press.

Beebe, J. R., & Sackris, D. (2016). Moral objectivism across the lifespan. *Philosophical Psychology*, *29*(6), 912-929.

Beebe, J. R., Qiaoan, R., Wysocki, T., & Endara, M. A. (2015). Moral objectivism in cross-cultural perspective. *Journal of Cognition and Culture*, *15*(3-4), 386-401.

Bergenholtz, C., Busch, J., & Praëm, S. K. (2021). Exclusion Criteria in Experimental Philosophy. *Erkenntnis*, *86*(6), 1531-1545.

Berniūnas, R. (2014). Delineating the moral domain in moral psychology. *Problemos*, (Supplement), 90-101.

Berniūnas, R. (2020). Mongolian yos surtakhuun and WEIRD "morality". *Journal of Cultural Cognitive Science*, *4*(1), 59-71.

Berniūnas, R., Beinorius, A., Dranseika, V., Silius, V., & Rimkevičius, P. (2021). The weirdness of belief in free will. *Consciousness and Cognition*, *87*, 103054. https://doi.org/10.1016/j.concog.2020.103054

Berniūnas, R., Silius, V., & Dranseika, V. (2022). Moralization East and West: Moralizing different transgressions among Chinese, Americans, and Lithuanians. *Asian Journal of Social Psychology*, *25*(2), 185-197.

Bervoets, J., Milton, D., & Van de Cruys, S. (2021). Autism and intolerance of uncertainty: an ill-fitting pair. *Trends in Cognitive Sciences*, *25*(12), 1009-1010.

Bex-Priestley, G. (2018). Error and the limits of quasi-realism. *Ethical Theory and Moral Practice*, *21*(5), 1051-1063.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382-386.

Birch, S. A., Brosseau-Liard, P. E., Haddock, T., & Ghrear, S. E. (2017). A 'curse of knowledge' in the absence of knowledge? People misattribute fluency when judging how common knowledge is among their peers. *Cognition*, *166*, 447-458.

Björnsson, G. (2012). Do 'objectivist' features of moral discourse and thinking support moral objectivism?. *The Journal of ethics*, *16*(4), 367-393.

Blackburn, S. (1991). Just causes. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *61*(1/2), 3-17.Blackburn, S. (1993). Essays in quasi-realism. Oxford, UK: Oxford University Press.

Blackburn, S. (2000). Ruling passions: A theory of practical reasoning. Oxford, UK: Oxford University Press.

Blair, J. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*(1), 1-29.

Blair, J., Jones, L., Clark, F., & Smith, M. (1995). Is the psychopath 'Morally Insane'? *Personality and Individual Differences, 19*(5), 741-752.

Blair, R. J. R. (1997). Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences, 22*(5), 731-739.

Bojanowski, J. (2012). Is Kant a moral realist?. *Kant Yearbook*, *4*(1), 1-22.

Bonard, C., Cova, F., & Humbert-Droz, S. (2021). De gustibus est disputandum: An empirical investigation of the folk concept of aesthetic taste. In *Perspectives on Taste* (pp. 77-108). Routledge.

Bonge, D. R., Schuldt, W. J., & Harper, Y. Y. (1992). The experimenter-as-fixed-effect fallacy. *The Journal of Psychology*, *126*(5), 477-486.

Booth, A, R., & Rowbottom, D. P. (Eds.) (2014). Intuitions. Oxford, UK: Oxford University Press.

Bourget, D. & Chalmers, D. J. (ms) Philosophers on Philosophy: The PhilPapers 2020 Survey.

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe?. *Philosophical Studies, 170*(3), 465-500.

Bower, B. (2005, December 4). The Pirahã Challenge. *ScienceNews* https://www.sciencenews.org/article/pirah%C3%A3-challenge (accessed June 10, 2022)

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Braun, V., & Clarke, V. (2014). What can "thematic analysis" offer health and wellbeing researchers?. *International Journal of Qualitative Studies on Health and Well-being*, *9*(1), 26152. https://doi.org/10.3402/qhw.v9.26152

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589-597.

Braun, V., & Clarke, V. (2021). Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, *21*(1), 37-47.

Brogaard, B. (2008). Moral contextualism and moral relativism. *The Philosophical Quarterly*, *58*(232), 385-409.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). New York, NY: Guilford Press.

Buchtel, E. E., Guan, Y., Peng, Q., Su, Y., Sang, B., Chen, S. X., & Bond, M. H. (2015). Immorality east and west: Are immoral behaviors especially harmful, or especially uncivilized?. *Personality and Social Psychology Bulletin*, *41*(10), 1382-1394.

Buckley, C. B. (1902). *An anecdotal history of Singapore (vol. 1)*. Singapore: Fraser & Neave, Limited.

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS One*, *10*(8), e0136589. https://doi.org/10.1371/journal.pone.0136589

Bush, L. S., & Moss, D. (2020). Misunderstanding metaethics: Difficulties measuring folk objectivism and relativism. *Diametros 17*(64): 6-21.

Bush, L.S. (2016). Cultural & moral relativism. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-16999-6_518-1

Bush, L.S., & Millhouse, T. (2016, June). *Operationalizing metaethics: The disagreement paradigm in empirical moral psychology* [conference session]. The 42nd Annual Meeting of the Society for Philosophy and Psychology, Austin, TX. https://www.socphilpsych.org/spp2016program.pdf

Çağaptay, S. (2014). *The rise of Turkey: The twenty-first century's first Muslim power*. Dulles, VA: Potomac Books.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, *9*(3), 240-244.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232-1254.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford, UK: Oxford University Press.

Cappelen, H. (2014). X-phi without intuitions?. In A. Booth and D. Rowbottom (Eds.). *Intuitions* (pp. 269-286). New York, NY: Oxford University Press.

Carnes, N. C., Allmon, B., Alva, J., Cousar, K. A., & Varnam, Z. D. (2022). How morality signals, benefits, binds, and teaches. *Journal of Experimental Social Psychology*, *101*, 104313. https://doi.org/10.1016/j.jesp.2022.104313

Carroll, L. (2022). *Jabberwocky*. Poetry Foundation. https://www.poetryfoundation.org/poems/42916/jabberwocky (Original work published in 1871)

Carter, D. B., & Patterson, C. J. (1982). Sex roles as social conventions: The development of children's conceptions of sex-role stereotypes. *Developmental Psychology*, *18*(6), 812-824.

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology nursing forum*, *41*(5), 545-547.

Caspar, E. A., Verdin, O., Rigoni, D., Cleeremans, A., & Klein, O. (2017). What do you believe in? French translation of the FAD-plus to assess beliefs in free will and determinism and their relationship with religious practices and personality traits. *Psychologica Belgica, 57*(1), 1-16. https://doi. org/10.5334/pb.321

Caspar, E. A., Vuillaume, L., Magalhães De Saldanha da Gama, P. A., & Cleeremans, A. (2017). The influence of (dis)belief in free will on immoral behavior. *Frontiers in Psychology, 8,* Article 20. https://doi.org/10.3389/fpsyg.2017.00020

Cassady, J. C., & Finch, W. H. (2014). Confirming the factor structure of the Cognitive Test Anxiety Scale: Comparing the utility of three solutions. *Educational Assessment*, *19*(3), 229-242.

Castensen, M. W., Sellmaier, S., Taylor, P. C. J., Deroy, O. (2022). Free will properly understood: Folk conceptions are incompatibilist, but not hard incompatibilist. *SSRN*. http://dx.doi.org/10.2139/ssrn.4024722

Caven, P. (2015). *Moral Disagreement-A Psychological Account and the Political Implications*. (uk.bl.ethos.658757) [Doctoral dissertation, University of Sheffield]. White Rose eTheses Online.

Caverni, J. P., Fabre, J. M., Gonzalez, M. (1990). *Cognitive biases*. Amsterdam, NL: Elsevier.

Chalmers, D. J. (2014). Intuitions in philosophy: A minimal defense. *Philosophical Studies*, *171*(3), 535-544.

Chan, H. M., & Yan, H. K. (2007). Is there a geography of thought for east-west differences? Why or why not?. *Educational Philosophy and Theory*, *39*(4), 383-403.

Chance, Z., & Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, *6*, 104-107.

Chappell, R. Y. (2007, June 19). *Context and relativism*. Philosophy, et cetera. https://www.philosophyetc.net/2007/06/context-and-relativism.html

Chartrand, L. (2022). Modeling and corpus methods in experimental philosophy. *Philosophy Compass*, e12837. https://doi.org/10.1111/phc3.12837

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral "ought" judgments for impossible actions: Toward an empirical refutation of "ought" implies "can". *Cognition*, *150*, 20-25.

Choi, I., Koo, M., & Choi, J. A. (2007). Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin*, *33*(5), 691-705.

Choshen-Hillel, S., Shaw, A., & Caruso, E. M. (2015). Waste management: How reducing partiality can promote efficient resource allocation. *Journal of Personality and Social Psychology*, *109*(2), 210-231.

Chudnoff, E. (2013). *Intuition*. Oxford, UK: Oxford University Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335-359.

Clarke, V., & Braun, V. (2013). Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, *26*(2), 120--123.

Climenhaga, N. (2018). Intuitions are used as evidence in philosophy. *Mind*, *127*(505), 69-104.

Cline, B. (2018). The tale of a moderate normative skeptic. *Philosophical Studies*, *175*(1), 141-161.

Cohen, Y. (2018). An analysis of recent empirical data on 'ought' implies 'can'. *Philosophia*, *46*(1), 57-67.

Cohnitz, D., & Häggqvist, S. (2010). The role of intuitions in philosophy. *Studia Philosophica Estonica*, *2.2*, 1-14. https://doi.org/10.12697/spe.2009.2.2.01

Colebrook, R. (2021). The irrationality of folk metaethics. *Philosophical Psychology*, *34*(5), 684-720.

Colebrook, R., & Sarkissian, H. (2018). Objectivity. In T. Shackelford & V. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science*. Springer International Publishing.

Collier-Spruel, L., Hawkins, A., Jayawickreme, E., Fleeson, W., & Furr, R. M. (2019). Relativism or tolerance? Defining, assessing, connecting, and distinguishing two moral personality features with prominent roles in modern societies. *Journal of Personality*, *87*(6), 1170-1188.

Compen, B., Pitthan, F., Schelfhout, W., & De Witte, K. (2022). How to elicit and cease herding behaviour? On the effectiveness of a warning message as a debiasing decision support system. *Decision Support Systems*, *152*, 113652. https://doi.org/10.1016/j.dss.2021.113652

Cooper, D. E. (1978). Moral relativism. *Midwest Studies in Philosophy*, *3*(1), 97-108.

Cosmides, L., & Tooby, J. (2016). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 91-115). New York, NY: Guilford.

Countering Moral Relativism (2015, April 21). *California Catholic conference.* https://cacatholic.org/article/countering-moral-relativism-spiritual-poverty-our-time

Cova, F., & Pain, N. (2012). Can folk aesthetics ground aesthetic realism?. *The Monist, 95*(2), 241-263.

Cova, F., Garcia, A., & Liao, S. Y. (2015). Experimental philosophy of aesthetics. *Philosophy Compass, 10*(12), 927-939.

Cova, F., Olivola, C. Y., Machery, E., Stich, S., Rose, D., Alai, M., ... & Zhu, J. (2019). De Pulchritudine non est Disputandum? A cross-cultural investigation of the alleged intersubjective validity of aesthetic judgment. *Mind & Language*, *34*(3), 317-338.

Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., & Lambert, N. M. (2016). Subjective correlates and consequences of belief in free will. *Philosophical Psychology*, *29*(1), 41-63.

Crone, D. L., & Levy, N. L. (2019). Are free will believers nicer people? (Four studies suggest not). *Social Psychological and Personality Science, 10*(5), 612-619.

Curry, O. S., Alfano, M., Brandt, M. J., & Pelican, C. (2021). Moral molecules: Morality as a combinatorial system. *Review of Philosophy and Psychology*, 1-20. https://doi.org/10.1007/s13164-021-00540-x

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106-124.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47-69.

da Vela, B., & Earley, B. (2015). Senatus and Sénat: the reception of the Roman Senate during the Radical Stage of the French Revolution (1792-4). *Classical Receptions Journal*, *7*(1), 46-63.

Damon, W. (1977). *The social world of the child*. San Francisco, CA: Jossey-Bass.

Dancy, J. (2004). *Ethics without principles*. Oxford, UK: Clarendon Press.

Dancy, J. (2017) Moral particularism. In E. N Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017 ed.). https://plato.stanford.edu/archives/win2017/entries/moral-particularism/

Darwall, S. (2006). Morality and practical reason: A Kantian approach. In D. Copp, (Ed.), *The Oxford handbook of ethical theory* (pp. 282-320). New York, NY: Oxford University Press.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association, 47*, 5-20.

Davis, T. (2021). Beyond objectivism: new methods for studying metaethical intuitions. *Philosophical Psychology*, *34*(1), 125-153.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science*, *41*, 382-402.

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA. *International Review of Social Psychology*, *32*(1), 1-12.

de Muynck, W. M. (2004) *Foundations of quantum mechanics, an empiricist approach*. Dordrecht, NL: Kluwer Academic Publishers.

Denman, D. C., Baldwin, A. S., Betts, A. C., McQueen, A., & Tiro, J. A. (2018). Reducing "I don't know" responses and missing survey data: Implications for measurement. *Medical Decision Making*, *38*(6), 673-682.

Dennett, D. C. (1984). Elbow room: The varieties of free will worth wanting. Cambridge, MA: MIT Press.

Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind & Language, 24*(4), 445-466.

Deutsch, M. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology, 1*(3), 447-460.

Deutsch, M. (2015). *The myth of the intuitive*. Cambridge, MA: MIT Press.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). Theory of mind: did evolution fool us?. *PloS One*, *9*(2), e87619. https://doi.org/10.1371/journal.pone.0087619

Dhaliwal, N. A., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes*, *164*, 27-51.

Díaz, R. (2019). Using fMRI in Experimental Philosophy: Exploring the Prospects. In E. Fischer & M. Curtis (Eds.) *Methodological advances in experimental philosophy* (pp. 131-152). London, UK: Bloomsbury Publishing.

Dink, J. W., & Rips, L. J. (2017). Folk teleology and its implications. In D. Rose (Ed.), *Experimental metaphysics* (pp. 207-235). London, UK: Bloomsbury Academic.

Dionysius of Halicarnassus. (1940). *Roman Antiquities, Volume III: Books 5-6.48* (E. Cary, Trans.). Cambridge, MA: Harvard University Press.

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of Learning and Motivation*, *50*, 307-338.

Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, *151*, 109547. https://doi.org/10.1016/j.paid.2019.109547

Doing reflexive TA. (n.d.) *Thematic analysis*. https://www.thematicanalysis.net/doing-reflexive-ta/

Dominici, L. (2018, July 24). *Moral relativism and its effects*. Seven Pillars Institute for Global Finance and Ethics. https://sevenpillarsinstitute.org/moral-relativism-and-its-effects/

Doral, M. (2010). Turkey: Another $1 Trillion Emerging Economy?. *Journal of Emerging Knowledge on Emerging Markets*, *2*(November 2010), 71-83.

Dorsey, D. (2006). A coherence theory of truth in ethics. *Philosophical Studies, 127*(3), 493-523.

Dranseika, V., Berniūnas, R., & Silius, V. (2018). Immorality and bu daode, unculturedness and bu wenming. *Journal of Cultural Cognitive Science, 2*(1), 71-84.

Dreier, J. (2004). Meta-ethics and the problem of creeping minimalism. *Philosophical Perspectives*, *18*, 23-44.

Drennan, J. (2003). Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing, 42*(1), 57-63.

Dupré, J. (2016). Against maladaptationism: or what's wrong with evolutionary psychology?. In M. Mazzotti (Ed.), *Knowledge as social order: Rethinking the sociology of Barry Barnes* (pp. 165-180). New York, NY: Routledge.

Dweck, C. (2006). Mindset: The new psychology of success. Philadelphia, PA: Psychology Press.

Dwyer, S. (1999). Moral competence. In K. Murasugi & R. Stainton (Eds.), *Philosophy and linguistics* (pp. 169-190). Boulder, CO: Westview Press.

Dwyer, S. (2004). How good is the linguistic analogy. In Carruthers, P., Laurence, S., & Stich, S. (Eds.), *The innate mind: Culture and cognition* (Vol. 2, pp. 237-256). New York, NY: Oxford University Press.

Dwyer, S., Huebner, B., & Hauser, M. D. (2010). The linguistic analogy: Motivations, results, and speculations. *Topics in Cognitive Science*, *2*(3), 486-510.

Earlenbaugh, J., & Molyneux, B. (2009). Intuitions are inclinations to believe. *Philosophical Studies*, *145*(1), 89-109.

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, *30*(3), 133-140.

Enoch, D. (2014). Why I am an objectivist about ethics (and why you are, too). *The ethical life*, 192-205.

Enoch, D. (2005). Why idealize?. *Ethics*, *115*(4), 759-787.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current anthropology*, *46*(4), 621-646.

Everett, D. (2008). Don't sleep, there are snakes: Life and language in the Amazonian jungle. New York, NY: Pantheon books.

Everett, D. L. (1983). *A língua Pirahã ea teoria da sintaxe: Descrição, perspectivas e teoria* [Doctoral dissertation, Unicamp Campinas].

Everett, D. L. (1986). *Pirahã*. In by D. Derbyshire and G. Pullum (Eds.), *Handbook of Amazonian languages* (Vol. 1, pp. 200-326). Berlin: Mouton de Gruyter.

Everett, D. L. (2009b). Pirahã culture and grammar: A response to some criticisms. *Language*, 405-442.

Everett, D. L. (2012). What does Pirahã grammar have to teach us about human language and the mind?. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(6), 555-563.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200-216.

Evers, D. (2014). Moral contextualism and the problem of triviality. *Ethical Theory and Moral Practice*, *17*(2), 285-297.

FAQs. (n.d.). *Thematic analysis*. https://www.thematicanalysis.net/faqs/

Fazekas, K. (2016). Special Relativity, Multiple B-series, and the Passage of Time. *American Philosophical Quarterly*, 215-229.

Feldman, G., Chandrashekar, S. P., & Wong, K. F. E. (2016). The freedom to excel: Belief in free will predicts better academic performance. *Personality and Individual Differences, 90*, 377-383.

Feldman, G., Farh, J. L., & Wong, K. F. E. (2018). Agency beliefs over time and across cultures: Free will beliefs predict higher job satisfaction. *Personality and Social Psychology Bulletin, 44*(3), 304-317.

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, *18*(1), 342-350.

Feltz, A., & Cokely, E. T. (2013). Predicting philosophical disagreement. *Philosophy Compass*, *8*(10), 978-989.

Feltz, A., cokely, E. T. (2008). *The fragmented folk: More evidence of stable individual differences in moral judgments and folk intuitions* [paper presentation]. In B.C. Love, K. McRae & V.M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 1771-1776). Austin, TX: Cognitive Science Society.

Feltz, A., Cokely, E. T., & Nadelhoffer, T. (2009). Natural compatibilism versus natural incompatibilism: Back to the drawing board. *Mind & Language*, *24*(1), 1-23.

Fessler, D. M., Barrett, H. C., Kanovsky, M., Stich, S., Holbrook, C., Henrich, J., Bolyanatz, A. H., Gervais, M. M., Gurven, M., Kushnick, G., Pisor, A. C., von Rueden, C., Laurence, S. (2015). Moral parochialism and contextual contingency across seven societies. *Proceedings of the Royal Society B*, *282*(1813), 20150907. https://doi.org/10.1098/rspb.2015.0907

Fessler, D. M., Holbrook, C., Kanovsky, M., Barrett, H. C., Bolyanatz, A. H., Gervais, M. M., Gurven, M., Henrich, J., Kushnick, G., Pisor, A. C., Stich, S., von Rueden, C., & Laurence, S. (2016). Moral parochialism misunderstood: a reply to Piazza and Sousa. *Proceedings of the Royal Society B*, *283*(1823), 20152628. https://doi.org/10.1098/rspb.2015.2628

Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science*, *24*, 365-393.

Finlay, S. (2008). The error in the error theory. *Australasian Journal of Philosophy*, *86*(3), 347-369.

Finlay, S. (2011). Errors upon errors: A reply to Joyce. *Australasian Journal of Philosophy*, *89*(3), 535-547.

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, *12*(3), 317-345.

Fischer, E., & Curtis, M. (Eds.). (2019). *Methodological advances in experimental philosophy*. New York, NY: Bloomsbury Academic.

Fischoff, B. (1981). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty* (pp. 331-339). New York, NY: Cambridge University Press.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cognitive Science*, *41*(4), 1119-1134.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Foot, P. (1978a). Moral arguments. In Foot (Ed.), *Virtues and vices and other essays in moral philosophy* (pp. 96-109). Oxford, UK: Basil Blackwell.

Foot, P. (1978b). Moral beliefs. In Foot (Ed.), *Virtues and vices and other essays in moral philosophy* (pp. 110-131). Oxford, UK: Basil Blackwell.

Formosa, P. (2013). Is Kant a moral constructivist or a moral realist?. *European Journal of Philosophy*, *21*(2), 170-196.

Forsyth, D. R. (1980). A taxonomy of ethical ideologies. *Journal of Personality and Social Psychology, 39*(1), 175-184.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, *108*(3), 819-824.

Fuqua, J. (2021). Ethical Mooreanism. *Synthese*, *199*(3), 6943-6965.

Gabennesch, H. (1990). The perception of social conventionality by children and adults. *Child Development*, *61*(6), 2047-2059.

Gaesser, B., Campbell, D., & Young, L. (2022). Moral psychology from the lab to the wild: Relief registries as a paradigm for studying real-world altruism. PsyArXiv. 10.31234/osf.io/2xfsz

Galef, J. (2021). *The scout mindset: Why some people see things clearly and others don't*. Canada: Penguin.

Gendler, T. S., & Liao, S. Y. (2016). The problem of imaginative resistance. In N. Carroll & J. Gibson (Eds.), *The Routledge companion to philosophy of literature* (pp. 405-418). New York, NY: Routledge.

Gibbard, A (2003). Thinking how to live. Cambridge, MA: Harvard University Press.

Gill, M. B. (2008). Metaethical variability, incoherence, and error. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality* (Vol. 2, pp. 387-402). Cambridge, MA: MIT Press.

Gill, M. B. (2009). Indeterminacy and variability in meta-ethics. *Philosophical Studies, 145*(2), 215-234.

Goh, C.M.C (2017). English as lingua franca: Singapore's common tongue. In O.S. Tan & E.L. Low (Eds.), *Lee Kuan Yew's educational legacy: The challenge of success* (pp. 57-67). Singapore: Springer.

Goldbeck-Wood, S. (1999). Evidence on peer review—scientific quality control or smokescreen?. *BMJ*, *318*(7175), 44-45.

Goldman, A. I. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien*, *74*(1), 1-26. https://doi.org/10.1163/9789401204651_002

Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106*(3), 1339-1366.

Goodwin, G. P., & Darley, J. M. (2010). The perceived objectivity of ethical beliefs: Psychological findings and implications for public policy. *Review of Philosophy and Psychology, 1*(2), 161-188.

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others?. *Journal of Experimental Social Psychology*, *48*(1), 250-256.

Google Ngram Viewer (2019a). '" moral realism "'. Google Books Ngram Viewer. https://books.google.com/ngrams/graph?content=%22+moral+realism+%22&year_start=1800&year_end=2019&corpus=26&smoothing=3&case_insensitive=true

Google Ngram Viewer (2019b). '" moral relativism "'. Google Books Ngram Viewer. https://books.google.com/ngrams/graph?content=%22+moral+relativism+%22&year_start=1800&year_end=2019&corpus=26&smoothing=3&case_insensitive=true

Gopnik, A., & Schwitzgebel, E. (1998). Whose concepts are they, anyway? The role of philosophical intuition in empirical psychology. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophy inquiry* (pp. 75-91). Lanham, MD: Rowman & Littlefield.

Gordon, D. S., & Lea, S. E. (2016). Who punishes? The status of the punishers affects the perceived success of, and indirect benefits from,"moralistic" punishment. *Evolutionary Psychology*, *14*(3), 1-14. https://doi.org/10.1177/1474704916658042

Gowans, C. (2021). Moral relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 ed.). https://plato.stanford.edu/archives/spr2021/entries/moral-relativism/

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029-1046.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55-130). San Diego, CA: Academic Press.

Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, *8*, 125-130.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385.

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600-1615.

Greco, J. (2014). Common Sense in Thomas Reid. *Canadian Journal of Philosophy 41*(Sup. 1): 142-155.

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology?. *Nature Reviews Neuroscience*, *4*(10), 846-850.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY: Penguin.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, *124*(4), 695-726.

Greene, J.D. (2008). The secret joke in Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, volume 3: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35-80). Cambridge, MA: MIT Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), S*yntax and semantics III: Speech acts* (pp. 41-58). New York: Academic Press.

Grimm, P. (2010). Social desirability bias. In J. Sheth & N. Malhotra (Eds.), *Wiley international encyclopedia of marketing*. https://doi.org/10.1002/9781444316568.wiem02057

Gruszczyński, R. (2020). Parts of falling objects: Galileo's thought experiment in mereological setting. *Erkenntnis*, 1-22. https://doi.org/10.1007/s10670-020-00263-y

Guerin, B., Leugi, G. B., & Thain, A. (2018). Attempting to overcome problems shared by both qualitative and quantitative methodologies: Two hybrid procedures to encourage diverse research. *The Australian Community Psychologist*, *29*(2), 74-90.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814-834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55-66.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. *Journal of Personality and Social Psychology*, *65*(4), 613-628.

Hawley, P. (2008). Moral absolutism defended. *The Journal of Philosophy*, *105*(5), 273-275.

Hawley, P. H., Stump, K. N., & Ratliff, J. (2010). Sidestepping the jingle fallacy: Bullying, aggression, and the importance of knowing the difference. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in North American schools* (pp. 121-136). New York, NY: Routledge.

Heiphetz, L., & Young, L. L. (2017). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition*, *167*, 78-90.

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2013). The development of reasoning about beliefs: Fact, preference, and ideology. *Journal of Experimental Social Psychology*, *49*(3), 559-565.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science*, *41*(3), 744-767.

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology*, *78*, 210-219.

Helwig, C. C. (2006). Rights, civil liberties, and democracy across cultures. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 185-210). Mahwah, NJ: Lawrence Erlbaum Associates.

Henne, P., Chituc, V., De Brigard, F., & Sinnott-Armstrong, W. (2016). An empirical refutation of 'ought' implies 'can'. *Analysis*, *76*(3), 283-290.

Henne, P., Semler, J., Chituc, V., De Brigard, F., & Sinnott-Armstrong, W. (2019). Against some recent arguments for 'ought' implies 'can': Reasons, deliberation, trying, and furniture. *Philosophia*, *47*(1), 131-139.

Henrich, J. (2022). *Q&A on WEIRD*. The WEIRDist people in the world. https://weirdpeople.fas.harvard.edu/qa-weird

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795-815.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied*, *5*(2), 205-221.

Hofer, H. & Mills, G. (1998). *Hyaenas : status survey and conservation action plan*. International Union for Conservation of Nature (IUCN).
https://policycommons.net/artifacts/1372774/hyaenas/1986984/

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Hofstede, G. H. (n.d.). *The 6-D model of national culture.* https://geerthofstede.com/culture-geerthofstede-gert-jan-hofstede/6d-model-of-national-culture/

Hok, H., Martin, A., Trail, Z., & Shaw, A. (2020). When children treat condemnation as a signal: The costs and benefits of condemnation. *Child Development*, *91*(5), 1439-1455.

Holleman, G. A., Hooge, I. T., Kemner, C., & Hessels, R. S. (2020). The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, *11*, Article 721. https://doi.org/10.3389/fpsyg.2020.00721

Hollos, M., Leis, P. E., & Turiel, E. (1986). Social reasoning in Ijo children and adolescents in Nigerian communities. *Journal of Cross-Cultural Psychology*, *17*(3), 352-374.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Hopster, J. (2019). The meta-ethical significance of experiments about folk moral objectivism. *Philosophical Psychology*, *32*(6), 831-852.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963-976.

Horvath, J., & Grundmann, T. (Eds.). (2012). *Experimental philosophy and its critics*. London, UK: Routledge.

Horvath, J., & Koch, S. (2021). Experimental philosophy and the method of cases. *Philosophy Compass*, *16*(1), e12716. https://doi.org/10.1111/phc3.12716

Horwich, P. (2012). *Wittgenstein's metaphilosophy*. Oxford, UK: Oxford University Press.

Hsieh, N. S. & Andersson H. (2021). Incommensurable values. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.).
https://plato.stanford.edu/archives/fall2021/entries/value-incommensurable/

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114.

Huebner, B., Lee, J. J., & Hauser, M. D. (2010). The moral-conventional distinction in mature moral competence. *Journal of Cognition and Culture, 10*(1-2), 1-26.

Huebner, B., Lee, J. J., & Hauser, M. D. (2010). The moral-conventional distinction in mature moral competence. *Journal of Cognition and Culture, 10*(1-2), 1-26.

Huemer, M. (2007). Compassionate phenomenal conservatism. *Philosophy and Phenomenological Research*, *74*(1), 30-55.

Huff, G., & Huff, G. (2020). The Second World War Japanese Occupation of Singapore. *Journal of Southeast Asian Studies*, *51*(1-2), 243-270.

Hughes, G. D. (2009). The impact of incorrect responses to reverse-coded survey items. *Research in the Schools*, *16*(2), 78-88.

Hunt, S. D., Sparkman Jr, R. D., & Wilcox, J. B. (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research*, *19*(2), 269-273.

Hursthouse, R. (2002). *On virtue ethics*. New York, NY: Oxford University Press.

Hursthouse, R. (2016). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018 ed.). https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166-184.

Iurino, K., & Saucier, G. (2020). Testing measurement invariance of the Moral Foundations Questionnaire across 27 countries. *Assessment*, *27*(2), 365-372.

James, W. (1890). *The principles of psychology*. New York, NY: Henry Holt and Company.

*Jamovi* (Version 2.3.21) [Computer Software]. https://www.jamovi.org

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521-537.

*JASP Team* (Version 0.17). JASP (Version 0.17) https://jasp-stats.org/

Jebari, J., & Huebner, B. (2018). From objectivized morality to objective morality. *Behavioral and Brain Sciences*, *41*. 10.1017/S0140525X18000109

Jian, Z., Zhang, W., Tian, L., Fan, W., & Zhong, Y. (2019). Self-deception reduces cognitive load: The role of involuntary conscious memory impairment. *Frontiers in Psychology*, *10*, Article 1718. https://doi.org/10.3389/fpsyg.2019.01718

Johansson, J., & Olson, J. (2015). Against pluralism in metaethics. In C. Daly (Ed.), *The Palgrave handbook of philosophical methods* (pp. 593-609). Hampshire, UK: Palgrave Macmillan.

Johansson-Stenman, O. (2012). Are most people consequentialists?. *Economics Letters*, *115*(2), 225-228.

Johnson, J. M., Bristow, D. N., & Schneider, K. C. (2004). Did you not understand the question or not? An investigation of negatively worded questions in survey research. *Journal of Applied Business Research (JABR)*, *20*(1), 75-86.

Johnson, S. G., Rodrigues, M., & Tuckett, D. (2021). Moral tribalism and its discontents: How intuitive theories of ethics shape consumers' deference to experts. *Journal of Behavioral Decision Making*, *34*(1), 47-65.

Jordan, J. J., & Kteily, N. S. (2022). *Reputation fuels moralistic punishment that people judge to be questionably merited.* PsyArXiv. 10.31234/osf.io/97nhj

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57-88.

Joyce, R. (2001). *The myth of morality*. Cambridge, UK: Cambridge University Press.

Joyce, R. (2006). *The evolution of morality*. Cambridge, MA: MIT Press.

Joyce, R. (2011a). Moral fictionalism. *Philosophy Now*, *82*, 14-17.

Joyce, R. (2011b). The error in 'The error in the error theory'. *Australasian Journal of Philosophy*, *89*(3), 519-534.

Joyce, R. (2012). Enough with the errors! A final reply to Finlay. [Unpublished manuscript].

Joyce, R. (2015). Moral objectivity and moral relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy archive.* https://stanford.library.sydney.edu.au/archives/spr2017/entries/moral-anti-realism/moral-objectivity-relativism.html

Joyce, R. (2019). Moral fictionalism: How to have your cake and eat it too. In R. Garner & R. Joyce (Eds.), *The end of morality: Taking moral abolitionism seriously* (pp. 150-166) New York, NY: Routledge.

Joyce, R. (2020). Moral and epistemic normativity: The guilty and the innocent. In C. Cowie & R. Rowland (Eds.), *Companions in guilt arguments in metaethics* (pp. 53-72). Abingdon, UK: Routledge.

Joyce, R. (2021). Moral anti-realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 ed.). https://plato.stanford.edu/entries/moral-anti-realism/

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54-69.

Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., Rennie, D., & PEER Investigators. (1998). Does masking author identity improve peer review quality?: A randomized controlled trial. *JAMA, 280*(3), 240-242.

Kagan, S. (1997). *Normative ethics. Dimensions of philosophy series.* Boulder, CO: Westview Press.

Kahane, G. (2013). Must metaethical realism make a semantic claim?. *Journal of Moral Philosophy, 10*(2), 148-178.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131-164.

Kahlke, R. M. (2014). Generic qualitative approaches: Pitfalls and benefits of methodological mixology. *International Journal of Qualitative Methods, 13*(1), 37-52.

Kalderon, M. E. (2005). Moral fictionalism. New York, NY: Oxford University Press.

Kallestrup, J. (2013). *Semantic externalism.* London, UK: Routledge.

Kelly, D., & Stich, S. (2007). Two theories about the cognitive architecture underlying morality. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 348-366). New York, NY: Oxford University Press.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*(2), 117-131.

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: benefits, critiques, & a survival guide. *EJIFCC, 25*(3), 227-243.

Kern, F. G., & Gleditsch, K. S. (2017). Exploring pre-registration and pre-analysis plans for qualitative inference. *ResearchGate.* https://www.researchgate.net/publication/319141144_Exploring_Pre-registration_and_Pre-analysis_Plans_for_Qualitative_Inference

Khoo, J., & Knobe, J. (2018). Moral disagreement and moral semantics. *Noûs, 52*(1), 109-143.

Killen, M., McGlothlin, H., & Lee-Kim, J. (2002). Between individuals and culture: Individuals' evaluations of exclusion from social groups. In H. Keller, Y. Poortinga, & A. Schölmerich (Eds.), *Between biology and culture: Perspectives on ontogenetic development* (pp. 159-190). Cambridge, UK: Cambridge University Press.

Kim, M., & Yuan, Y. (2015). No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001. *Episteme*, *12*(3), 355-361.

*King James Version Bible.* (2022). King James Bible Online. https://www.kingjamesbibleonline.org/Romans-Chapter-2/

Kirchin, S. (Host). (2022, June 26). PGS - Metaethics - Part ONE - Moral realism. In *Philosophy gets schooled.* https://www.podbean.com/ew/pb-riik3-125c463Kirwin, C. (2021). Value realism and idiosyncrasy. Unpublished manuscript. Retrieved from https://elucidations.vercel.app/value_realism_and_idiosyncrasy.pdf

Kissinger-Knox, A., Aragon, P., & Mizrahi, M. (2018). Ought implies can," framing effects, and "empirical refutations. *Philosophia*, *46*(1), 165-182.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385-418.

Knobe, J. & S. Nichols (2008). An experimental philosophy manifesto. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 3-14). Oxford, UK: Oxford University Press.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190-194.

Knobe, J. (2007). Experimental philosophy. *Philosophy Compass*, *2*(1), 81-92.

Knobe, J. (2016). Experimental philosophy is cognitive science. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 37-52) Malden, MA: Wiley-Blackwell.

Knobe, J., & Nichols, S. (2017). Experimental philosophy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017 ed.). https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *Annual Review of Psychology*, *63*, 81-99.

Knowles, E. S., & Nathan, K. T. (1997). Acquiescent responding in self-reports: cognitive style or social concern?. *Journal of Research in Personality*, *31*(2), 293-301.

Kohlberg, L., & Kramer, R. (1969). Continuities and discontinuities in childhood and adult moral development. *Human Development*, *12*(2), 93-120.

Komatsu, L. K., & Galotti, K. M. (1986). Children's reasoning about social, physical, and logical irregularities: A look at two worlds. *Child Development, 57*(2), 413-420.

Koonin, E. V. (2009). The Origin at 150: is a new evolutionary synthesis in sight?. *Trends in Genetics*, *25*(11), 473-475.

Koons, R. C. (2022). A dynamic B theory of time. In C. J. Austin, A. Marmodoro, & A. Roselli (Eds.), *Powers, time, and free will* (pp. 69-82). Cham, CH: Springer.

Korman, D. and Carmichael, C. (2017). What do the folk think about composition and does it matter?. In D. Rose (Ed..), *Experimental metaphysics* (pp. 187-206). London, UK: Bloomsbury Academic.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75-84.

Kvale, S. (1994). Ten standard objections to qualitative research interviews. *Journal of Phenomenological Psychology*, *25*(2), 147-173.

Lassiter, D. (2008). Semantic externalism, language variation, and sociolinguistic accommodation. *Mind & Language*, *23*(5), 607-633.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods* (Vol. 2). Thousand Oaks, CA: Sage.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, *6*(1), 2158244016636433.

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2021). Religious affiliation and conceptions of the moral domain. *Social Cognition*, *39*(1), 139-165.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Levy, N. (2003). Descriptive relativism: Assessing the evidence. *Journal of Value Inquiry*, *37*, 165-177.

Lewis, D. (1983). *Philosophical papers: Volume I*. New York, NY: Oxford University Press.

Lewkowicz, D. J. (2001). The concept of ecological validity: What are its limitations and is it bad to be invalid?. *Infancy*, *2*(4), 437-450.

Li, J., Zhao, Y., Lin, L., Chen, J., & Wang, S. (2018). The freedom to persist: Belief in free will predicts perseverance for long-term goals among Chinese adolescents. *Personality and Individual Differences, 121*, 7-10.

Liddle, J. R., Bush, L. S., & Shackelford, T. K. (2011). An introduction to evolutionary psychology and its application to suicide terrorism. *Behavioral Sciences of Terrorism and Political Aggression*, *3*(3), 176-197.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 281-288.

Lillehammer, H. (2007). Davidson on value and objectivity. *Dialectica*, *61*(2), 203-217.

Lim, D., & Chen, J. (2018). Is compatibilism intuitive?. *Philosophical Psychology*, *31*(6), 878-897.

Lindauer, M., & Southwood, N. (2021). How to cancel the Knobe effect: the role of sufficiently strong moral censure. *American Philosophical Quarterly*, *58*(2), 181-186.

Linnebo, O. (2018). Platonism in the philosophy of mathematics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 ed.). https://plato.stanford.edu/entries/platonism-mathematics/

Lockhart, K. L., Abrahams, B., & Osherson, D. N. (1977). Children's understanding of uniformity in the environment. *Child Development, 48*(4), 1521-1531.

Loeb, D. (2003). Gastronomic realism—A cautionary tale. *Journal of Theoretical and Philosophical Psychology*, *23*(1), 30-49.

Loeb, D. (2008). Moral incoherentism: How to pull a metaphysical rabbit out of a semantic hat. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality* (Vol. 2, pp. 355-386). Cambridge, MA: MIT Press.

Lopez, J., & Hillygus, D. S. (2018). Why so serious?: Survey trolls and misinformation. *Why so serious.* [Unpublished manuscript]. http://dx.doi.org/10.2139/ssrn.3131087

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, *31*, 128-159.

Lycan, W. G. (2019). *On evidence in philosophy*. Oxford, UK: Oxford University Press.

Machery, E. & Stich, S. (2022). The Moral/conventional distinction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 ed.). https://plato.stanford.edu/archives/sum2022/entries/moral-conventional/

Machery, E. (2017). *Philosophy within its proper bounds*. Oxford, UK: Oxford University Press.

Machery, E. (2018). Morality: A historical invention. In K. J. Gray & J. Graham (Eds.), Atlas of moral psychology (pp. 259-265). New York, NY: The Guilford Press.

Machery, E., & Mallon, R. (2010). Evolution of morality. In J. M. Doris & The Moral Psychology Research Group (Eds.). *The moral psychology handbook* (pp. 3-47). New York, NY: Oxford University Press.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1-B12.

Machery, E., & Stich, S. (2022). The moral/conventional distinction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 ed.). https://plato.stanford.edu/archives/sum2022/entries/moral-conventional/

MacIntyre, A. (1957). What morality is not. *Philosophy*, *32*(123), 325-335.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. London, UK: Penguin.

Madans, J., Miller, K., Maitland, A., & Willis, G. B. (2011). *Question evaluation methods: Contributing to the science of data quality*. Hoboken, NJ: John Wiley & Sons.

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, *6*, Article 2, 1-7. https://doi.org/10.3389/fpsyg.2015.00002

Mainwood, P. (2020). Richard Feynman said, "I think I can safely say that nobody understands quantum mechanics." What is your opinion or perspective about this? What did he mean?. *Quora.* https://www.quora.com/Richard-Feynman-said-I-think-I-can-safely-say-that-nobody-understands-quantum-mechanics-What-is-your-opinion-or-perspective-about-this-What-did-he-mean/answer/Paul-Mainwood

Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research*, *79*(2), 332-356.

Martí, G. (2009). Against semantic multi-culturalism. *Analysis*, *69*(1), 42-48.

Martin, R. A. (2007). *The psychology of humor: An integrative approach*. Burlington, MA: Elsevier Academic Press.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51-69.

Mavrogenis, A. F., Quaile, A., & Scarlat, M. M. (2020). The good, the bad and the rude peer-review. *International Orthopaedics*, *44*(3), 413-415.

Maxwell, N. (2006). Special relativity, time, probabilism and ultimate reality. *Philosophy and Foundations of Physics*, *1*, 229-245.

May, J., & Holton, R. (2012). What in the world is weakness of will?. *Philosophical Studies, 157*(3), 341-360.

McCambridge, J., De Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS One*, *7*(6), e39116. https://doi.org/10.1371/journal.pone.0039116

McCarthy, D. (2016). Probability in ethics. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of philosophy and probability* (pp. 705-737). Oxford, UK: Oxford University Press.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science*, *21*(8), 1141-1149.

McGraw, A. P., Schiro, J. L., Fernbach, P. M. (2015). Not a problem: A downside of humorous appeals. *Journal of Marketing Behavior, 1*(2), 187-208.

McGraw, A. P., Williams, L. E., & Warren, C. (2014). The rise and fall of humor: Psychological distance modulates humorous responses to tragedy. *Social Psychological and Personality Science*, *5*(5), 566-572.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

McNaughton, D. (1988). *Moral vision: An introduction to ethics*. Oxford, UK: Blackwell.

McPherson Frantz, C. (2006). I AM being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology*, *28*(2), 157-167.

Medinnus, G. R. (1962). Objective responsibility in children: A comparison with the Piaget data. *The Journal of Genetic Psychology*, *101*(1), 127-133.

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates?. *PloS One*, *15*(6), e0235403. https://doi.org/10.1371/journal.pone.0235403

Menold, N. (2020). Double barreled questions: An analysis of the similarity of elements and effects on measurement quality. *Journal of Official Statistics*, *36*(4), 855-886.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Merritt, J. (2016, March 25). The death of moral relativism. *The Atlantic*. http://www.theatlantic.com/politics/archive/2016/03/the-death-of-moral-relativism/475221/

Mihailov, E. (2022). Measuring Impartial Beneficence: A Kantian Perspective on the Oxford Utilitarianism Scale. *Review of Philosophy and Psychology*, 1-16. https://doi.org/10.1007/s13164-021-00600-2

Mijović-Prelec, D., & Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 227-240.

Miller, C. (2009). The conditions of moral realism. *Journal of Philosophical Research*, *34*, 123-155.

Miller, J. G., & Bersoff, D. M. (1988). When do American children and adults reason in social conventional terms?. *Developmental Psychology*, *24*(3), 366-375.

Milo, R. (1995). Contractarian constructivism. *The Journal of Philosophy*, *92*(4), 181-204.

Mizrahi, M. (2015). Ought, can, and presupposition: An experimental study. *Methode: Analytic Perspectives, 4*(6), 232-243.

Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, *93*, 1-17.

Monroe, A. E., & Malle, B. F. (2015). Free will without metaphysics. In A. R. Mele (Ed.), *Surrounding free will* (pp. 25-48). Oxford, UK: Oxford University Press.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760-775.

Moody-Adams, M. M. (2009). Fieldwork in familiar places: Morality, culture, and philosophy. Cambridge, MA: Harvard University Press.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379-387.

Moomal, Z., & Henzi, S. P. (2000). The evolutionary psychology of deception and self-deception. *South African Journal of Psychology*, *30*(3), 45-51.

Moral parochialism misunderstood: a reply to Piazza and Sousa. *Proceedings of the Royal Society B*, *283*(1823), 20152628. https://doi.org/10.1098/rspb.2015.2628

Morreall, J. (2009). *Comic relief: A comprehensive philosophy of humor*. Malden, MA: Wiley-Blackwell.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, *40*(2), 120-123.

Moss, D. (2017). Experimental philosophy, folk metaethics and qualitative methods. *Teorema: Revista Internacional de Filosofía 36*(3), 185-203.

Moss, D., & Bush, L. S. (2021). Measuring metaaesthetics: Challenges and ways forward. *New Ideas in Psychology*, *62*, 100866. https://doi.org/10.1016/j.newideapsych.2021.100866

Moss, D., Montealegre, A., Bush, L.S., Caviola, L., & Pizarro, D. (n.d.). *The reputational implications of relativism and objectivism* [Unpublished manuscript].

Munier, B. R. (1988). A guide to decision-making under uncertainty. In B. R. Munier (Ed.), *Risk, decision and rationality* (pp. 1-34). Dordrecht, NL: D. Reidel Publishing Company.

Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434-467.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678-701.

Myers, R. H. (2004). Finding value in Davidson. *Canadian Journal of Philosophy*, *34*(1), 107-136.

Nadelhoffer, T., & Monroe, A. (2022). Experimental philosophy of free will and moral responsibility: An introduction. In T. Nadelhoffer & A. Monroe (Eds.), *Advances in experimental philosophy of free will and responsibility* (pp. 1-11). London, UK: Bloomsbury Academic.

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations*, *10*(2), 123-149.

Nadelhoffer, T., Murray, S., & Murry, E. (2021). Intuitions About Free Will and the Failure to Comprehend Determinism. *Erkenntnis*, 1-22. https://doi.org/10.1007/s10670-021-00465-y

Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (2020). Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science*, *44*(8), e12873. https://doi.org/10.1111/cogs.12873

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, *25*, 27-41.

Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition, 129*(3), 652-661.

Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, *6*(1-2), 215-237.

Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford handbook of free will* (2nd ed., pp. 555-576). New York, NY: Oxford University Press.

Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189-216). New York, NY: Palgrave-Macmillan.

Nahmias, E., Morris, S. G., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive?. *Philosophy and Phenomenological Research*, *73*(1), 28-53.

Nahmias, E., Morris, S., & Nadelhoffer, T. (2004). The phenomenology of free will. *Journal of Consciousness Studies*, *11*(7-8), 162-179.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561-584.

Navarro-Plaza, M. C., Gomis-Pont, S., Mirete-Fructuoso, M., Molina-Rodríguez, S., & Martinez, L. M. (2020). To what extend is trading ecological validity for experimental control justified in moral research?. PsyArXiv. https://doi.org/10.31234/osf.io/eykcd

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language 85*(2), 355-404.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1-19.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*(2), 151-166.

Nichols, S. (2004). After objectivity: An empirical study of moral judgment. *Philosophical Psychology*, *17*(1), 3-26.

Nichols, S. (2008). How can psychology contribute to the free will debate? In K. Baer, J. C, Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 10-31). Oxford, UK: Oxford University Press.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, *331*(6023), 1401-1403.

Nichols, S. (2012). The indeterminist intuition. *The Monist, 95*(2), 290-307.

Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, *23*(3), 293-312.

Nichols, S., & Folds-Bennett, T. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition*, *90*(2), B23-B32.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*(4), 663-685.

Nichols, S., Stich, S., & Weinberg, J. (2003). Metaskepticism: Meditations in ethno-epistemology. In S. Luper (Ed.), *The skeptics* (pp. 227-247), Burlington VT: Ashgate.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

Nisan, M. (1987). Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology, 23*(5), 719-725.

Nisan, M. (1988). A story of a pot: or A cross-cultural comparison of basic moral evaluations: A response to the critique by Turiel, Nucci, and Smetana (1988). *Developmental Psychology, 24*(1), 144-146.

Nisbett, R. (2004). *The geography of thought: How Asians and Westerners think differently... and why*. New York, NY: Simon & Schuster.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108*(2), 291-310.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653-684.

Nucci, L. & Nucci, M. (1982). Children's social interactions in the context of moral and conventional transgressions. *Child Development, 53(2)*, 403-412.

Nucci, L. & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*(2), 400-407.

Nucci, L. P. (2001). *Education in the moral domain*. Cambridge, UK: Cambridge University Press.

Nucci, L. P., Turiel, E., & Encarnacion-Gawrych, G. (1983). Children's social interactions and social concepts: Analyses of morality and convention in the Virgin Islands. *Journal of Cross-Cultural Psychology*, *14*(4), 469-487.

Nucci, L., & Turiel, E. (1993). God's word, religious rules, and their relation to Christian and Jewish children's concepts of morality. *Child Development, 64*(5), 1475-1491.

Nucci, L., Camino, C., & Sapiro, C. M. (1996). Social class effects on northeastern Brazilian children's conceptions of areas of personal choice and social regulation. *Child Development*, *67*(3), 1223-1242.

Office of International Religious Freedom. (2021, May 12). *2020 report on International Religious Freedom: Brazil*. U.S. Department of State. https://www.state.gov/reports/2020-report-on-international-religious-freedom/brazil/

Olsen, W. 2004. Triangulation in social research: qualitative and quantitative methods can really be mixed. In M. Holborn (Ed.), *Developments in sociology: An annual review* (Vol. 20, pp. 103-121). Ormskirk, UK: Causeway Press.

Olson, J. (2014). *Moral error theory: History, critique, defense*. Oxford, UK: Oxford University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. 10.1126/science.aac4716

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776-783.

Ort, A., & Fahr, A. (2020). The effectiveness of a positively vs. negatively valenced PSA against sexually transmitted diseases–Evidence from an experimental study. *SCM Studies in Communication and Media*, *9*(3), 341-366.

Oxford University Press. (2021). *Objective*, adj. and n. In OED online. www.oed.com/view/Entry/129634

Paden, R. (1994). The natural history of student relativism. *Journal of Thought*, *29*(2), 47-58.

Parfit, D. (2011). *On what matters* (Vol. 2). Oxford, UK: Oxford University Press.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, *23*(10), 3162-3180.

Paul, L. A. (2010). Temporal experience. *The Journal of Philosophy, 107*(7), 333-359.

Peters, U. (2022). What is the function of confirmation bias?. *Erkenntnis*, *87*, 1351-1376.

Peterson, J. (2018). *12 Rules for Life: An Antidote to Chaos*. Toronto, ON: Random House.

Pethokoukis, J. (2011). *Paul Ryan Speaks! Part two! Moral relativism, the EU debt crisis, inflation and more ....* American Enterprise Institute. https://www.aei.org/economics/paul-ryan-speaks-part-two-moral-relativism-the-eu-debt-crisis-inflation-and-more/

Pfister, J. (2019). Classification of strategies for dealing with student relativism and the epistemic conceptual change strategy. *Teaching Philosophy*, *42*(3), 221-246.

Phillips, J., Nyholm, S., & Liao, S. Y. (2014). The good in happiness. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 1, 253-293). Oxford, UK: Oxford University Press.

Piaget, J. (1997). *The moral judgment of the child* (M. Gabain, Trans.). New York, NY: Simon & Schuster. (Original work published 1932)

Piazza, J., & Sousa, P. (2016). When injustice is at stake, moral judgements are not parochial. *Proceedings of the Royal Society B*, *283*(1823), 20152037. https://doi.org/10.1098/rspb.2015.2037

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, *107*(supplement_2), 8993-8999.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). American Psychological Association.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903.

Pölzler, T. (2017). Revisiting folk moral realism. *Review of Philosophy and Psychology*, *8*(2), 455-476.

Pölzler, T. (2018a). Can the empirical sciences contribute to the moral realism/anti-realism debate?. *Synthese*, *195*(11), 4907-4930.

Pölzler, T. (2018b). How to measure moral realism. *Review of Philosophy and Psychology*, *9*(3), 647-670.

Pölzler, T. (2018c). *Moral reality and the empirical sciences*. New York, NY: Routledge.

Pölzler, T. (2021). Insufficient effort responding in experimental philosophy. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 4, pp. 214-256). Oxford, UK: Oxford University Press.

Pölzler, T., & Wright, J. C. (2019). Empirical research on folk moral objectivism. *Philosophy Compass*, *14*(5), e12589. https://doi.org/10.1111/phc3.12589

Pölzler, T., & Cole Wright, J. C. (2020a). An empirical argument against moral non-cognitivism. *Inquiry*, 1-29. https://doi.org/10.1080/0020174X.2020.1798280

Pölzler, T., & Wright, J. C. (2020b). Anti-realist pluralism: A new approach to folk metaethics. *Review of Philosophy and Psychology*, *11*(1), 53-82.

Pölzler, T., Zijlstra, L., & Dijkstra, J. (2022). Moral progress, knowledge and error: Do people believe in moral objectivity?. *Philosophical Psychology*, 1-37.

Population Density (n.d.) The world bank. https://data.worldbank.org/indicator/EN.POP.DNST

Potter, N. (1997). The synthetic a priori proposition of Kant's ethical philosophy. In B. S. Byrd, J. C. Joerden, & J. Hruschka (Eds.), *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics* (Vol. 5, pp. 437-459). Berlin, DE: Duncker & Humblot.

Povee, K., & Roberts, L. D. (2014). Qualitative research in psychology: Attitudes of psychology students and academic staff. *Australian Journal of Psychology*, *66*(1), 28-37.

Pravato, G. (2020). Indeterminacy and Normativity. *Erkenntnis*, 1-23. https://doi.org/10.1007/s10670-020-00293-6

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, *11*(1), 37-43.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381.

Pust, J. (2017). Intuitions. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). https://plato.stanford.edu/archives/sum2019/entries/intuition/

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90.

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European Journal of Education Studies 3*(9), 369-387.

Quintelier, K. J., & Fessler, D. M. (2012). Varying versions of moral relativism: The philosophy and psychology of normative relativism. *Biology & Philosophy*, *27*(1), 95-113.

Quintelier, K. J., De Smet, D., & Fessler, D. M. (2013). The moral universalism-relativism debate. *KLESIS–Revue Philosophique*, *27*, 211-262.

Quintelier, K., De Smet, D., & Fessler, D. (2014). Agent versus appraiser moral relativism: an exploratory study. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 209-230). London, UK: Bloomsbury Academic.

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416-426.

Rabb, N., Han, A., Nebeker, L., & Winner, E. (2020). Expressivist to the core: Metaaesthetic subjectivism is stable and robust. *New Ideas in Psychology*, *57*, 100760. https://doi.org/10.1016/j.newideapsych.2019.100760

Rabkin, E. S. (1979). Metalinguistics and Science Fiction. *Critical Inquiry*, *6*(1), 79-97.

Radziwill, N. (2015, March 16). One-proportion Z test in R. *Quality and innovation*. https://qualityandinnovation.com/2015/03/16/one-proportion-z-test-in-r/

Rahim, L. Z. (2010). *Singapore in the Malay world: Building and breaching regional bridges*. New York, NY: Routledge.

Rahman, M. S. (2020). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review. *Journal of Education and Learning 6*(1), 102-112.

Rai, T. S., & Holyoak, K. J. (2013). Exposure to moral relativism compromises moral behavior. *Journal of Experimental Social Psychology*, *49*(6), 995-1001.

Reudell, R. (2021, April 8). *Why moral relativism is dangerous*. A Philosopher's Stone. https://medium.com/the-philosophers-stone/why-moral-relativism-is-dangerous-9b1e5da62bd7

Reuter, K., & Brun, G. (2022) Empirical studies in truth and the project of re-engineering truth. *Pacific Philosophical Quarterly*. https://doi.org/10.1111/papq.12370

Reynolds, N., Diamantopoulos, A., & Schlegelmilch, B. (1993). Pre-testing in questionnaire design: A review of the literature and suggestions for further research. *Market Research Society Journal*, *35*(2), 1-11.

Rich, M. C. (1979). Verbal reports on mental processes: Issues of accuracy and awareness. *Journal for the Theory of Social Behaviour, 9*(1), 29-37.

Rose, D., & Nichols, S. (2019). From punishment to universalism. *Mind & Language*, *34*(1), 59-72.

Rose, D., & Schaffer, J. (2017). Folk mereology is teleological. In D. Rose (Ed.), *Experimental metaphysics* (pp. 135-186). London, UK: Bloomsbury Academic.

Rose, D., Schaffer, J., & Tobia, K. (2020). Folk teleology drives persistence judgments. *Synthese*, *197*(12), 5491-5509.

Rosenthal, G. S. (1978). *The many faces of Judaism: Orthodox, Conservative, Reconstructionist, & Reform*. New York: Behrman House.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, *112*(1), 159-174.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Ryazanov, A. A., Knutzen, J., Rickless, S. C., Christenfeld, N. J., & Nelkin, D. K. (2018). Intuitive probabilities and the limitation of moral imagination. *Cognitive Science*, *42*, 38-68.

Ryskin, R. A., & Brown-Schmidt, S. (2014). Do adults show a curse of knowledge in false-belief reasoning? A robust estimate of the true effect size. *PloS One*, *9*(3), e92406. https://doi.org/10.1371/journal.pone.0092406

Sarkissian, H. (2016). Aspects of Folk Morality. In J. Sytsma, & W. Buckwalter (Eds.), A companion to experimental philosophy (pp. 212-224). West Sussex, UK: Wiley-Blackwell.

Sarkissian, H., & Phelan, M. (2019). Moral objectivism and a punishing God. *Journal of Experimental Social Psychology*, *80*, 1-7.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal?. *Mind & Language*, *25*(3), 346-358.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language*, *26*(4), 482-505.

Satris, S. A. (1986). Student relativism. *Teaching Philosophy*, *9*(3), 193-205.

Sauer, H. (2019). The argument from agreement: How universal values undermine moral realism. *Ratio*, *32*(4), 339-352.

Savitsky, K., & Gilovich, T. (2003). The illusion of transparency and the alleviation of speech anxiety. *Journal of Experimental Social Psychology*, *39*(6), 618-625.

Sayre-McCord, G. (2014). Metaethics. In E. D. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2014 ed.). https://plato.stanford.edu/archives/sum2014/entries/metaethics

Sayre-McCord, G. (2015). Moral realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). https://plato.stanford.edu/archives/sum2021/entries/moral-realism/

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford, UK: Oxford University Press.

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*, 1-9. https://doi.org/10.15626/MP.2019.1645

Schlagel, R. H. (1974). Contra Wittgenstein. *Philosophy and Phenomenological Research*, *34*(4), 539-550.

Schmidt, M. F., Gonzalez-Cabrera, I., & Tomasello, M. (2017). Children's developing metaethical judgments. *Journal of Experimental Child Psychology*, *164*, 163-177.

Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, *23*(50), 844-849.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J., & Henrich, J. (2018, June 22). *The origins of WEIRD psychology*. [Unpublished manuscript]. https://doi.org/10.31234/osf.io/d6qhu

Schwardmann, P., & Van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055-1061.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195-202.

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, *6*(1), 38-47.

Semler, J., & Henne, P. (2019). Recent experimental work on "ought" implies "can". *Philosophy Compass*, *14*(9), e12619. https://doi.org/10.1111/phc3.12619

Shafer-Landau, R. (2003). *Moral realism: A defense*. Oxford, UK: Clarendon Press.

Shakespeare, W. (2020). *The Tragedy of Julius Caesar*. B. A. Mowat & P. Westerine (Eds.). Folger Shakespeare Library. (Original work published 1599)

Shaw, A., Choshen-Hillel, S., & Caruso, E. M. (2016). The development of inequity aversion: Understanding when (and why) people give others the bigger piece of the pie. *Psychological Science*, *27*(10), 1352-1359.

Shoemaker, D. & Tobia, K. (2022). Personal identity. In M. Vargas & J. M. Doris (Eds.), *Oxford handbook of moral psychology* (543-563). Oxford, UK: Oxford University Press.

Shweder, R. A. (1990). In defense of moral realism: Reply to Gabennesch. *Child Development*, *61*(6), 2060-2067.

Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1-83). Chicago, IL: University of Chicago Press.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119-169). New York, NY: Routledge.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Sinclair, N. (2012). Moral realism, face-values and presumptions. *Analytic Philosophy*, *53*(2), 158-179.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4-31). New York, NY: Psychology Press.

Sinnott-Armstrong, W. (2009). Mixed-up meta-ethics. *Philosophical Issues 19*(1), 235-256.

Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. *The Monist*, *95*(3), 355-377.

Sinnott-Armstrong, W., & Wheatley, T. (2014). Are moral judgments unified?. *Philosophical Psychology*, *27*(4), 451-474.

Skitka, L. J., Bauman, C. W. and Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength of something more?. *Journal of Personality and Social Psychology*, *88*(6), 895-917.

Smetana, J. (1993). Understanding of social rules. In M. Bennett (Ed.), *The child as psychologist: An introduction to the development of social cognition* (pp. 111-141). New York, NY: Guilford Press.

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*, 1333-1336.

Smetana, J. G., & Braeges, J. L. (1990). The development of toddlers' moral and conventional judgments. *Merrill-Palmer Quarterly, 36*(3), 329-346.

Smetana, J. G., Kelly, M., & Twentyman, C. T. (1984). Abused, neglected, and nonmaltreated children's conceptions of moral and social-conventional transgressions. *Child Development, 55*(1), 277-287.

Smith, M. (1994). The moral problem. Malden, MA: Blackwell.

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*(4), 178-182.

Smith, R. (2010). Classical peer review: an empty gun. *Breast Cancer Research*, *12*(4), 1-4.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review, 85*(4), 355-362.

Solzhenitsyn, A. (2009). *Détente, democracy and dictatorship* (3rd ed.). New York, NY: Routledge.

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass*, *5*(2), 199-212.

Song, M. J., Smetana, J. G., & Kim, S. Y. (1987). Korean children's conceptions of moral and conventional transgressions. *Developmental Psychology*, *23*(4), 577-582.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, *132*(1), 99-107.

Sosa, E. (2009). A defense of the use of intuitions in philosophy. In D. Murphy & M. Bishop (Eds.) *Stich and his critics* (pp. 101-112). West Sussex, UK: Wiley-Blackwell.

Sosa, E. (2014). Intuitions: Their nature and probative value. In A. R. Booth & D. P. Rowbottom (Eds.) *Intuitions* (pp. 36-49). Oxford, UK: Oxford University Press.

Sousa, P., Allard, A., Piazza, J., & Goodwin, G. P. (2021). Folk moral objectivism: The case of harmful actions. *Frontiers in Psychology*, *12*, Article 2776. https://doi.org/10.3389/fpsyg.2021.638515

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*(5), 495-518.

Stanford, P. (2018). The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation. *Behavioral and Brain Sciences, 41*, E95. doi:10.1017/S0140525X17001911

Starmans, C., & Bloom, P. (2018a). If you become evil, do you die?. *Trends in Cognitive Sciences*, *22*(9), 740-741.

Starmans, C., & Bloom, P. (2018b). Nothing personal: What psychologists get wrong about identity. *Trends in Cognitive Sciences*, *22*(7), 566-568.

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition, 124*(3), 272-283.

Sterelny, K., & Fraser, B. (2016). Evolution and moral realism. *British Journal for the Philosophy of Science, 68*(4), 981-1006.

Stevenson, C. L. (1963). *Facts and values*. New Haven, CT: Yale University Press.

Stevenson, C. L. (1963). *Facts and values: Studies in ethical analysis*. New Haven, CT: Yale University Press.

Stich, S. (2018). The moral domain. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 547-555). New York, NY: Guilford Press.

Stich, S. (2019). The quest for the boundaries of morality. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge Handbook of Moral Epistemology* (pp. 15-38). New York, NY: Routledge.

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 5-21) Malden, MA: Wiley-Blackwell.

Stojanovic, I., & McNally, L. (2022). Are moral predicates subjective? A corpus study. In D. Bordonaba (Ed.), *Experimental philosophy of language: Perspectives, methods, and prospects* [Unpublished manuscript]. Springer. 10.13140/RG.2.2.23636.88965

Stoljar, D. Damnjanovic, N. (2007). The deflationary theory of truth. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2010 ed.). https://stanford.library.sydney.edu.au/archives/spr2010/entries/truth-deflationary/

Stroebe, W., Gadenne, V., & Nijstad, B. A. (2018). Do our psychological laws apply only to college students?: External validity revisited. *Basic and Applied Social Psychology, 40*(6), 384-395.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159-171.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26*(9), 1469-1479.

Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition, 119*(2), 295-300.

Stuenkel, O. (2011). Identity and the concept of the West: the case of Brazil and India. *Revista Brasileira de Política Internacional, 54*(1), 178-195.

Suárez Álvarez, J., Pedrosa, I., Lozano, L. M., García Cueto, E., Cuesta Izquierdo, M., & Muñiz Fernández, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*(2), 149-158.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research, 76*(1), 138-155.

Swee-Hock, S. (2012). *The population of Singapore* (3rd ed.). Pasir Panjang, Singapore: ISEAS Publishing.

Sytsma, J., Bishop, R., & Schwenkler, J. (2022). Has the Side-Effect Effect been cancelled?(No, not yet.). *Synthese, 200*(5), 1-15.

Sytsma, J., & Livengood, J. (2012). Experimental philosophy and philosophical disputes. *Essays in Philosophy, 13*(1), 145-161.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53-55.

Taylor, S. E. (1982). The availability bias in social perception and interaction. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 190-200). New York, NY: Cambridge University Press.

Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity and Peer Review, 5*(1), 1-14.

Tenney, M. (2015, March 11). *The dangers of moral relativism*. United Families International. https://www.unitedfamilies.org/life/abortion/the-dangers-of-moral-relativism/

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2017). Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General, 146*(11), 1586-1605.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2020). Theory of Mind network activity is associated with metaethical judgment: An item analysis. *Neuropsychologia, 143*, 107475. https://doi.org/10.1016/j.neuropsychologia.2020.107475

Thompson, K. (2022). Qualitative methods show that surveys misrepresent "ought implies can" judgments. *Philosophical Psychology*, 1-29. https://doi.org/10.1080/09515089.2022.2036714

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., & Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Science, 64*(3), 1458-1476.

Tisak, M. (1995). Domains of social reasoning and beyond. *Annals of Child Development, 11*(1), 95-130.

Tisak, M. S., & Turiel, E. (1988). Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology, 24*(3), 352-357.

Tisak, M. S., Crane-Ross, D., & Tisak, J. (2000). Mothers' and teachers' home and school rules: Young children's conceptions of authority in context. *Merrill-Palmer Quarterly, 46*, 168-187.

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs, 44*(3), 197-217.

Tosi, J., & Warmke, B. (2020). *Grandstanding: The use and abuse of moral talk*. New York, NY: Oxford University Press.

Trainer, F. E. (1983). Ethical objectivism-subjectivism: a neglected dimension in the study of moral thought. *Journal of Moral Education, 12*(3), 192-207.

Trask, L. (1997). Scare quotes. *University of Sussex*. https://www.sussex.ac.uk/informatics/punctuation/quotes/scare

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. New York, NY: Basic Books.

Turek, F. [@DrFrankTurek]. (2020, August 3). *If there is no objective morality then love is no better than murder* [Tweet]. Twitter. https://twitter.com/drfrankturek/status/1290347092486086658

Turiel, E. (1974). Conflict and transition in adolescent moral development. *Child Development, 45*(1), 14-29.

Turiel, E. (1979). Distinct conceptual and developmental domains: social convention and morality. In H. Howe & C. Keasey (Eds.), *Nebraska symposium on motivation, 1977: Social cognitive development*. Lincoln: University of Nebraska Press.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. New York, NY: Cambridge University Press.

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. New York, NY: Cambridge University Press.

Turiel, E. (2012). Moral reasoning, cultural practices, and social inequalities. *Innovación Educativa, 12*(59), 17-32.

Turiel, E., Nucci, L. P., & Smetana, J. G. (1988). A cross-cultural comparison about what? A critique of Nisan's (1987) study of morality and convention. *Developmental Psychology, 24*(1), 140-143.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232.

United Nations Development Programme (2020). *Human development report 2020: The next frontier: Human development and the Anthropocene*. New York, NY: United Nations.

United States Census Bureau (2021, October 8). *Urban areas facts*. United States Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/ua-facts.html

United States Census Bureau (2022, February 24). *Census bureau releases new educational attainment data*. United States Census Bureau. https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html

Uskul, A. K., Nisbett, R. E., & Kitayama, S. (2008). Ecoculture, social interdependence, and holistic cognition: Evidence from farming, fishing, and herding communities in Turkey. *Communicative & Integrative Biology*, *1*(1), 40-41.

Uttich, K., Tsai, G., & Lombrozo, T. (2014). Exploring metaethical commitments: Moral objectivity and moral progress. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 108-208). London, UK: Bloomsbury.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383-403.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science-Cambridge*, *17*(6), 476-477.

Van Bavel, J. J., Packer, D. J., Haas, I. J., & Cunningham, W. A. (2012). The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, universal evaluations of the same actions. *PloS One*, *7*(11), e48693. https://doi.org/10.1371/journal.pone.0048693

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492.

van den Berg, R. G. (2022). SPSS RM ANOVA – 2 Within-subjects factors. *SPSS Tutorials*. https://www.spss-tutorials.com/spss-repeated-measures-anova-2-within-subjects-factors/

van Dongen, N., Colombo, M., Romero, F., & Sprenger, J. (2021). Intuitions about the reference of proper names: A meta-analysis. *Review of Philosophy and Psychology*, *12*(4), 745-774.

Van Inwagen, P. (1997). Materialism and the psychological-continuity account of personal identity. *Philosophical Perspectives*, *11*, 305-319.

Van Roojen, M. (2015). *Metaethics: A contemporary introduction*. New York, NY: Routledge.

Van Roojen, M. (2018). Moral cognitivism vs. non-cognitivism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018 ed.). https://plato.stanford.edu/entries/moral-cognitivism/

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70.

Väyrynen, P. (2021). Thick ethical concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 ed.). https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/

Veluwenkamp, H. (2017). Parfit's and Scanlon's Non-Metaphysical Moral Realism as Alethic Pluralism. *Ethical Theory and Moral Practice*, *20*(4), 751-761.

Viciana, H., Hannikainen, I. R., & Gaitan Torres, A. (2019). The dual nature of partisan prejudice: Morality and identity in a multiparty system. *PloS One*, *14*(7), e0219509. https://doi.org/10.1371/journal.pone.0219509

Viciana, H., Hannikainen, I. R., & Rodríguez-Arias, D. (2021). Absolutely right and relatively good: Consequentialists see bioethical disagreement in a relativist light. *AJOB Empirical Bioethics*, *12*(3), 190-205.

Villars, C. N. (1986). The paradox of Schrodinger's cat. *Physics Education*, *21*(4), 232-237.

Vogt, P. J., & Goldman, A. (Hosts). (2016, May 12). On the inside (No. 64) [Audio podcast episode]. In *Reply all*. Gimlet. https://gimletmedia.com/shows/reply-all/brho4v/64-on-the-inside

Wagner, J. M., Pölzler, T., & Wright, J. C. (2021). Implicit metaethical intuitions: Validating and employing a new IAT procedure. *Review of Philosophy and Psychology*, 1-31. https://doi.org/10.1007/s13164-021-00572-3

Wainryb, C., Shaw, L. A., Langley, M., Cottam, K., & Lewis, R. (2004). Children's thinking about diversity of belief in the early school years: Judgments of relativism, tolerance, and disagreeing persons. *Child Development*, *75*(3), 687-703.

Walters, G. (2020, April 3). *Why young people should reject moral relativism to lead a better life*. Universal Owl. https://universalowl.com/2020/04/03/why-young-people-should-reject-moral-relativism-to-lead-a-better-life/

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, *29*(1/2), 429-460.

Welsh, M., Begg, S., & Bratvold, R. (2007). Efficacy of bias awareness in debiasing oil and gas judgments. D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1647-1652). Cognitive Science Society. https://hdl.handle.net/2440/41191

Wetzel, E., & Roberts, B. W. (2020). Commentary on Hussey and Hughes (2020): Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(4), 505-508.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review, 87*(1), 105-112.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202-217.

Wike, E. L., & Church, J. D. (1976). Comments on Clark's" The language-as-fixed-effect fallacy". *Journal of Memory and Language*, *15*(3), 249-255.

Wikforss, Å. (2008). Semantic externalism and psychological externalism. *Philosophy Compass*, *3*(1), 158-181.

Willig, C. (2008). Introducing qualitative research in psychology. New York, NY: Open University Press.

Willis, G. B. (2004). Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks, CA: Sage Publications.

Wineland, D. J. (2013). Nobel Lecture: Superposition, entanglement, and raising Schrödinger's cat. *Reviews of Modern Physics*, *85*(3), 1103-1114.

Wolff, J. (2018). An introduction to moral philosophy. New York, NY: W.W. Norton & Company.

Wong, D. B. (1995). Pluralistic relativism. *Midwest Studies in Philosophy*, *20*, 378-399.

Wong, D. B. (2006). *Natural moralities: A defense of pluralistic relativism*. New York, NY: Oxford University Press.

Wright, J. C. (2015). Meta-ethical pluralism: A cautionary tale about cohesive moral communities. *Behavioral and Brain Sciences*, *38*, e163, 44-44. doi:10.1017/S0140525X14001344

Wright, J. C. (2018). The fact and function of meta-ethical pluralism: Exploring the evidence. In T. Lombrozo, J. Knobe, S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 2, pp. 119-150). Oxford, UK: Oxford University Press.

Wright, J. C., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin*, *34*(11), 1461-1476.

Wright, J. C., Grandjean, P. T., & McWhite, C. B. (2013). The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology*, *26*(3), 336-361.

Wright, J. C., McWhite, C. B., & Grandjean, P. T. (2014). The cognitive mechanisms of intolerance: Do our meta-ethical commitments matter?. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 1, pp. 28-61). New York, NY: Oxford University Press.

Wygoda Cohen, S. (2021). Mind Independence versus Mind Nongroundedness: Two Kinds of Objectivism. *Ethics*, *132*(1), 180-203.

Yang, Z., Fu, D., Qi, Y., Zheng, Y., Li, Q., & Liu, X. (2019). Humor affects fairness considerations in the gain and loss contexts. *Frontiers in Psychology*, *9*, 2679. https://doi.org/10.3389/fpsyg.2018.02679

Yasenchuk, K. (1997). Moral realism and the burden of argument. *The Southern Journal of Philosophy*, *35*(2), 247-264.

Yau, J., & Smetana, J. G. (2003). Conceptions of moral, social-conventional, and personal events among Chinese preschoolers in Hong Kong. *Child Development*, *74*(3), 647-658.

Yilmaz, O., & Bahçekapili, H. G. (2015a). Without God, everything is permitted? The reciprocal influence of religious and meta-ethical beliefs. *Journal of Experimental Social Psychology*, *58*, 95-100.

Yilmaz, O., & Bahçekapili, H. G. (2015b). Without God, everything is permitted? The reciprocal influence of religious and meta-ethical beliefs. *Journal of Experimental Social Psychology*, *58*(Suppl. A), 95-100.

Yilmaz, O., & Bahçekapili, H. G. (2018). Meta-ethics and the mortality: Mortality salience leads people to adopt a less subjectivist morality. *Cognition*, *179*, 171-177.

Yilmaz, O., Bahcekapili, H. G., Harma, M., & Sevi, B. (2020). Intergroup tolerance leads to subjective morality, which in turn is associated with (but does not lead to) reduced religiosity. *Archive for the Psychology of Religion*, *42*(2), 232-243.

Young, L., & Durwin, A. J. (2013a). Moral realism as moral motivation: The impact of meta-ethics on everyday decision-making. *Journal of Experimental Social Psychology*, *49*(2), 302-306.

Young, L., & Durwin, A. J. (2013b). Moral realism as moral motivation: The impact of meta-ethics on everyday decision-making. *Journal of Experimental Social Psychology*, *49*(2, Suppl.), 302-306.

Zahn-Waxler, C., & Robinson, J. (1995). Empathy and guilt: Early origins of feelings of responsibility. In J. P. Tangney, & K. W. Fisher (Eds.), Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride (pp. 143-173). New York, NY: Guilford Press.

Zhu, J. H. (1996). 'I don't know' in public opinion surveys in China: Individual and contextual causes of item non-response. *Journal of Contemporary China*, *5*(12), 223-244.

Zijlstra, L. (2019). Folk moral objectivism and its measurement. *Journal of Experimental Social Psychology, 84*, 103807. https://doi.org/10.1016/j.jesp.2019.04.005

Zijlstra, L. (2021). Are people implicitly moral objectivists?. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00593-y

# Appendix A:
# Glossary

**Agent relativism**
A form of relativism which relativizes moral truth to the goals, standards, or values of agents, rather than those judging the agent. For instance, whether it would be morally right or wrong for Alex to steal would depend on Alex's standards, not the moral standards of whoever is judging Alex. (Note that this description focuses on the deontic properties of actions but could also be used for evaluative claims, judgments about character traits, thick moral concepts, and nonmoral norms.)

**Appraiser relativism**
A form of relativism which relativizes moral truth to whoever is judging an action, rather than the agent performing the action. For instance, whether it would be morally right or wrong for Alex to steal would depend on the standards of whoever is judging Alex's actions, not Alex's standards (Note that this description focuses on the deontic properties of actions but could also be used for evaluative claims, judgments about character traits, thick moral concepts, and nonmoral norms.)

**Clear intended interpretation**
An interpretation of the given stimuli that appears consistent with researcher intent and is sufficiently clear to be confidently classified as an instance of an intended interpretation.

**Clear interpretation**
An interpretation of the given stimuli that is sufficiently clear that it can be confidently classified as an instance of either an intended or unintended interpretation.

**Cognitivism**
The metaethical position that moral claims are propositional, and are therefore capable of being true or false.

**Commitment**
(See *philosophical commitment*)

**Conflation**
In the context of research on folk metaethics, conflations occur when researchers or participants conflate a particular metaethical concept or distinction with another concept or distinction (whether it be a different metaethical concept/distinction or a non-metaethical concept/distinction). (See *formal conflation* and *informal conflation*)

**Correspondence theory of truth**
A view of *truth* which holds that claims are true insofar as they *represent* (or "correspond") *reality*. In the context of moral claims, a correspondence-theoretic conception of truth would hold that moral claims are true insofar as they *represent moral facts* (Pölzler, 2018b, p. 662).

**Cultural relativism**

A form of relativism that relativizes moral truth to the moral standards of different cultures.

**Cultural relativism**

The metaethical position that moral claims are true or false relative to different cultural standards.

**Descriptive metaethics**

A branch of metaethics that seeks to offer the best account of the metaethical stances and commitments of moral thought and language. Could be described more simply as *the study of folk metaethics.*

**Descriptive relativism**

The descriptive claim that different individuals or societies have different moral values. In its stronger form, descriptive relativism refers to the notion that there are widespread and ineliminable fundamental moral disagreements, i.e., moral disagreements that cannot be reduced to differences in nonmoral beliefs.

**Determinate**

The position that there is a discoverable fact of the matter about whether competing propositions with respect to a given topic are true or best supported by reasons and evidence.

**Disagreement paradigm**

A common paradigm used to assess folk moral realism and antirealism. The disagreement paradigm presents a participant with moral (and potentially nonmoral) disagreements and asks them to judge whether one or both people who disagree can be (or are) correct. Response options are intended to reflect different realist and antirealist positions.

**Error theory | The error theory**

A metaethical position which holds that all first-order moral claims are false. Typically, error theories maintain that all first-order moral claims are false because they are implicitly committed to one or more false presuppositions, e.g., all moral claims implicitly presuppose that God exists, but God does not exist, so all such claims are false.

**Evaluative standard ambiguity**

An ambiguity that arises whenever there is insufficient information to know which moral framework a given moral claim could (if relativized) be indexed to. For instance, if we are told that Alex and Sam disagree about whether stealing is morally wrong, a cultural relativist cannot judge whether they can both be correct or whether one must be mistaken without knowing which culture(s) Alex and Sam are from, or which cultures Alex and Sam are referencing when making their claims.

**External validity**

The degree to which findings obtained in a given study justify inferences about people and situations outside the context of the study.

**Folk philosophy**
The philosophical stances and commitments of ordinary people. (See *ordinary people*)

**Forced choice**
A feature of studies in which participants are required to choose from among a limited set of response options that do not represent the full range of possible responses.

**Formal conflation**
In the context of research on folk metaethics, this occurs when stimuli erroneously conflate metaethical concepts or distinctions with other, unintended concepts or distinctions. For instance, researchers may use an item intended to measure *moral realism*, but accidentally use an item that a person competent with the relevant distinctions would recognize as an expression of universalism, or at least to be sufficiently ambiguous so as to plausibly be interpreted this way. Formal conflations are not the same as informal conflations. In principle, a competent person could spot a formal conflation and ignore it. Conversely, even if there is no formal conflation, someone might nevertheless conflate one concept/distinction with another concept/distinction.

**Fundamental moral disagreement**
Any moral disagreement that results from a genuine and ineliminable difference in moral values, and thus cannot be reduced to a difference in nonmoral beliefs.

**Indeterminacy / indeterminism**
The position that there is *no* discoverable fact of the matter about whether competing propositions with respect to a given topic are true or best supported by reasons and evidence.

**Indeterminacy-Variability Thesis | IV Thesis**
The hypothesis that folk metaethics is some combination of indeterminate or variable. Initially proposed by Michael Gill as an alternative to the UD assumption. (See *indeterminacy* and *variability*.)

**Informal conflation**
This occurs whenever a participant conflates a metaethical concept/distinction with a different concept/distinction, not as a result of experimenter error, but due to the participant failing to interpret face valid stimuli as intended.

**Intended interpretation**
An interpretation of the given stimuli that *appears* consistent with researcher intent. Intended interpretations may be clear or unclear, so responses coded as intended interpretations may not genuinely reflect a genuine instance of an intended interpretation.

**Interpersonal variability**
Metaethical pluralism with respect to the moral stances or commitments of different individuals or groups.

**Interpretative variation**

An index of the degree to which participants differ from one another in how they interpret the stimuli presented in a study. High interpretative variation would occur when participants frequently interpret stimuli differently from one another.

**Interpretative variation**

Systematic variation in how participants interpret experimental stimuli. For instance, some participants answering questions about a "bank" may interpret the term *bank* to refer to a riverbank, while others may interpret *bank* to refer to a financial institution. Interpretative variation may or may not be significant and may or may not have meaningful implications for a given study.

**Intrapersonal variability**

Metaethical pluralism with respect to the moral stances or commitments of an individual.

**Magic: The Gathering | MTG**

A collectible card game in which players take on the role of dueling wizards who summon monsters and cast spells. The goal of the game is to defeat all opponents. Typically, this involves reducing the life total of the other player (or players) to zero, though there are a variety of alternative ways to win.

**Metaethical indeterminacy**

The descriptive metaethical position that ordinary people have no determinate metaethical stances and/or commitments.

**Metaethical pluralism**

The descriptive metaethical position that ordinary people have different metaethical stances or commitments. For example, some people endorse realism or speak like realists, while others endorse antirealism or speak like antirealists. Pluralism can be both interpersonal and intrapersonal. Pluralism is not the mere recognition that some ordinary people may speak or think in unconventional ways, but that substantive and ineliminable elements of ordinary moral thought and language are best characterized by traditionally competing philosophical accounts (e.g., cognitivism and noncognitivism).

**Metaethical quietism**

The view that the dispute between realism and antirealism is a pseudoproblem resulting from linguistic and conceptual confusions. Technically an antirealist position, since quietism rejects realism, quietism nevertheless rejects much of the substantive content of conventional antirealist accounts as well.

**Metaethical variability**

The descriptive metaethical position that people do not have uniform metaethical stances or commitments, but instead exhibit variable metaethical stances or commitments.

**Mind-independence**

Another term for stance-independence.

**Modal operator scope ambiguity**
This occurs whenever the scope of a modal operator is ambiguous in a given sentence, such that it is subject to more than one possible interpretation. For instance, "Alex and Sam could both win the game" is ambiguous between an inclusive reading and an exclusive reading. On an exclusive reading, Alex *or* Sam could win, *but not both* (they could be playing a game that only allows one person to win). On an inclusive reading, Alex and Sam could both win *at the same time* (perhaps they are playing on the same team). Typically, such ambiguities are resolved by context. For instance, "Alex and Sam could both be tall" would typically be interpreted in inclusive terms, while "Alex and Sam could both win the gold medal for the 200m breaststroke at the next Olympics" would be exclusive.

**Moral antirealism**
The metaethical position that there are no stance-independent moral facts.

**Moral disagreement**
Any instance in which two people hold conflicting moral stances about a first-order moral claim. Such disagreements do not require either person to be aware of the other's position or for them to be actively engaged in a dispute. "Disagreement" simply refers to the difference in first-order moral stances, *not* to some social phenomenon, such as a debate or conflict.

**Moral domain**
Norms and evaluations that are distinctively *moral* and thus could in principle be distinguished from nonmoral norms. The moral domain could in principle be unified by one or more metanormative characteristics or its distinctness could be primitive (i.e., incapable of being reduced or defined by appeal to other terms or concepts).

**Moral grandstanding**
Strategically exploiting moral discourse to enhance social status (Grubbs et al., 2019).

**Moral incoherentism | incoherentism**
A form of moral antirealism advanced by Don Loeb (2008). Incoherentism is the metaethical position that (1) (the descriptive metaethical / linguistic thesis) folk metaethics is committed to conflicting metaethical presuppositions (e.g., both realism and antirealism) and that (2) (the metaphysical thesis) along with additional considerations, this suggests that there are no stance-independent moral facts (this glosses over the explanation and arguments for the view). For example, if a person asserting that "murder is wrong" is committed both to the notion that this claim refers to a stance-independent moral fact *and* that it *doesn't*, such a claim would be incoherent.

**Moral realism**
The claim that there are stance-independent moral facts. (Note: moral realism is sometimes described in other ways. It can be broader, including all positions which hold that there are moral facts, or narrower, by including additional theses such as the claim that we possess at least some moral knowledge, or that we are capable of knowing at least some moral facts. I do not employ these uses of *moral realism* here.)

**Moral relativism**

The metaethical position that moral claims are true or false relative to different standards of evaluation, such as the standards of individuals or groups.

**Naturalism | Metaethical naturalism**

The claim that moral facts are a type of *natural fact*. Natural facts are facts that are identical with, reducible to, or consistent with the facts described by the natural sciences.

**Noncognitivism**

The metaethical position that moral claims are not propositional, and are therefore incapable of being true or false.

**Non-naturalism | Metaethical non-naturalism**

The claim that moral facts cannot be reduced to *natural facts*.

**Nonmoral differences**

Differences that are not attributable to different moral standards. For instance, two people with the same moral standards could disagree about the moral status of abortion because they disagree about empirical details related to the viability of the fetus outside the womb or its capacity for pain or conscious experience.

**Nonphilosophers**

People without significant training in academic philosophy.

**Normative antirealism**

The claim that there are no stance-independent normative facts.

**Normative conflation**

Normative conflations are a distinctive form of conflation that occurs in studies designed to evaluate metaethical stances or commitments. Participants presented with stimuli intended to solicit metaethical stances or commitments may instead interpret stimuli to be asking questions about their normative moral standards. When this occurs, their responses may reflect their normative moral standards, rather than their metaethical standards. Such instances render any measures predicated on an intended interpretation invalid.

**Normative domain**

A general term for referring to different normative/evaluative domains, such as *moral norms*, *epistemic norms*, *prudential norms*, and so on.

**Normative entanglement**

The conflation between metaethical and normative considerations, or mistaken inferences about necessary conceptual relations between metaethics and normative ethics, e.g., someone may insist that

one cannot have normative moral standards unless they endorse moral realism, or that if you endorse moral antirealism, that you cannot object to someone harming you.

This occurs whenever metaethical claims are embedded (implicitly or explicitly) in normative claims or claims with potentially significant normative implications. Normative entanglement may be used as a rhetorical strategy to imply that people who disagree with the speaker hold immoral or undesirable stances and attitudes. For instance, a moral realist may ask a moral antirealist:

"Do you think it's objectively wrong to torture babies for fun?"

There is no way to respond to this question with a simple "yes" or "no" without expressing a stance towards moral realism *and* a direct or implied normative position on the moral status of torturing babies for fun. This is because this question embeds a metaethical claim inside a normative claim, and simultaneously asks about both:

(1) Metaethical claim: Torturing babies for fun is *stance-independently* morally wrong
(2) Normative claim: Torturing babies for fun is morally wrong

An antirealist cannot respond with a "yes," since this would be a concession to (1), which they do not accept. This only leaves them with the option to respond with "no." Technically, responding with a "no" only entails a rejection of (1). However, note that to respond with a "yes" simultaneously affirms both (1) and (2). To deny (1) may therefore be interpreted to *pragmatically imply* a rejection of (2).

Note, also, that in addition to expressing (1) and (2), this claim also carries pragmatic implications about a respondent's non-metaethical and non-propositional *attitudes* towards the action in question. To respond with a "no," may imply more than merely that one does not hold the normative moral stance that torturing babies is morally wrong, but that one does not personally object to torturing babies for fun, but is ambivalent, indifferent, or even in support of doing so. As a result, to respond with a "no" may give the impression that the antirealist has extremely deviant and antisocial standards, or may lack such standards altogether.

Moral realists often employ questions like these in order to imply that moral antirealists reject (2), and that they therefore have no objections to, and are not opposed to, various moral transgressions. The effectiveness of this ploy often revolves around how realistic and how serious the transgression is. As a result, using normative entanglement as a rhetorical strategy is typically most effective when the person employing the strategy appeals to an extremely serious moral transgression that causes widespread repugnance (such as abusing animals), but that isn't too outlandish or implausible to lose its rhetorical force (such as joining the dark side and using the Force to slaughter Wookiees).

**Normative realism**
The claim that there are stance-independent normative facts.

## Normative relativism

The moral stance that we should tolerate people and cultures with different moral standards. Not a genuine form of "relativism," but a concept frequently conflated with relativism as a metaethical position and closely connected to it.

## Objectivism

A common term often used interchangeably with *stance-independence*.

## Objectivism

A term commonly used in metaethics. Used interchangeably with "realism" and "stance-independence."

## Ordinary people

Nonphilosophers, i.e., people without significant training in academic philosophy.

## Philosophical commitment

An account of a speech act or behavior that is best explained in accordance with a particular philosophical position.

## Philosophical induction

Inducing ordinary people to engage in philosophy to such an extent that they are no longer ordinary people.

## Philosophical stance

A belief about the truth of a philosophical position.

## Reflexive thematic analysis (RTA)

A method of thematic analysis pioneered by Braun and Clarke which purports to be methodologically distinct from other forms of thematic analysis. This approach emphasizes the active, fluid, and evolving process of coding and rejects the notion of a single, objectively correct analysis. As such, RTA tends to reject the value of interrater reliability (FAQs, n.d.).

## Relation-designating account

Relation designating accounts are those for which the truth conditions of a given set of claims are contingent, and are made with respect to, a fixed standard of evaluation, such that all of those claims are stance-dependent but all relate to the same standard. As such, the truth conditions for a given statement *cannot* vary across speakers, appraisers, and so on. For instance, suppose you endorsed the theory that whether an action is morally right or wrong depends on the moral standards of the current ruler. All statements, such as "murder is morally wrong" and "kindness is morally good" relate to the standards of this ruler. As such, they are stance-dependently true or false, in that they are *made true* by the ruler's standards, but since there is only one current ruler with one set of moral standards, the moral rules cannot be true relative to one standard, but false relative to another.

**Relativism**
The claim that moral claims are true or false relative to the moral standards of different individuals or groups. Relativism may also be construed as a semantic thesis about the meaning of moral claims, namely, that they contain implicit indexicals that allow for the truth of moral claims to vary in accordance with the moral standards they relativized to.

**Response option**
One of a set of responses available to participants participating in a study, e.g., "true" and "false" would be standard response options to a true or false question.

**Scope ambiguity**
(See *modal operator scope ambiguity*)

**Spontaneous theorizing**
A phenomenon which occurs whenever the experimental context causes participants who held no determinate stance or commitment prior to participating in the study adopt or express a stance or commitment due to the experience of participating in the study itself.

**Stance**
(See *philosophical stance*)

**Stance-dependent**
A feature of the truthmaking properties of a given proposition. A proposition is stance-dependent insofar as it is made true by a stance.

**Stance-independent**
A feature of the truthmakers of a given proposition or set of propositions whereby that proposition (or those propositions) is *not* made true by real or hypothetical stances, such as the stances of individuals or groups.

**Stance-independent**
A feature of the truthmaking properties of a given proposition. A proposition is stance-independent insofar as it is made not true by a stance.

**Strategic metaethical pluralism**
A form of metaethical pluralism in which people adopt or express a particular metaethical stance or commitment to achieve social goals, such as persuasion or signaling desirable character traits. May explain apparent inconsistencies with people's metaethical views by redescribing them in rhetorical or argumentative terms.

**Subjectivism**
A form of moral relativism that relativizes moral truth to the moral standards of different individuals.

**Subjectivism**
A form of relativism which relativizes moral claims to the standards of individuals.

**Thematic analysis**
A method of analysis used in qualitative research that centers on the use of *themes* to organize and evaluate recurring patterns in a given body of data (FAQs, n.d.).

**Themes**
A distinct, recurring pattern within a given dataset (FAQs, n.d.).

**Training paradigms**
Studies that employ extensive instructions intended to teach participants about the concepts and distinctions captured by the psychological constructs the study seeks to measure. Such studies often include training exercises (such as practice questions) and comprehension checks, and may exclude participants that fail at these tasks.

**Unclear interpretation**
An interpretation of the given stimuli that is not sufficiently clear to be confidently classified as an instance of either an intended or unintended interpretation.

**Uniformism**
The claim that there is a single determinate account of folk metaethics.

**Uniformity-Determinacy assumption | UD assumption**
The assumption that there is a uniform and determinate account of folk metaethics. An idea originally put forward by Michael Gill, who attributed the UD assumption to much of 20th century descriptive metaethics.

**Unintended interpretation**
An interpretation of the given stimuli that does not appear consistent with researcher intent. Intended interpretations may be clear or unclear, so it is possible that a participant whose response was coded as unintended did interpret the stimuli as intended.

**Universalism**
The claim that there is a single correct moral standard (Ayars & Nichols, 2018). Alternatively, universalism could also be used to convey the notion that a given moral principle or rule applies to everyone, and is thereby concerned with the *scope* of the moral standard in question (Goodwin & Darley, 2008).

**WEIRD**
An acronym proposed by Henrich and colleagues to refer to the standard research subjects in most psychological research. WEIRD stands for **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic.

# APPENDIX B:
## Bibliography of Folk Metaethics Studies

*The following bibliography provides a complete list of all published studies which address folk moral realism and antirealism in some way in chronological order. This list was updated on February 3, 2023, and consists only of those studies that, to the best of my knowledge, provide empirical data. It is possible I've missed some studies.*

## Bibliography

Trainer, F. E. (1983). Ethical objectivism-subjectivism: a neglected dimension in the study of moral thought. *Journal of Moral Education, 12*(3), 192-207.

Nichols, S., & Folds-Bennett, T. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition, 90*(2), B23-B32.

Nichols, S. (2004). After objectivity: An empirical study of moral judgment. *Philosophical Psychology, 17*(1), 3-26.

Wainryb, C., Shaw, L. A., Langley, M., Cottam, K., & Lewis, R. (2004). Children's thinking about diversity of belief in the early school years: Judgments of relativism, tolerance, and disagreeing persons. *Child Development, 75*(3), 687-703.

Cova, F., Ravat, J. (2008). Common sense and moral objectivism: Global or local? A case of experimental philosophy. *KLESIS - Revue philosophique, 9*, 180-202.

Feltz, A., & Cokely, E. T. (2008). The fragmented folk: more evidence of stable individual differences in moral judgments and folk intuitions. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1771-1776). Austin, TX: Cognitive science society.

Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106*(3), 1339-1366.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language, 26*(4), 482-505.

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others?. *Journal of Experimental Social Psychology, 48*(1), 250-256.

Quintelier, K. J., & Fessler, D. M. (2012). Varying versions of moral relativism: The philosophy and psychology of normative relativism. *Biology & Philosophy, 27*(1), 95-113.

Young, L., & Durwin, A. J. (2013a). Moral realism as moral motivation: The impact of meta-ethics on everyday decision-making. *Journal of Experimental Social Psychology, 49*(2), 302-306.

Rai, T. S., & Holyoak, K. J. (2013). Exposure to moral relativism compromises moral behavior. *Journal of Experimental Social Psychology, 49*(6), 995-1001.

Quintelier, K. J., De Smet, D., & Fessler, D. M. (2013). The moral universalism-relativism debate. *KLESIS–Revue Philosophique, 27*, 211-262.

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2013). The development of reasoning about beliefs: Fact, preference, and ideology. *Journal of Experimental Social Psychology, 49*(3), 559-565.

Wright, J. C., Grandjean, P. T., & McWhite, C. B. (2013). The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology, 26*(3), 336-361.

Quintelier, K., De Smet, D., & Fessler, D. (2014). Agent versus appraiser moral relativism: an exploratory study. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 209-230). London, UK: Bloomsbury Academic.

Uttich, K., Tsai, G., & Lombrozo, T. (2014). Exploring metaethical commitments: Moral objectivity and moral progress. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 108-208). London, UK: Bloomsbury.

Wright, J. C., McWhite, C. B., & Grandjean, P. T. (2014). The cognitive mechanisms of intolerance: Do our meta-ethical commitments matter?. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 1, pp. 28-61). New York, NY: Oxford University Press.

Beebe, J. R. (2014). How different kinds of disagreement impact folk metaethical judgments. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 167-187). London, UK: Bloomsbury.

Beebe, J. (2015). The empirical study of folk metaethics. Etyka, 50, 11-28.

Beebe, J., Qiaoan, R., Wysocki, T., & Endara, M. A. (2015). Moral objectivism in cross-cultural perspective. *Journal of Cognition and Culture, 15*(3-4), 386-401.

Yilmaz, O., & Bahcekapili, H. G. (2015). Without God, everything is permitted? The reciprocal influence of religious and meta-ethical beliefs. *Journal of Experimental Social Psychology, 58*, 95-100.

Sarkissian, H. (2016). Aspects of Folk Morality. In J. Sytsma, & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 212-224). West Sussex, UK: Wiley-Blackwell.

Beebe, J. R., & Sackris, D. (2016). Moral objectivism across the lifespan. *Philosophical Psychology, 29*(6), 912-929.

Heiphetz, L., & Young, L. L. (2017). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition, 167*, 78-90.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cognitive science, 41*(4), 1119-1134.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2017). Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General*, *146*(11), 1586-1605.

Schmidt, M. F., Gonzalez-Cabrera, I., & Tomasello, M. (2017). Children's developing metaethical judgments. *Journal of Experimental Child Psychology*, *164*, 163-177.

Khoo, J., & Knobe, J. (2018). Moral disagreement and moral semantics. *Noûs, 52*(1), 109-143.

Wright, J. C. (2018). The fact and function of meta-ethical pluralism: Exploring the evidence. In T. Lombrozo, J. Knobe, S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 2, pp. 119-150). Oxford, UK: Oxford University Press.

Yilmaz, O., & Bahçekapili, H. G. (2018). Meta-ethics and the mortality: Mortality salience leads people to adopt a less subjectivist morality. *Cognition, 179*, 171-177.

Colebrook, R. T. (2018). Toward a science of morals. [Doctoral dissertation, City University of New York.] CUNY Academic Works. https://academicworks.cuny.edu/gc_etds/2648

Ayars, A., & Nichols, S. (2020). Rational learners and metaethics: Universalism, relativism, and evidence from consensus. *Mind & Language, 35*(1), 67-89.

Johnson, S. G., Rodrigues, M., & Tuckett, D. (2021). Moral tribalism and its discontents: How intuitive theories of ethics shape consumers' deference to experts. *Journal of Behavioral Decision Making, 34*(1), 47-65.

Rose, D., & Nichols, S. (2019). From punishment to universalism. *Mind & Language, 34*(1), 59-72.

Hopster, J.K.G. (2019). Moral objectivity: Origins and foundations (Quaestiones Infinitae, volume 119) [Doctoral dissertation, Utrecht University]. Utrecht University Repository. https://dspace.library.uu.nl/handle/1874/374619

Sarkissian, H., & Phelan, M. (2019). Moral objectivism and a punishing God. *Journal of Experimental Social Psychology*, 80, 1-7.

Becklloyd, D. and Sytsma, J.(2019). *Simulating metaethics: Consensus and the independence of moral beliefs*. PhilSci Archive. http://philsci-archive.pitt.edu/16461/

Zijlstra, L. (2019). Folk moral objectivism and its measurement. *Journal of Experimental Social Psychology, 84*, 103807. https://doi.org/10.1016/j.jesp.2019.04.005

Viciana, H., Hannikainen, I. R., & Gaitan Torres, A. (2019). The dual nature of partisan prejudice: Morality and identity in a multiparty system. *PloS One*, *14*(7), e0219509. https://doi.org/10.1371/journal.pone.0219509

Collier-Spruel, L., Hawkins, A., Jayawickreme, E., Fleeson, W., & Furr, R. M. (2019). Relativism or tolerance? Defining, assessing, connecting, and distinguishing two moral personality features with prominent roles in modern societies. *Journal of Personality, 87*(6), 1170-1188.

Davis, T. (2021). Beyond objectivism: new methods for studying metaethical intuitions. *Philosophical Psychology, 34*(1), 125-153.

Pölzler, T., & Cole Wright, J. C. (2020a). An empirical argument against moral non-cognitivism. *Inquiry*, 1-29. https://doi.org/10.1080/0020174X.2020.1798280

Pölzler, T., & Wright, J. C. (2020b). Anti-realist pluralism: A new approach to folk metaethics. *Review of Philosophy and Psychology, 11*(1), 53-82.

Stanley, M. L., Marsh, E. J., & Kay, A. C. (2020). Structure-seeking as a psychological antecedent of beliefs about morality. *Journal of Experimental Psychology: General, 149*(10), 1908-1918.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2020). Theory of Mind network activity is associated with metaethical judgment: An item analysis. *Neuropsychologia, 143*, 107475. https://doi.org/10.1016/j.neuropsychologia.2020.107475

Yilmaz, O., Bahcekapili, H. G., Harma, M., & Sevi, B. (2020). Intergroup tolerance leads to subjective morality, which in turn is associated with (but does not lead to) reduced religiosity. *Archive for the Psychology of Religion, 42*(2), 232-243.

Sousa, P., Allard, A., Piazza, J., & Goodwin, G. P. (2021). Folk moral objectivism: The case of harmful actions. *Frontiers in Psychology, 12*, Article 2776. https://doi.org/10.3389/fpsyg.2021.638515

Sytsma, J., Muldoon, R., & Nichols, S. (2021). The meta-wisdom of crowds. *Synthese, 199*(3), 11051-11074.

Viciana, H., Hannikainen, I. R., & Rodríguez-Arias, D. (2021). Absolutely right and relatively good: Consequentialists see bioethical disagreement in a relativist light. *AJOB Empirical Bioethics, 12*(3), 190-205.

Wagner, J. M., Pölzler, T., & Wright, J. C. (2021). Implicit Metaethical Intuitions: Validating and Employing a New IAT Procedure. *Review of Philosophy and Psychology*, 1-31. https://doi.org/10.1007/s13164-021-00572-3

Zijlstra, L. (2021). Are people implicitly moral objectivists?. *Review of Philosophy and Psychology*, 1-19. https://doi.org/10.1007/s13164-021-00593-y

Cova, F. (2022, June 15). The need for measure calibration in experimental philosophy: the case of measures of folk objectivism. https://doi.org/10.31234/osf.io/kcft9

Lewry, C., & Lombrozo, T. (2022). Ethical Explanations. In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44), 42-48. https://escholarship.org/uc/item/2rv1r2s0

Chvaja, R., Horský, J., Lang, M., & Kundt, R. (2022). Positive Association Between Ritual Performance and Perceived Objectivity of Moral Norms. *The International Journal for the Psychology of Religion*, 1-21. https://doi.org/10.1080/10508619.2022.2121454

Pölzler, T., Zijlstra, L., & Dijkstra, J. (2022). Moral progress, knowledge and error: Do people believe in moral objectivity?. Philosophical Psychology, 1-37.

Rodríguez-Pérez, A., Rodríguez-Torres, R., Betancor, V., Chen-Xia, X. J., & Rodríguez-Gómez, L. (2022). Are Civility Norms Morality Norms' Little Sister? The Truth Value That Lay Thinking Associates with Civility and Morality Social Norms. Social Sciences, 11(12), 568.

Wright, J. C., & Pölzler, T. (2022). Should morality be abolished? An empirical challenge to the argument from intolerance. *Philosophical Psychology, 35*(3), 350-385.

# APPENDIX C:
## Metaethics scales

### Relativism Subscale of the Ethics Position Questionnaire (EPQ)
*Forsyth (1980)*

**Source**

Forsyth, D. R. (1980). A taxonomy of ethical ideologies. *Journal of Personality and Social psychology*, *39*(1), 175-184.

**Scale**

A = strongly disagree, B = disagree, C = no opinion or neutral, D = agree, E = strongly agree

1.  There are no ethical principles that are so important that they should be part of any code of ethics.
2.  What is ethical varies from one situation and society to another.
3.  Moral standards should be seen as being individualistic; what one person considers to be moral may be judged to be immoral by another person.
4.  Different types of moralities cannot be compared as to "rightness."
5.  Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.
6.  Moral standards are simply personal rules which indicate how a person should behave, and are not to be applied in making judgments of others.
7.  Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their own individual codes.
8.  Rigidly codifying an ethical position that prevents certain types of actions could stand in the way of better human relations and adjustment.
9.  No rule concerning lying can be formulated; whether a lie is permissible or not totally depends on the situation.
10. Whether a lie is judged to be moral or immoral depends upon the circumstances surrounding the action.

### Objectivism/Subjectivism scale (TOS)
*Trainer (1983)*

**Source**

Trainer, F. E. (1983). Ethical objectivism-subjectivism: a neglected dimension in the study of moral thought. *Journal of Moral Education*, *12*(3), 192-207.

**Scale**

Dichotomous choice between realist and antirealist response.
A = Realist response, B = Antirealist response

1A. You can say without any doubt in some situations that something is right or wrong, and you say that people who don't agree with you are wrong.
1B. It isn't possible for anyone to be really sure what is right or wrong. You can only say that others with different opinions are wrong.

2A. Some things are wrong no matter what anyone thinks and people should be told this if they don't know.
2B. People should be free to work out for themselves what they think is right or wrong.

3A. You can be quite sure that some of your basic moral principles are right.
3B. You can't be sure that your basic moral principles are right. Even 'It is wrong to kill' is just an assumption or an item of belief on your part.

4A. Some values or actions are objectively right, they are right in fact, whether or not individuals think so.
4B. All judgements about right and wrong state nothing more objective than the ideas or attitudes of individuals.

5A. There are fixed and eternal moral laws of nature.
5B. All moral laws can be made, altered or rejected by man.

6A. Moral laws can't be changed by human beings any more than the laws of physics can be.
6B. Moral laws can be changed by man because they are only man-made rules in the first place.

7A. Human beings can only discover moral laws; we can't make them. Just as we can't make up true laws of science to suit ourselves neither can we make up true moral laws.
7B. Moral laws are made up by humans as ways they choose for regulating behaviour.

8A. In any problem to do with justice there is in fact a just solution, even though no one may be able to say what it is.
8B. You can only talk about what you or others would accept as a just or fair solution, not about what is in fact the just solution.

9A. There would still be fundamental human rights even if no one on earth thought there were.
9B. Rights are entirely created by man.

10A. lt is possible to know that your basic moral principles are the right criteria for evaluating things.
10B. Your basic moral principles can only be your best guess at the criteria for evaluating things; you can never know whether yours are the right criteria.

11A. My moral principles are not just things I prefer. They are things that ought to be preferred because they are morally Important.

11B. My moral principles can't be any more than the values I choose or prefer to live by; others choose other values and I can't say my choices are in fact the right ones.

12A. lt is possible to claim that your moral principles are better than those of some other people, such as a thief or a sadist.

12B. You can't claim that your moral principles are better than anyone elses.

13A. Some things are in fact morally better than others no matter what anyone thinks or prefers.

13B. Our basic moral values are only our personal preferences. You might prefer kindness to cruelty but you can't say it is in fact morally better.

14A. We can say much more than that we do not like this. We could say cruelty to animals is in fact morally bad and should not be done whether or not anyone likes to do it.

14B. If we found someone who likes to torture animals for the fun of it we could not say this was a morally bad thing to do. We could only say that his interest disgusts us and that we wish he would not practise it.

15A. Honesty is in fact morally better than cheating. There is more involved here than my liking for one and my dislike of another.

15B. I can't say that even my most important values are objectively morally better than any other values at all. There is, for example, no objective difference in the moral quality of honesty and cheating. All I can say is that I like one and not the other.

16A. Whether or not you disapprove and whether or not men make laws against it, murder is in fact morally wrong.

16B. I cannot condemn even murder as being a morally bad action, because it breaks no moral law. There are no moral laws of nature, there are only laws men make up. All I can say are things like, I don't approve of murder and most people don't so they make laws against it.

17A. You can say more than that you approve of the first person. There is a real moral difference in the quality of their actions and this does not depend on what people think or approve.

17B. If one person works hard for charity while another loafs, and another works hard for a terrorist organization, you can't say there is any real moral difference in the quality of these actions, or that the first is in fact morally better than the other. You can only say you approve of the first and disapprove of the other.

## 3-item face valid metaethics scale (YB3)
*Yilmaz & Bahçekapili (2015b)*

**Source**

Yilmaz, O., & Bahçekapili, H. G. (2015b). Without God, everything is permitted? The reciprocal influence of religious and meta-ethical beliefs. *Journal of Experimental Social Psychology*, *58*(Suppl. A), 95-100.

**Scale**

(Not specified)

1. There are no ethical principles that are so important that they should be a part of any code of ethics.
2. Different types of moralities cannot be compared as to "rightness."
3. Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.

## New Meta-Ethics Questionnaire (NMQ)
*Yilmaz & Bahçekapili (2015a)*

**Source**

Yilmaz, O., & Bahçekapili, H. G. (2015a). Without God, everything is permitted? The reciprocal influence of religious and meta-ethical beliefs. *Journal of Experimental Social Psychology, 58*, 95-100.

**Scale**

1 = strongly disagree, 7 = strongly agree

1. What is moral varies on the basis of context and society.
2. Moral standards are personal, therefore something morally acceptable to one person might be immoral for another person.
3. Since moral rules are not absolute, no definite judgments about them are possible.
4. Different cultures adopt different values and no moral law is right or wrong in an absolute sense.
5. We can agree on 'what is moral for everyone' because what is moral and immoral is self-evident.
6. If morality were to differ from person to person, it would be impossible for people to live together.
7. Since the moral laws I believe in are universally true, they can be applied to everyone in the world regardless of culture, race or religion.
8. If a moral law is right and good for others, it is also right and good for us.

### 3-Item Moral Objectivism Scale (MO3)
*Sarkissian & Phelan (2019)*

**Source**

Sarkissian, H., & Phelan, M. (2019). Moral objectivism and a punishing God. *Journal of Experimental Social Psychology*, *80*, 1-7.

**Scale**

(Not specified)

1. There exists a single moral code that is applicable to everyone, regardless of any individual person's beliefs or cultural identity.
2. If two people really disagree about a particular moral problem then at most one of them can be correct, since moral problems cannot have multiple correct answers.
3. It is possible to compare different cultures by a single, universal standard of moral rightness.


### Moral Relativism Scale (MRS)
*Collier-Spruel et al. (2019)*

**Source**

Collier-Spruel, L., Hawkins, A., Jayawickreme, E., Fleeson, W., & Furr, R. M. (2019). Relativism or tolerance? Defining, assessing, connecting, and distinguishing two moral personality features with prominent roles in modern societies. *Journal of Personality*, *87*(6), 1170-1188.

**Scale**

[1 = Strongly Disagree, 5 = Strongly Agree]

1. Different people can have opposing views on what is moral and immoral without anyone being wrong
2. People can disagree on what is morally right without anyone being wrong
3. Two different cultures could have dissimilar moral rules and both be "right"
4. One's own culture determines whether that person's actions are "right" or "wrong"
5. The viewpoint of one's culture determines whether their actions are morally right
6. There is a moral standard that all actions should be held to, even if cultures disagree [reverse coded]
7. Each person is the final authority on whether his or her actions really are morally correct
8. An action is only morally wrong if a person believes it is morally wrong
9. There are moral rules that apply to everyone regardless of personal beliefs [reverse coded]
10. The same moral standards should be followed by people from all cultures [reverse coded]

**Folk Moral Objectivism Scale (FMO)**
*Zijlstra (2019)*

**Source**

**Scale**

[1 = Strongly disagree, 6 = Strongly agree]

*No truth subscale*

1. Other than what people believe, are brought up to believe, or want to believe about it, there are no facts about what is morally right and wrong
2. All ideas about what is morally right and morally wrong are products of individuals, cultures, and communities and nothing more
3. What people believe to be morally right and wrong are merely social conventions that could have been different
4. It is an illusion to think that anything is really morally true or false

*Relativism subscsale*

5. When two people have opposing beliefs about a moral issue, it is not necessarily the case that either or both are wrong
6. There is not one but many different answers to the question of what is morally right and wrong and these can be equally correct
7. What is ultimately morally right and wrong is different for people with different moral views and from different cultures and societies
8. What is morally right and wrong is relative to the moral beliefs of an individual, culture, or society

*Universalism subscale*

9. What is ultimately morally right or wrong is the same for all people at all times and places
10. Although people or cultures sometimes ignore moral concerns, moral norms apply anywhere and everywhere
11. What is morally right and wrong for me here and now is also morally right and wrong for people elsewhere, even for people living in different countries and part of different cultures
12. Despite the diversity of moral views between individuals, cultures, and societies, there are moral norms that should apply universally

*Absolutism subscale*

13. Although people disagree about what is morally right and wrong, I believe in the existence of specific moral principles that can settle any moral disagreement

14. Certain actions are morally wrong and they remain morally wrong even in the rare case that no one believes so

15. There are absolute moral rules that apply to all people, including those who do not acknowledge these principles

16. There is, in all circumstances, one correct answer about what is the morally right thing to do

*Divine Command Theory (DCT) subscale*

17. The correct answer to any moral issue can be found in a sacred book or text (for example, the Bible, the Qur'an, the Torah, or another)

18. The only actions that are ultimately morally right or wrong are those actions that God prescribes

19. God is the only true source of knowledge about what is morally right or wrong

20. Without the existence of God, nothing is truly morally right or wrong

# Objectivity of Morality Scale (JRT5)
*Johnson, Rodrigues, & Tuckett (2020)*

**Source**

Johnson, S. G., Rodrigues, M., & Tuckett, D. (2021). Moral tribalism and its discontents: How intuitive theories of ethics shape consumers' deference to experts. *Journal of Behavioral Decision Making, 34*(1), 47-65.

**Scale**

(Not specified)

*Normativity subscale*

1. Every good person on earth, regardless of culture, holds these beliefs.
2. The truth of these beliefs is self-evident.
3. A society could not survive without its citizens holding these beliefs

*Subjectivity subscale*

4. If someone strongly disagreed with you about one of these beliefs, it is possible that neither you nor the other person are mistaken [reverse coded]
5. There are no clearly true or false answers to these questions. [reverse coded]

**Open Science Statement**

Data, code, and materials can be found on the Open Science Network at: https://osf.io/2afky/

**Qualitative Comprehension Paradigms**
(Open response)

- ➢ Study 1A: Reanalysis of Goodwin & Darley (2008) Experiment 1
- ➢ Study 1B: Reanalysis of Goodwin & Darley (2008) Experiment 2
- ➢ Study 1C: New test of disagreement paradigm
- ➢ Study 2: Interpreting other people's responses to metaethics stimuli
- ➢ Study 3: Charity & noncognitivism
- ➢ Study 4: Direct realism and relativism
- ➢ Study 5: Interpretation of the Moral Relativism Scale (MRS)
- ➢ Study 6: YB3 and New-Metaethics Questionnaire (NMQ)
- ➢ Study 7: Folk Moral Objectivism Scale (FMO)

**Study 1A & 1B:**

These studies were a reanalysis of the studies conducted by Goodwin and Darley (2008) and will not be presented here.

**Study 1C:**
**New test of the disagreement paradigm**

**Study 1C**

When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?

○ They can both be correct

○ At least one of them must be incorrect

Please briefly explain why you chose this response.

| |
|---|

## Study 2:
## Interpreting other people's responses to metaethics stimuli

*(Note: These materials were included as part of a larger study. I have extracted the relevant materials from the study, but the actual formatting of the study was somewhat different. It included other measures prior to participants being presented with open response questions, so these questions did not immediately follow the measures, as they appear here.)*

**Study 2A**

The respondent was asked the following question:

**When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?**

**John:**
**"When people disagree about a particular moral issue there can be at most only one correct answer".**

Think about the person who answered this question.

In your own words, what do you think the respondent means in the statement above?

**Study 2B**

The respondent was asked the following question:

**When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?**

**John:**
**"When people disagree about a particular moral issue each can be correct according to their own moral standards".**

Think about the person who answered this question.

In your own words, what do you think the respondent means in the statement above?

```




```

## Study 3:
## Charity & noncognitivism

*(Note: These materials were included as part of a larger study. I have extracted the relevant materials from the study, but the actual formatting of the study was somewhat different. It included other measures prior to participants being presented with open response questions, so these questions did not immediately follow the measures, as they appear here.)*

### Study 3A

In your own words, what does it mean to say that "it is a fact that some charities do more good than others, not a matter of personal beliefs or values"?

```


```

### Study 3B

In your own words, what does it mean to say that "the truth about which charities do the most good depends on the beliefs and values of each individual"?

```


```

### Study 3C

In your own words, what does it mean to say that "there is no fact of the matter about which charities do the most good"?

```
                                    ┌────────────────────────────────────────────────────┐
                                    │                                                    │
                                    │                                                    │
                                    └────────────────────────────────────────────────────┘
```

## Study 4:
## Direct realism and relativism

**Study 4A**

**[Objectivism | Realism | Murder]**

In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is objective?

```
┌────────────────────────────────────────────────────┐
│                                                    │
│                                                    │
│                                                    │
│                                                    │
│                                                    │
│                                                    │
└────────────────────────────────────────────────────┘
```

**[Concrete | Relativism | Murder]**

In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is relative?

```
┌────────────────────────────────────────────────────┐
│                                                    │
│                                                    │
│                                                    │
│                                                    │
│                                                    │
│                                                    │
└────────────────────────────────────────────────────┘
```

**[Concrete | Realism | Abortion]**

In your own words, what does it mean to say that the truth of the moral claim "abortion is morally wrong" is objective?

```
┌────────────────────────────────────────────────────┐
│                                                    │
└────────────────────────────────────────────────────┘
```

**[Concrete | Relativism | Abortion]**

In your own words, what does it mean to say that the truth of the moral claim "abortion is morally wrong" is relative?

**Study 4B**

**[Abstract | Realism]**

In your own words, what does it mean to say that moral truth is objective?

**[Abstract | Relativism]**

In your own words, what does it mean to say that moral truth is relative?

**Study 5:**
**Interpretation of the Moral Relativism Scale (MRS)**

*[instructions]*
In this study you will be asked to rate how much you agree or disagree with a statement. Then you will be asked a few questions about the statement.

{--------------------------------Page break--------------------------------}

*[Participants randomly assigned to one of the following statements for the measures below or to items on the moral tolerance scale (MTS) from Collier-Spruel et al. (2019) which isn't shown]*

*[Relativism 1]*
**Different people can have opposing views on what is moral and immoral without anyone being wrong.**

*[Relativism 2]*
**People can disagree on what is morally right without anyone being wrong.**

*[Relativism 3]*
**Two different cultures could have dissimilar moral rules and both be "right."**

*[Relativism 4]*
**One's own culture determines whether that person's actions are "right" or "wrong."**

*[Relativism 5]*
**The viewpoint of one's culture determines whether their actions are morally right.**

*[Relativism 6] [Reverse-coded]*
**There is a moral standard that all actions should be held to, even if cultures disagree.**

*[Relativism 7]*
**Each person is the final authority on whether his or her actions really are morally correct.**

*[Relativism 8]*

380

**An action is only morally wrong if a person believes it is morally wrong.**

*[Relativism 9] [Reverse-coded]*
**There are moral rules that apply to everyone regardless of personal beliefs.**

*[Relativism 10] [Reverse-coded]*
**The same moral standards should be followed by people from all cultures.**

*[Agree]*
Please rate how much you agree or disagree with the following statement.

**[Statement]**

[1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree]

*[Why]*
Please explain why you chose this response.

```
                                                                    
                                                                    
                                                                    
                                                                    
```

*[Explain]*
In your own words, please explain what this statement means:

**[Statement]**

```
                                                                    
                                                                    
                                                                    
                                                                    
```

**Study 6:**
**YB3 and New-Metaethics Questionnaire (NMQ)**

*(Note: Participants were randomly assigned to items from the EPQ, NMQ, YB3, or FMO in one study, which was all collected at the same time, so the items collected in study 6 and 7 were conducted as part of the same study. Materials for the EPQ are not shown here)*

*[Participants randomly assigned to items from the YB3 or NMQ]*

*[YB3 1]*

**There are no ethical principles that are so important that they should be a part of any code of ethics.**

*[YB3 2]*

**Different types of moralities cannot be compared as to "rightness."**

*[YB3 3]*

**Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.**

*(Note: Items NMQ 2 and NMQ 5 were not run because they also appear on the EPQ, so rerunning them would be redundant)*

*[NMQ 1]*

**Since what is moral varies on the basis of context and society, there is no one true morality.**

*[NMQ 3]*

**Since moral rules are not true or false in an absolute sense, moral debates are bound to remain inconclusive.**

*[NMQ 4]*

**Different cultures may adopt different values and thus it is impossible to compare cultures on the basis of an objective standard.**

*[NMQ 6]*

**What makes it possible for people to live together in harmony is the fact that fundamental moral rules do not differ from person to person.**

*[NMQ 7]*

**Since moral laws are universally true, they can be applied to everyone in the world regardless of culture, race or religion.**

*[NMQ 8]*

**Fundamental moral principles are universally valid; therefore they can be transferred from one society to another without difficulty.**

*[Agree]*
Please rate how much you agree or disagree with the following statement.

**[Statement]**

[1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree]

*[Why]*
Please explain why you chose this response.

```
┌────────────────────────────────────────────┐
│                                            │
│                                            │
│                                            │
│                                            │
└────────────────────────────────────────────┘
```

*[Explain]*
In your own words, please explain what this statement means:

**[Statement]**

```
┌────────────────────────────────────────────┐
│                                            │
│                                            │
│                                            │
│                                            │
└────────────────────────────────────────────┘
```

**Study 7:**
**Folk Moral Objectivism Scale (FMO)**

*(Note: Participants were randomly assigned to items from the EPQ, NMQ, YB3, or FMO in one study, which was all collected at the same time, so the items collected in study 6 and 7 were conducted as part of the same study. Materials for the EPQ are not shown here)*

*[Participants randomly assigned to one of the following items from the FMO]*

*[FMO Relativism 1]*
**When two people have opposing beliefs about a moral issue, it is not necessarily the case that either or both are wrong.**

*[FMO Relativism 2]*
**There is not one but many different answers to the question of what is morally right and wrong and these can be equally correct**

*[FMO Relativism 3]*
**What is ultimately morally right and wrong is different for people with different moral views and from different cultures and societies**

*[FMO Relativism 4]*

What is morally right and wrong is relative to the moral beliefs of an individual, culture, or society

*[FMO Universalism 1]*
What is ultimately morally right or wrong is the same for all people at all times and places

*[FMO Universalism 2]*
Although people or cultures sometimes ignore moral concerns, moral norms apply anywhere and everywhere

*[FMO Universalism 3]*
What is morally right and wrong for me here and now is also morally right and wrong for people elsewhere, even for people living in different countries and part of different cultures

*[FMO Universalism 4]*
Despite the diversity of moral views between individuals, cultures, and societies, there are moral norms that should apply universally

*[FMO Absolutism 1]*
Although people disagree about what is morally right and wrong, I believe in the existence of specific moral principles that can settle any moral disagreement

*[FMO Absolutism 2]*
Certain actions are morally wrong and they remain morally wrong even in the rare case that no one believes so

*[FMO Absolutism 3]*
There are absolute moral rules that apply to all people, including those who do not acknowledge these principles

*[FMO Absolutism 4]*
There is, in all circumstances, one correct answer about what is the morally right thing to do

*[FMO DCT 1]*
The correct answer to any moral issue can be found in a sacred book or text (for example, the Bible, the Qur'an, the Torah, or another)

*[FMO DCT 2]*
The only actions that are ultimately morally right or wrong are those actions that God prescribes

*[FMO DCT 3]*
God is the only true source of knowledge about what is morally right or wrong

*[FMO DCT 4]*
**Without the existence of God, nothing is truly morally right or wrong**

*[Agree]*
Please rate how much you agree or disagree with the following statement.

**[Statement]**

[1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree]

*[Why]*
Please explain why you chose this response.

```
┌─────────────────────────────────────────────┐
│                                             │
│                                             │
│                                             │
│                                             │
└─────────────────────────────────────────────┘
```

*[Explain]*
In your own words, please explain what this statement means:

**[Statement]**

```
┌─────────────────────────────────────────────┐
│                                             │
│                                             │
│                                             │
│                                             │
└─────────────────────────────────────────────┘
```

**Quantitative Comprehension Paradigms**
(Multiple choice, Checkbox, Likert scale)

➢ Study 1: Moral Relativism Scale (MRS)
➢ Study 2: Bush & Moss
➢ Study 3: Fisher et al. Explicit
➢ Study 4: Young & Durwin
➢ Study 5: Checkbox
➢ Study 6: Likert Scale

**Study 1:**
**Moral Relativism Scale (MRS)**

**INSTRUCTIONS**

On the next page, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.


{--------------------------------Page break--------------------------------}


**[conditions] [randomized, between subjects, one per participant]**
Below, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [antirealism] [disagreement | relativism] [MRS]*
**"Different people can have opposing views on what is moral and immoral without anyone being wrong."**

- **Since moral truth is determined by what each individual believes, people who hold conflicting moral views can both be correct**
- Different people can have conflicting views about what is right or wrong without either of them having made an unjustified mistake
- People with different moral views can each be correct about some aspects of a moral issue without either being completely incorrect
- When people hold conflicting moral views sometimes there is no way to know which view is correct or incorrect
- People with opposing moral views cannot be wrong because it is not possible for people to be right or wrong about their moral views
- People have a right to their moral views even if other people hold opposing moral views



*[condition 2] [antirealism] [disagreement | dependent truth] [MRS]*
 **"The viewpoint of one's culture determines whether their actions are morally right."**

- **The truth about whether an action is morally right can only be judged according to the standards of different cultures**

386

- People from different cultures can be justified in holding different moral beliefs
- Each culture adapts its views about what is morally right to fit with their traditions and meet local needs
- Since each culture has different views about what is morally right, a person's actions will be judged as moral by some cultures and immoral by other cultures
- Different cultures have the right to hold different moral beliefs, so it isn't acceptable to impose our moral values on them
- The attitudes of each person's culture influence whether their actions are judged as moral or immoral

*[condition 3] [realism | universalism]*
**"There are moral rules that apply to everyone regardless of personal beliefs."**

- **The truth of some moral rules is not determined by what people believe is morally right or wrong**
- Some moral rules apply to all people regardless of the specific circumstances they are in
- Everyone should be held to the same moral standards even if they do not agree with those standards
- People should be held accountable for breaking some moral rules even if they believe their actions were justified
- We should judge everyone according to a strict moral code even if we believe that there ought to be an exception in some cases
- Some actions are always right or wrong regardless of the intentions of the person committing the action

*[condition 4] [realism | universal/normative]*
**"The same moral standards should be followed by people from all cultures."**

- **There is a correct set of moral standards that does not depend on what cultures think is right or wrong**
- We would be better off if people from every culture agreed to follow the same moral standards
- All cultures should be held to the same moral standards, so we should not tolerate cultures with moral beliefs that conflict with those standards
- People from all cultures should adhere to the same moral standards even if they do not agree with those standards
- We should not consider a person's cultural background when we judge their actions as morally right or wrong
- No matter what culture you come from, there are certain things everyone should know are right or wrong

**Study 2:**
**Fisher et al. Explicit**

*[instructions]*

**Instructions**

On the next page, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

*We are not asking you to choose the option that you agree with the most.* We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.


{---------------------------------Page break--------------------------------}


**[conditions] [random assignment] [one condition each]**

Below, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [realism] [explicit] [Fisher]*
**"Moral truth is objective, so there is an objectively true answer to whether violence should be seen as an appropriate response in conflicts."**

- **The truth about whether violence is an appropriate response in conflicts does not depend on the standards or values of individuals or cultures**
- Whether violence is an appropriate response in conflicts depends on the circumstances
- We should see violence as appropriate in all situations or no situations, rather than thinking it is appropriate in some situations but not others
- The truth about whether violence should be seen as an appropriate response in conflicts applies to everyone, regardless of their situation
- We can appeal to facts and data instead of our subjective opinions to determine whether violence is an appropriate response in conflicts
- Although people can be biased in judging actions as right or wrong, if we can set aside our biases we can judge in each case whether violence is an appropriate response to conflict

*[condition 2] [relativism] [explicit] [Fisher]*
**"Moral truth is relative, so the truth about whether violence should be seen as an appropriate response in conflicts is relative."**

- **Claims about whether violence is an appropriate response to conflicts can only be true or false according to different moral frameworks**
- Different people and societies have different perspectives, so there more than one point of view about whether violence should be seen as an appropriate response in conflicts,

- There is no way to be sure whether violence is an appropriate response in conflicts because we don't have a single agreed-upon method for measuring the outcomes of violent responses
- Whether violence should be seen as an appropriate response in conflicts depends on the context, so it may be appropriate in some cases but not others
- Since each person and society has different views about whether violence is an appropriate response in conflicts is relative, the use of violence will be judged as moral by some people and immoral by other people
- There are no moral truths about whether violence should be seen as an appropriate response in conflicts, just different people's feelings about what's right or wrong

{--------------------------------Page break--------------------------------}

*[comprehension check]*

In response to the question on the previous page, which of the following were you doing?

- **Selecting the single statement which I think best matches the meaning of the bolded sentence**
- Selecting the single statement that I most agree with
- Selecting every statement that I agree with
- Selecting every statement which I think best matches the meaning of the bolded sentence

{--------------------------------Page break--------------------------------}

*[age]*

What is your age?

_____

*[gender]*

What is your gender?

○ Male

○ Female

○ Other

**Study 3:**
**Young & Durwin**

*[instructions]*

**<u>Instructions</u>**

389

On the next page, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

*We are not asking you to choose the option that you agree with the most.* We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.


{--------------------------------Page break--------------------------------}

**[conditions] [random assignment] [one condition each]**

Below, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [realism] [Y&D]*
**"Some things are just morally right or wrong, good or bad, wherever you happen to be from in the world."**

- **There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.**
- There are moral truths that are wrong in every time, place, and situation without exception
- There are some actions that are considered morally wrong in all societies
- Some moral rules apply to all people regardless of the specific circumstances they are in
- We should all be judged according to a strict moral code regardless of where we come from
- People from all cultures should adhere to the same moral standards even if they do not agree with those standards

*[condition 2] [antirealism] [Y&D]*
**"Our morals and values are shaped by our culture and upbringing, so there are no absolute right answers to any moral question."**

- **Whether an answer to a moral question is correct or incorrect can only be judged according to the standards of different individuals or cultures**
- People have different life experiences and points of view, so we cannot impose a single moral standard on everyone without considering their particular circumstances
- No questions about what is morally right or wrong have a single correct answer that applies in all circumstances, because the same action can be right in some circumstances but wrong in others
- Each person's views about what is morally right or wrong is shaped by their unique background and experiences, so we frequently reach different conclusions about moral questions

- There are no moral truths, just different opinions about what is right or wrong
- Everyone's perspective on morality is different, so there's no way to know for sure whether an action is always right or wrong

{---------------------------------Page break---------------------------------}

*[comprehension check]*

In response to the question on the previous page, which of the following were you doing?

- **Selecting the single statement which I think best matches the meaning of the bolded sentence**
- Selecting the single statement that I most agree with
- Selecting every statement that I agree with
- Selecting every statement which I think best matches the meaning of the bolded sentence

{---------------------------------Page break---------------------------------}

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

O Male

O Female

O Other

### Study 4:
### Bush & Moss Scale

*[instructions]*

### Instructions

On the next page, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

*We are not asking you to choose the option that you agree with the most.* We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.

**[conditions] [random assignment] [one condition each]**

Below, you will see a statement in bold and a set of multiple choice options. *Please choose the option that you believe is the best interpretation of the meaning of the statement.*

We are not asking you to choose the option that you agree with the most. We only want to know which of the options most closely matches the meaning of the statement. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [realism] [B&M]*
**"There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise."**

- **Moral truth does not depend on cultural standards or personal values**
- The truth about what is morally right or wrong applies to all people regardless of the specific circumstances they are in
- Everyone should be held to the same moral standards even if they do not agree with those standards
- There are objective ways to measure whether some moral rules have good or bad outcomes, so individuals and cultures who think differently could be mistaken
- People should be held accountable for breaking some moral rules even if they believe their actions were justified
- Some actions are always right or wrong regardless of the intentions of the person committing the action

*[condition 2] [antirealism] [relativism] [B&M]*
**"The truth of all moral claims can vary depending on the moral standards of different individuals or cultures."**

- **Facts about what is morally right or wrong are made true by the standards and values of people and societies**
- No claims about what is morally right or wrong can be true in all circumstances, because the same action can be right in some circumstances but wrong in others
- Different people and societies have different beliefs about what is morally right or wrong
- Each society adapts its views about what is morally right to fit with their traditions and meet local needs
- There are no moral truths, just different opinions about what is right or wrong
- Since each person and society has different views about what is morally right or wrong, we can't be sure there is a single correct answer to every moral claim

*[comprehension check]*

In response to the question on the previous page, which of the following were you doing?

- **Selecting the single statement which I think best matches the meaning of the bolded sentence**
- Selecting the single statement that I most agree with
- Selecting every statement that I agree with
- Selecting every statement which I think best matches the meaning of the bolded sentence

{--------------------------------Page break--------------------------------}

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

○ Male
○ Female
○ Other

**Study 5:**
**Checkbox**

*[Instructions]*

<u>**Instructions**</u>

On the next page, you will see a sentence in bold and a set of additional statements. *Please select every statement which you believe matches the meaning of the bolded sentence.*

*We are not asking you to choose the statements that you agree with.* We only want to know how well each of these statements match the meaning of the bolded sentence. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.

{--------------------------------Page break--------------------------------}

**[conditions] [random assignment] [one condition each]**

Below, you will see a sentence in bold and a set of additional statements. *Please select every statement which you believe matches the meaning of the bolded sentence.*

*We are not asking you to choose the statements that you agree with.* We only want to know how well each of these statements match the meaning of the bolded sentence. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [antirealism] [ relativism | disagreement] [MRS]*
**"Different people can have opposing views on what is moral and immoral without anyone being wrong."**

- **Since moral truth is determined by what each individual believes, people who hold conflicting moral views can both be correct**
- Different people can have conflicting views about what is right or wrong without either of them having made an unjustified mistake
- People with different moral views can each be correct about some aspects of a moral issue without either being completely incorrect
- When people hold conflicting moral views sometimes there is no way to know which view is correct or incorrect
- People with opposing moral views cannot be wrong because it is not possible for people to be right or wrong about their moral views
- People have a right to their moral views even if other people hold opposing moral views

*[condition 2] [antirealism] [relativism | dependent truth] [MRS]*
**"The viewpoint of one's culture determines whether their actions are morally right."**

- **The truth about whether an action is morally right can only be judged according to the standards of different cultures**
- People from different cultures can be justified in holding different moral beliefs
- Each culture adapts its views about what is morally right to fit with their traditions and meet local needs
- Since each culture has different views about what is morally right, a person's actions will be judged as moral by some cultures and immoral by other cultures
- Different cultures have the right to hold different moral beliefs, so it isn't acceptable to impose our moral values on them
- The attitudes of each person's culture influence whether their actions are judged as moral or immoral

*[condition 3] [realism] [B&M]*
**"There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise."**

- **Moral truth does not depend on cultural standards or personal values**

- The truth about what is morally right or wrong applies to all people regardless of the specific circumstances they are in
- Everyone should be held to the same moral standards even if they do not agree with those standards
- There are objective ways to measure whether some moral rules have good or bad outcomes, so individuals and cultures who think differently could be mistaken
- People should be held accountable for breaking some moral rules even if they believe their actions were justified
- Some actions are always right or wrong regardless of the intentions of the person committing the action

*[condition 4] [realism] [universal conflation] [MRS]*
**"There are moral rules that apply to everyone regardless of personal beliefs."**

- **The truth of some moral rules is not determined by what people believe is morally right or wrong**
- Some moral rules apply to all people regardless of the specific circumstances they are in
- Everyone should be held to the same moral standards even if they do not agree with those standards
- People should be held accountable for breaking some moral rules even if they believe their actions were justified
- We should judge everyone according to a strict moral code even if we believe that there ought to be an exception in some cases
- Some actions are always right or wrong regardless of the intentions of the person committing the action

*[condition 5] [realism] [universal and normative conflation] [MRS]*
**"The same moral standards should be followed by people from all cultures."**

- **There is a correct set of moral standards that does not depend on what cultures think is right or wrong**
- We would be better off if people from every culture agreed to follow the same moral standards
- All cultures should be held to the same moral standards, so we should not tolerate cultures with moral beliefs that conflict with those standards
- People from all cultures should adhere to the same moral standards even if they do not agree with those standards
- We should not consider a person's cultural background when we judge their actions as morally right or wrong
- No matter what culture you come from, there are certain things everyone should know are right or wrong

*[condition 6] [antirealism] [relativism] [B&M]*

**"The truth of all moral claims can vary depending on the moral standards of different individuals or cultures."**

- **Facts about what is morally right or wrong are made true by the standards and values of people and societies**
- No claims about what is morally right or wrong can be true in all circumstances, because the same action can be right in some circumstances but wrong in others
- Different people and societies have different beliefs about what is morally right or wrong
- Each society adapts its views about what is morally right to fit with their traditions and meet local needs
- There are no moral truths, just different opinions about what is right or wrong
- Since each person and society has different views about what is morally right or wrong, we can't be sure there is a single correct answer to every moral claim

{---------------------------------Page break---------------------------------}

*[Comprehension check]*
In response to the questions above, which of the following were you doing:

- **Selecting every statement which I think matches the meaning of the bolded sentence**
- Selecting the single statement that I most agree with
- Selecting every statement that I agree with
- Selecting the single statement which I think best matches the meaning of the bolded sentence

{---------------------------------Page break---------------------------------}

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

○ Male
○ Female
○ Other

**Study 6:**
**Likert scale**

*[Instructions]*

**Instructions**

On the next page, you will see a sentence in bold and a set of additional statements. *Please select the degree to which you agree or disagree that the additional statements match the meaning of the bolded sentence.*

*We are not asking you to select how much you agree or disagree with the statements.* We only want to know how well each of these statements match the meaning of the bolded sentence. There are no correct or incorrect answers. We are only interested in what you think.

These instructions will appear on the next page for your convenience.

{---------------------------------Page break---------------------------------}

**[conditions] [random assignment] [one condition each] [each condition will include the same correct/incorrect response options that the item was paired with above, and will include two additional items: each realism item will include the other two realism items, and each antirealism item will include the other two antirealism items]**

Below, you will see a sentence in bold and a set of additional statements. *Please select the degree to which you agree or disagree that the additional statements match the meaning of the bolded sentence.*

*We are not asking you to select how much you agree or disagree with the statements.* We only want to know how well each of these statements match the meaning of the bolded sentence. There are no correct or incorrect answers. We are only interested in what you think.

*[condition 1] [realism] [universal] [MRS]*
**"There are moral rules that apply to everyone regardless of personal beliefs."**

*[condition 2] [realism] [universal and normative conflation] [MRS]*

**"The same moral standards should be followed by people from all cultures."**

*[condition 3] [realism] [B&M]*
**"There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise."**

*[condition 4] [antirealism] [relativism | disagreement] [MRS]*
**"Different people can have opposing views on what is moral and immoral without anyone being wrong."**

*[condition 5] [antirealism] [relativism | dependent truth] [MRS]*
**"The viewpoint of one's culture determines whether their actions are morally right."**

*[condition 6] [antirealism] [relativism] [B&M]*
**"The truth of all moral claims can vary depending on the moral standards of different individuals or cultures."**

*[1 = strongly disagree, 7 = strongly agree]*

{--------------------------------Page break--------------------------------}

*[Comprehension check]*
In response to the questions above, which of the following were you doing:

- **Indicating how much I agree or disagree that each statement matches the meaning of the bolded sentence**
- Indicating my agreement with the single statement I agree with most, and indicating disagreement with every other statement
- Indicating how much I agree with each statement
- Indicating my agreement with the single statement which I think best matches the meaning of the bolded sentence, and indicating disagreement with all other statements

{--------------------------------Page break--------------------------------}

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

○ Male
○ Female
○ Other

**Forced Choice & Spontaneous Theorizing Paradigm**
(Quantum mechanics)

**Study 1:**
**Quantum Mechanics**

*[instructions]*

**<u>Instructions: Part 1</u>**

In this study, we will present you with different perspectives about an issue. The topic you are presented with may be unfamiliar to you. However, no prior knowledge is necessary or expected. Please answer to the best of your ability.


{---------------------------------Page break---------------------------------}

*[quantum mechanics instructions]*

We will now ask you some questions about quantum mechanics. On subsequent pages we will ask you follow-up questions about specific theories related to quantum mechanics.

**[familiarity questions]**

*[fam_1] [randomized]*
Each of the following terms is related to quantum mechanics. For each term, please indicate if you are familiar with that term. You may select more than one.

- The Copenhagen interpretation
- The Many Worlds interpretation
- Mach–Zehnder interferometer
- Quantum harmonic oscillator
- Dirac equation
- Finite potential well
- Time-symmetric theory
- Bohm's interpretation

*[fam_2]*
What is the highest level of formal education **in physics** that you have completed?

○ None

○ Some coursework in grade school

○ Some coursework in high school

○ Some coursework in college

○ Undergraduate degree in physics (BS, BA, etc.)

○ Some graduate coursework in physics

○ Graduate degree in physics (MS, PhD, etc.)

*[fam_3]*
How well do you understand different interpretations of quantum mechanics?
*[1 = Very poorly, 7 = Very well]*

What level of expertise most closely reflects your knowledge of quantum mechanics?

○ No knowledge (I may have seen or heard some of these terms before but I have no further understanding of quantum mechanics)

○ Little knowledge (I know a little, so I understand about as much as someone who has not studied the topic or has at most some passing familiarity with it)

○ Novice (I have some knowledge, could perform some related mathematical equations and could answer basic questions about theoretical physics)

○ Proficient (I have significant knowledge, equivalent to a college student who has completed considerable coursework in physics and could perform complex mathematical equations, and could answer advanced questions about theoretical physics)

○ Expert (I have extensive knowledge, equivalent to comprehensive understanding of the mathematical equations underlying quantum mechanics, and have mastered the most advanced mathematical equations and questions about theoretical physics)

*[position]*
Do you have an opinion about which interpretation of quantum mechanics is correct?

○ Yes
○ No

{----------------------------------Page break----------------------------------}

## Quantum mechanics

Quantum mechanics is a branch of physics that deals with atomic and subatomic phenomena. It is a highly successful theory that makes precise predictions and has inspired new technologies. Despite this success, scientists disagree about how to interpret it as a literal description of the world.

One disagreement concerns the nature of some subatomic events. These events appear to have more than one possible outcome, but when we measure them, we only observe one outcome. There are two competing explanations for what occurs when we observe one of these outcomes.

### The Copenhagen Interpretation of Quantum Mechanics

According to the Copenhagen interpretation, we only observe one outcome because there is only one outcome. According to this view, such events have probabilistic outcomes. For example, a subatomic particle could have a 50% chance of moving to the left, and a 50% chance of moving to the right. We will not know which way it will move until we observe it. Once we observe it, there is an equal chance it ends up moving in either direction, like a coin flip. There is no way to know whether the particle would go to the left or the right in advance, because the result is random. Thus, the Copenhagen interpretation holds that some subatomic events have random outcomes.

### The Many Worlds Interpretation of Quantum Mechanics

According to the Many Worlds interpretation, we observe one outcome but there is actually more than one outcome. According to this view, events do not have probabilistic outcomes. Whenever a subatomic particle could move to either the left or the right, instead of it only moving in one or the other direction, the universe splits into two different universes. The particle moves to the left in one of the universes, and to the right in the other universe. Each of those universes contains an exact copy of us and everything else in the universe, with the only difference being the direction the particle moves. In one universe, we observe the particle move to the left, while in the other, we observe the particle move to the right. The reason the particle seems to have moved to either the left or the right is because we are in one of these universes, and cannot observe the other universe. But there is a copy of us in the other universe that saw the particle move in the opposite direction. As a result, the outcome of subatomic events is never random. Instead, all possible ways the particle could move are realized in different universes. Since many subatomic events occur every second, there are countless other universes, each slightly different from the one you are in.

{---------------------------------Page break---------------------------------}

### Instructions: Part 2

Now we will ask you what you think about the **Copenhagen** and **Many Worlds** interpretations of quantum mechanics.

*[correct]*
Think about the Copenhagen and Many Worlds interpretations of quantum mechanics. Which of these interpretations is correct?

- The **Copenhagen** interpretation is the correct interpretation of quantum mechanics. Whenever a subatomic event appears to have more than one possible outcome, only one of those outcomes randomly occurs.
- The **Many Worlds** interpretation is the correct interpretation of quantum mechanics. Whenever a subatomic event appears to have more than one possible outcome, all possible outcomes are realized. The universe splits into a different universe for each possible outcome, and each of those possibilities occurs in one of those universes.

*[everyday]*
How often does the topic of quantum mechanics come up in conversation in your everyday life?

○ Never (It has never come up in conversation)

○ Rarely (a few times in my life)

○ Sometimes (several times a year)

○ Often (several times a month)

○ Frequently (several times a week)

*[ordinary_agree]*
Do you think most ordinary people have an opinion on how to interpret quantum mechanics?

○ Yes, most ordinary people believe the Copenhagen interpretation is correct

○ Yes, most ordinary people believe the Many Worlds interpretation is correct

○ No, most ordinary people have no opinion about which interpretation is correct

*[ordinary_claims]*
When ordinary people make claims about events in the past, present, or future, do their claims presuppose a particular view about how to interpret quantum mechanics?

○ Yes, when people talk about events they presuppose that the Copenhagen interpretation is correct

○ Yes, when people talk about events they presuppose that the Many Worlds interpretation is correct

○ No, when people talk about events they do not presuppose any particular interpretation of quantum mechanics is correct

*[ordinary_practical]*
How much of an impact do people's opinions about how to interpret quantum mechanics have on their everyday life? (For instance, what kinds of decisions they make and how successful they are in pursuing their goals)

○ No impact

○ Very little impact

○ Some impact

○ Moderate impact

○ Enormous impact

{---------------------------------Page break---------------------------------}

Please indicate which of the following statements about the Copenhagen and Many Worlds interpretations is true or false.

*[comp_1]*
According to the Copenhagen interpretation, some events have random outcomes.

- **True**
- False

*[comp_2]*
According to the Many Worlds interpretation, if we observe a subatomic particle go to the left, there is another universe where it went to the right.

- **True**
- False

*[comp_3]*
According to the Copenhagen interpretation, events can have more than one outcome.

- True
- **False**

*[comp_4]*
According to the Many Worlds interpretation, we can observe all possible outcomes of subatomic events.

- True
- **False**

*[comp_5]*
Quantum mechanics is a branch of physics that primarily studies very large objects, such as planets and stars.

- True
- **False**

*[comp_6]*
Experts who study quantum mechanics disagree about which interpretation is correct.

- **True**
- False

{---------------------------------Page break---------------------------------}

*[forced_choice]*
In this study, you were asked to choose between the Copenhagen interpretation and the Many Worlds interpretation. However, it is possible that you do not agree with either of these positions. Did you feel that you had to choose between one of these two answers, even though it did not accurately reflect what you thought?

○ Yes
○ No

*[age]*
What is your age?

_____

*[gender]*

What is your gender?

O Male

O Female

O Other

## Third Person & Normative Entanglement Paradigms

➢ Study 1: Third Person Paradigm (pilot)
➢ Study 2: Third Person Paradigm with concrete moral issues
➢ Study 3: Third Person Paradigm (nonmoral domains)
➢ Study 4: Normative entanglement

### Study 1:
### Third Person Paradigm (pilot)

*[instructions]*

**[other conditions] [randomized, between subjects, one per participant]**

*[condition 1][no statement]*

Instructions: Part 1

Please read the following questions carefully. You will be asked several questions about what you think **a typical person in your society** believes. Try to interpret the statements in as normal and non-exceptional a way as possible.

*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

*[condition 2] [wrong]*

Instructions: Part 1

You will be asked several questions about what you think **a typical person in your society** believes. Try to interpret the statements in as normal and non-exceptional a way as possible.

Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"That is morally wrong."**

*[condition 3] [right]*

You will be asked several questions about what you think **a typical person in your society** believes. Try to interpret the statements in as normal and non-exceptional a way as possible.

Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"That is morally right."**

*Think about the person in this scenario. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[All conditions]**

*[other] [objectivism 1]*
They believe that when two people disagree about a moral issue at least one of them must be incorrect.

*[other] [objectivism 2]*
They believe there are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

*[other] [objectivism 3]*
They believe moral truth is independent of cultural standards and personal beliefs.

*[other] [universalism 1]*
They believe that there is one standard of moral truth by which all people should be judged.

*[other] [universalism 2]*
They believe the same moral rules apply to everyone.

*[other] [universalism 3]*
They believe there is a single set of moral standards for all people and cultures.

*[other] [relativism 1]*
They believe that when two people disagree about a moral issue they can both be correct.

*[other] [relativism 2]*
They believe that moral truth is simply what is true to each individual.

*[other] [relativism 3]*
They believe that things are only morally right or wrong according to different points of view.

*[other] [noncognitivism 1]*
They believe that moral views cannot be true or false.

*[other] [noncognitivism 2]*
They believe that judgments about whether an action was morally right or wrong can be correct or incorrect.

*[other] [noncognitivism 3]*
They believe that judgments about morality are about what is true or false and not just expressions of feelings.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

{---------------------------------Page break---------------------------------}

**[self condition]**

*[instructions]*

## Instructions: Part 2

Now we will ask you some questions about what **<u>you</u>** believe about the statements we just asked about.

{---------------------------------Page break---------------------------------}

*[instructions, repeated]*

## Instructions: Part 2

Now we will ask you some questions about what **<u>you</u>** believe about the statements we just asked about.[117]

*How strongly do you agree or disagree with the following statements?*

*[self] [objectivism 1]*
When two people disagree about a moral issue at least one of them must be incorrect.

---

[117] Instructions were provided on their own page once, between two page breaks, in order to prompt participants to recognize that the questions are shifting from judgments about someone else to judgments about themselves. Instructions were then repeated on the page with the items. That is why these same instructions appear twice here, with page breaks.

*[self] [objectivism 2]*
There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

*[self] [objectivism 3]*
Moral truth is independent of cultural standards and personal beliefs.

*[self] [universalism 1]*
There is one standard of moral truth by which all people should be judged.

*[self] [universalism 2]*
The same moral rules apply to everyone.

*[self] [universalism 3]*
There is a single set of moral standards for all people and cultures.

*[self] [relativism 1]*
When two people disagree about a moral issue they can both be correct.

*[self] [relativism 2]*
Moral truth is simply what is true to each individual.

*[self] [relativism 3]*
Things are only morally right or wrong according to different points of view.

*[self] [noncognitivism 1]*
Moral views cannot be true or false.

*[self] [noncognitivism 2]*
Judgments about whether an action was morally right or wrong can be correct or incorrect.

*[self] [noncognitivism 3]*
Judgments about morality are about what is true or false and not just expressions of feelings.

*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*

{--------------------------------Page break--------------------------------}

*[demographics]*

*[gender]*
What is your gender?

○ Male
○ Female

○ Other

*[age]*
What is your age?

_____

*[politics]*
Here is a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale?

*[1 = Extremely liberal, 2 = Liberal, 3 = Slightly liberal, 4 = Moderate, 5 = Slightly conservative, 6 = Conservative, 7 = Extremely conservative]*

*[religiosity]*
How religious do you consider yourself?

*[1 = Not religious at all, 7 = Very religious]*

**Study 2:**
**Third Person Paradigm with concrete moral issues**

*[instructions] [all conditions]*

**<u>Instructions: Part 1</u>**

On the pages that follow you will be presented with scenarios describing people and will then be asked how likely you think it is that they believe certain statements. Please read these scenarios and statements carefully. There are no right or wrong answers, please just provide your thoughtful first response.

{---------------------------------Page break---------------------------------}

*[condition 1] [no statement]*

Please read the following questions carefully. You will be asked several questions about what you think a typical person in your society believes. Try to interpret the statements in as normal and non-exceptional a way as possible.

*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[metaethics measures]**

*[other] [objectivism 3]*
They believe moral truth is independent of cultural standards and personal beliefs.

*[other] [universalism 2]*
They believe the same moral rules apply to everyone.

*[other] [relativism 3]*
They believe that things are only morally right or wrong according to different points of view.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

**[character measures]**

*[moral character]*
How would you rate the <u>moral character</u> of this person?

*[1 = Very morally bad, 2 = Moderately morally bad, 3 = Somewhat morally bad, 4 = Neither morally good nor morally bad, 5 = Somewhat morally good, 6 = Moderately morally good, 7 = Very morally good]*

*[empathy]*
How <u>empathic</u> would you consider this person?

*[1 = Not empathic at all, 7 = Very empathic]*

*[moral seriousness]*
How <u>seriously</u> do you think this person takes morality?

*[1 = Not seriously at all, 7 = Very seriously]*

*[partner preference]*
Please rate the extent to which this person would be a good person to have as a social partner (such as a coworker, neighbor, or close friend).

*[1 = Not good at all, 7 = Very good]*

{--------------------------------Page break--------------------------------}

*[concrete conditions] [order randomized, within subjects, all participants assigned to all conditions]*

*[condition 2] [meat]*

## Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"It is morally wrong to eat meat."**

*Think about the person who made this statement. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[metaethics measures]**

[Same as condition 1]

**[character measures]**

[Same as condition 1]

{--------------------------------Page break--------------------------------}

*[condition 3] [grade]*

## Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"It is morally wrong for a professor to give a bad grade to a student just because they dislike the student."**

*Think about the person who made this statement. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[metaethics measures]**

[Same as condition 1]

**[character measures]**

[Same as condition 1]

{--------------------------------Page break--------------------------------}

*[condition 4] [mock]*

## Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"It is morally wrong for a person to go to a funeral to mock the deceased person in front of their family."**

*Think about the person who made this statement. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[metaethics measures]**

[Same as condition 1]

**[character measures]**

[Same as condition 1]

{--------------------------------Page break--------------------------------}

*[condition 5] [alcohol]*

## Scenario

A person is having a discussion about moral issues. During this discussion the person says:

**"It is morally wrong for a woman who knows she is pregnant to drink alcohol."**

*Think about the person who made this statement. Please rate how likely or unlikely it is that they hold the following beliefs.*

**[metaethics measures]**

[Same as condition 1]

**[character measures]**

[Same as condition 1]

{--------------------------------Page break--------------------------------}

**[self condition]**

*[instructions]*

## Instructions: Part 2

Now we will ask you some questions about what **you** believe about the statements we just asked about.

{--------------------------------Page break--------------------------------}

*[instructions, repeated]*

Now we will ask you some questions about what **you** believe about the statements we just asked about.

*How strongly do you agree or disagree with the following statements?*

*[self] [objectivism 3]*
Moral truth is independent of cultural standards and personal beliefs.

*[self] [universalism 2]*
The same moral rules apply to everyone.

*[self] [relativism 3]*
Things are only morally right or wrong according to different points of view.

*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*

{---------------------------------Page break---------------------------------}

*[demographics]*

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

○ Male
○ Female
○ Other

**Study 3:**
**Third Person Paradigm (nonmoral domains)**

*[instructions]*

**Instructions: Part 1**

Please read the following questions carefully. You will be asked several questions about what you think a typical person in your society believes. Try to interpret the statements in as normal and non-exceptional a way as possible. There are no right or wrong answers, please just provide your thoughtful first response.

**[conditions] [randomized, between subjects, one per participant]**

**[condition 1] [morality]**

*[other]*

*[instructions]*
*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

*[other] [moral] [objectivism 1]*
They believe that when two people disagree about a moral issue at least one of them must be incorrect.

*[other] [moral] [objectivism 2]*
They believe there are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

*[other] [moral] [objectivism 3]*
They believe moral truth is independent of cultural standards and personal beliefs.

*[other] [moral] [universalism 1]*
They believe that there is one standard of moral truth by which all people should be judged.

*[other] [moral] [universalism 2]*
They believe the same moral rules apply to everyone.

*[other] [moral] [universalism 3]*
They believe there is a single set of moral standards for all people and cultures.

*[other] [moral] [relativism 1]*
They believe that when two people disagree about a moral issue they can both be correct.

*[other] [moral] [relativism 2]*
They believe that moral truth is simply what is true to each individual.

*[other] [moral] [relativism 3]*
They believe that things are only morally right or wrong according to different points of view.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

*[instructions]*

**Instructions: Part 2**

Now we will ask you some questions about what **you** believe about the statements we just asked about.

*[self] [moral] [objectivism 1]*
When two people disagree about a moral issue at least one of them must be incorrect.

*[self] [moral] [objectivism 2]*
There are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

*[self] [moral] [objectivism 3]*
Moral truth is independent of cultural standards and personal beliefs.

*[self] [moral] [universalism 1]*
There is one standard of moral truth by which all people should be judged.

*[self] [moral] [universalism 2]*
The same moral rules apply to everyone.

*[self] [moral] [universalism 3]*
There is a single set of moral standards for all people and cultures.

*[self] [moral] [relativism 1]*
When two people disagree about a moral issue they can both be correct.

*[self] [moral] [relativism 2]*
Moral truth is simply what is true to each individual.

*[self] [moral] [relativism 3]*
Things are only morally right or wrong according to different points of view.

*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*


**[condition 2] [taste]**

*[other]*

*[instructions]*
*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

*[other] [taste] [objectivism 1]*
They believe that when two people disagree about whether a painting is beautiful at least one of them must be incorrect.

*[other] [taste] [objectivism 2]*
They believe there are facts about which foods taste good or bad that are true even if some people or societies think otherwise.

*[other] [taste] [objectivism 3]*
They believe the truth about which food, music, and art are best is independent of cultural standards and personal beliefs.

*[other] [taste] [universalism 1]*
They believe that there is one standard of aesthetic truth by which all art and music should be judged.

*[other] [taste] [universalism 2]*
They believe that all art and music is subject to the same standards for judging whether it is good or bad.

*[other] [taste] [universalism 3]*
They believe there is a single standard for judging how good or bad food tastes that applies to all people and cultures.

*[other] [taste] [relativism 1]*
They believe that when two people disagree about whether a work of art is beautiful they can both be correct.

*[other] [taste] [relativism 2]*
They believe that truths about which people, landscapes, and architecture are most beautiful is simply what is true to each individual.

*[other] [taste] [relativism 3]*
They believe that claims about what is beautiful or ugly can only be true according to different points of view.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

{---------------------------------Page break---------------------------------}

*[self]*

*[instructions]*

*[self] [taste] [objectivism 1]*
When two people disagree about whether a painting is beautiful at least one of them must be incorrect.[118]

*[self] [taste] [objectivism 2]*
There are facts about which foods taste good or bad that are true even if some people or societies think otherwise.

*[self] [taste] [objectivism 3]*
The truth about which food, music, and art are best is independent of cultural standards and personal beliefs.

*[self] [taste] [universalism 1]*
There is one standard of aesthetic truth by which all art and music should be judged.

*[self] [taste] [universalism 2]*
All art and music is subject to the same standards for judging whether it is good or bad.

*[self] [taste] [universalism 3]*
There is a single standard for judging how good or bad food tastes that applies to all people and cultures.

*[self] [taste] [relativism 1]*
When two people disagree about whether a work of art is beautiful they can both be correct.

*[self] [taste] [relativism 2]*
Truths about which people, landscapes, and architecture are most beautiful is simply what is true to each individual.

*[self] [taste] [relativism 3]*
Claims about what is beautiful or ugly can only be true according to different points of view.

*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*

**[condition 3] [social convention]**

*[other]*

*[instructions]*
*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

---

[118] Note that in the study conducted and reported here, the wrong item was used for [self] [taste] [objectivism 1]. In place of the correct measure, which is featured here, I mistakenly used [self] [taste] [relativism 3].

*[other] [convention] [objectivism 1]*
They believe that when two people disagree about cultural practices, such as appropriate work attire, that at least one of them must be incorrect.

*[other] [convention] [objectivism 2]*
They believe there are facts about which customs and conventions are best even if some people or societies believe otherwise.

*[other] [convention] [objectivism 3]*
They believe that there are facts about which social conventions (such as shaking hands or making eye contact) are correct regardless of the preferences or values of different individuals or cultures.

*[other] [convention] [universalism 1]*
They believe that all people should adopt the same social conventions and customs, regardless of their culture or background.

*[other] [convention] [universalism 2]*
They believe the same social conventions and customs apply to everyone.

*[other] [convention] [universalism 3]*
They believe the same social norms (such as rules about how to greet others and how to dress) apply to everyone in the world.

*[other] [convention] [relativism 1]*
They believe that when two people disagree about which social conventions (such as greeting others by bowing or shaking hands) are correct they can both be correct.

*[other] [convention] [relativism 2]*
They believe that truths about social conventions are simply what is true according to different cultures and social groups.

*[other] [convention] [relativism 3]*
They believe that facts about which customs and practices are appropriate can vary according to the standards of different communities.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

*[self]*

*[instructions]*

417

*[self] [convention] [objectivism 1]*
When two people disagree about cultural practices, such as appropriate work attire, at least one of them must be incorrect.[119]

*[self] [convention] [objectivism 2]*
There are facts about which customs and conventions are best even if some people or societies believe otherwise.

*[self] [convention] [objectivism 3]*
There are facts about which social conventions (such as shaking hands or making eye contact) are correct regardless of the preferences or values of different individuals or cultures.

*[self] [convention] [universalism 1]*
All people should adopt the same social conventions and customs, regardless of their culture or background.

*[self] [convention] [universalism 2]*
The same social conventions and customs apply to everyone.

*[self] [convention] [universalism 3]*
The same social norms (such as rules about how to greet others and how to dress) apply to everyone in the world.

*[self] [convention] [relativism 1]*
When two people disagree about which social conventions (such as greeting others by bowing or shaking hands) are correct they can both be correct.

*[self] [convention] [relativism 2]*
Truths about social conventions are simply what is true according to different cultures and social groups.

*[self] [convention] [relativism 3]*
Facts about which customs and practices are appropriate can vary according to the standards of different communities.


*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*



**[condition 4] [facts]**

---

[119] Note that the version of the item used in Chapter 6, Study 3 had an error. The item used there stated, "When two people disagree about cultural practices, such as appropriate work attire, [that] at least one of them must be incorrect." The inclusion of "that" is a mistake that should be excluded from future uses of this measure.

*Think about a typical person in your society. Please rate how likely or unlikely it is that they hold the following beliefs.*

They believe that when two people disagree about matters of science or history that at least one of them must be incorrect.

They believe there are facts about science that are true even if some people or societies believe otherwise.

They believe that claims that describe what the world is like, such as the claims made in geography, history, and physics, are correct regardless of the preferences or values of different individuals or cultures.

They believe that facts about science and math are true everywhere and at all times, rather than only being true in some places and times.

They believe the way the world is (not simply the way it seems) is the same for everyone.

They believe the same scientific facts (such as the chemical composition of water and the shape of the earth) should be recognized by everyone in the world.

They believe that when two people disagree about matters of science or history, they can both be correct.

They believe that truths about what the world is like are simply what is true according to different cultures and social groups.

They believe that if the people from a particular society share a belief about what the world is like, their beliefs make that claim about what the world is like true.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

*[self] [facts] [objectivism 1]*
When two people disagree about matters of science or history at least one of them must be incorrect.[120]

*[self] [facts] [objectivism 2]*
There are facts about science that are true even if some people or societies believe otherwise.

*[self] [facts] [objectivism 3]*
Claims that describe what the world is like, such as the claims made in geography, history, and physics, are correct regardless of the preferences or values of different individuals or cultures.

*[self] [facts] [universalism 1]*
Facts about science and math are true everywhere and at all times, rather than only being true in some places and times.

*[self] [facts] [universalism 2]*
The way the world is (not simply the way it seems) is the same for everyone.

*[self] [facts] [universalism 3]*
The same scientific facts (such as the chemical composition of water and the shape of the earth) should be recognized by everyone in the world.

*[self] [facts] [relativism 1]*
When two people disagree about matters of science or history, they can both be correct.

*[self] [facts] [relativism 2]*
Truths about what the world is like are simply what is true according to different cultures and social groups.

*[self] [facts] [relativism 3]*
If the people from a particular society share a belief about what the world is like, their beliefs make that claim about what the world is like true.

*[1 = Strongly disagree, 2 = Moderately disagree, 3 = Somewhat disagree, 4 = Neither agree nor disagree, 5 = Somewhat agree, 6 = Moderately agree, 7 = Strongly agree]*


{---------------------------------Page break---------------------------------}


*[demographics]*

---

[120] Note that the version of the item used in Chapter 6, Study 3 had an error. The item used there stated, "When two people disagree about matters of science or history [that] at least one of them must be incorrect." The inclusion of "that" is a mistake that should be excluded from future uses of this measure.

*[age]*
What is your age?

_____

*[gender]*
What is your gender?

○ Male
○ Female
○ Other

**Study 4:**
**Normative entanglement**

**[conditions] [randomized, between subjects, one per participant]**

*[instructions]*

## Instructions: Part 1

In the scenario below, you will be presented with a discussion between two people. You will then be asked several questions about what you think about a particular person, given what they say in a discussion.

Scenario

Alex and Sam are having a discussion about moral issues. During this discussion, the following interaction takes place:

*[condition specific wording]*

*Think about what Sam said in this scenario. Please rate how likely or unlikely it is that Sam holds the following beliefs.*

*[condition 1] [realism | funeral]*

Alex: "Sam, do you believe it is objectively wrong for a person to go to a funeral to mock the deceased person in front of their family?"

**Sam: "Yes, I believe it is objectively wrong for a person to go to a funeral to mock the deceased person in front of their family."**

*[condition 2] [realism | alcohol]*

Alex: "Sam, do you believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol?"

**Sam: "Yes, I believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol."**

*[condition 3] [realism | torture]*

Alex: "Sam, do you believe it is objectively wrong to torture and kill a baby just for fun?"

**Sam: "Yes, I believe it is objectively wrong to torture and kill a baby just for fun."**

*[condition 4] [antirealism | no cancellation | funeral]*

Alex: "Sam, do you believe it is objectively wrong for a person to go to a funeral to mock the deceased person in front of their family?"

**Sam: "No, I don't believe it is objectively wrong for a person to go to a funeral to mock the deceased person in front of their family."**

*[condition 5] [antirealism | no cancellation | alcohol]*

Alex: "Sam, do you believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol?"

**Sam: "No, I don't believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol."**

*[condition 6] [antirealism | no cancellation | torture]*

Alex: "Sam, do you believe it is objectively wrong to torture and kill a baby just for fun?"

**Sam: "No, I don't believe it is objectively wrong to torture and kill a baby just for fun."**

*[condition 7] [antirealism | cancellation | funeral]*

Alex: "Sam, do you believe it is objectively wrong for a person to go to a funeral to mock the deceased person in front of their family?"

**Sam: "No, I don't believe it is objectively wrong to go to a funeral to mock the deceased person in front of their family. But I do think it is morally wrong. I just don't believe morality is objective. I still find it to be terrible and I am deeply opposed to it."**

*[condition 8] [antirealism | cancellation | alcohol]*

Alex: "Sam, do you believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol?"

**Sam: "No, I don't believe it is objectively wrong for a woman who knows she is pregnant to drink alcohol. But I do think it is morally wrong. I just don't believe morality is objective. I still find it to be terrible and I am deeply opposed to it."**

*[condition 9] [antirealism | cancellation | torture]*

Alex: "Sam, do you believe it is objectively wrong to torture and kill a baby just for fun?"

**Sam: "No, I don't believe it is objectively to torture and kill a baby just for fun. But I do think it is morally wrong. I just don't believe morality is objective. I still find it to be terrible and I am deeply opposed to it."**

*[All conditions]*

*[Metaethics]*

*[Sam] [objectivism 1]*
Sam believes that when two people disagree about a moral issue at least one of them must be incorrect.

*[Sam] [objectivism 2]*
Sam believes there are facts about what is morally right or wrong that are true even if some people or societies believe otherwise.

*[Sam] [objectivism 3]*
Sam believes moral truth is independent of cultural standards and personal beliefs.

*[Sam] [universalism 1]*
Sam believes that there is one standard of moral truth by which all people should be judged.

*[Sam] [universalism 2]*
Sam believes the same moral rules apply to everyone.

*[Sam] [universalism 3]*
Sam believes there is a single set of moral standards for all people and cultures.

*[Sam] [relativism 1]*
Sam believes that when two people disagree about a moral issue they can both be correct.

*[Sam] [relativism 2]*
Sam believes that moral truth is simply what is true to each individual.

*[Sam] [relativism 3]*
Sam believes that things are only morally right or wrong according to different points of view.

*[1 = Very unlikely, 2 = Moderately unlikely, 3 = Somewhat unlikely, 4 = Neither likely nor unlikely, 5 = Somewhat likely, 6 = Moderately likely, 7 = Very likely]*

*[character judgments]*

*[moral character]*
How <u>morally good or bad</u> is Sam?

*[1 = Very morally bad, 2 = Moderately morally bad, 3 = Somewhat morally bad, 4 = Neither morally good nor bad, 5 = Somewhat morally good, 6 = Moderately morally good, 7 = Very morally good]*

*[trustworthiness]*
How <u>trustworthy</u> is Sam?

*[1 = Very untrustworthy, 2 = Moderately untrustworthy, 3 = Somewhat untrustworthy, 4 = Neither trustworthy nor untrustworthy, 5 = Somewhat trustworthy, 6 = Moderately trustworthy, 7 = Very trustworthy]*

*[empathy]*
How <u>empathic</u> is Sam?

*[1 = Not empathic at all, 7 = Very empathic]*

*[judgmentalness]*
How <u>judgmental</u> is Sam?

*[1 = Not judgmental at all, 7 = Very judgmental]*

*[commitment]*
How <u>committed</u> is Sam to their moral values?

*[1 = Not committed at all, 7 = Very committed]*

*[political views]*
How <u>liberal or conservative</u> is Sam?

*[1 = Very liberal, 2 = Moderately liberal, 3 = Somewhat liberal, 4 = Neither liberal nor conservative, 5 = Somewhat conservative, 6 = Moderately conservative, 7 = Very conservative]*

*[religiosity]*
How <u>religious</u> is Sam?

*[1 = Not religious at all, 7 = Very religious]*

{---------------------------------Page break---------------------------------}

*[demographics]*

*[age]*

What is your age?

_____

*[gender]*
What is your gender?

○ Male

○ Female

○ Other

# APPENDIX E:
## Metaethics & the moral conventional distinction

### E1.1 The MCD is neither universal nor innate

The *moral/conventional task* (MCT) is the primary experimental paradigm used to support the claim that children and adults distinguish moral norms from conventional norms, and was developed for the explicit purpose of detecting this distinction in children. Before administering the MCT to participants, experimenters develop a list of prototypical transgressions of moral and conventional norms, which experimenters select based on their own *a priori* classification. Since the MCT was designed for children, prototypical moral norms have typically consisted of acts of aggression exhibited in school (e.g., one child hitting another child) and other common transgressions among children (e.g., refusing to share a toy; Smetana, 1981), though more recent versions include more adult-appropriate transgressions (e.g., stealing cars or inciting riots in Huebner et al. 2010, see also Fessler et al., 2015; Kelly et al., 2007). Since moral norms are thought of as non-arbitrary rules that apply to all communities and cannot be overridden by local authorities, they differ from prototypical conventional norms, which are conceived of as arbitrary social conventions that promote social coordination by reinforcing behavioral uniformity within a social group, regardless of whether those norms generalize to other groups (Nucci & Turiel, 1978). Like moral norms, conventional norms used in research have likewise tended to depict classroom behavior, such as designated seating in classrooms and putting toys away (Kelly et al., 2007; Nucci & Turiel, 1978).[121]

---

[121] To provide one concrete example of the precise set of items used, Smetana (1981, p. 1334) used the following items:

*Moral norms*
(1) one child hitting another child
(2) a child not sharing a toy
(3) a child shoving another child
(4) a child throwing water at another child
(5) a child taking another child's apple

*Conventional norms*
(6) a child not participating in show and tell
(7) a child not sitting in the designated place (on a rug) during story time
(8) a child not saying grace before snack
(9) a child putting a toy away in the incorrect place
(10) a child not placing her belongings in the designated place

Participants are presented with descriptions of moral and conventional transgressions and are asked a series of questions designed to elicit the predicted pattern of responses for each:

(1) *Does the transgression generalize to people in other times and places* (i.e., is it also impermissible for them)?
(2) *Is the transgression wrong regardless of what authorities or custom dictate?*[122]
(3) *Why is the transgression impermissible?*[123]

Kelly and stich point out that over sixty studies have employed the MCT, with most finding that participants reliably judge prototypical moral transgressions to generalize to other times and places, to be independent of authority, and to be justified in terms of harm, justice, or rights, while prototypical conventional transgressions exhibit none of these traits. As they put it, prototypical moral norms exhibit a *signature response pattern*, such that these characteristics reliably cluster together (the *signature moral pattern*), while the absence of this pattern applies to conventional norms (the *signature conventional pattern*).

Evidence of the MCD not only appear to be highly replicable, but generalizes to a wide range of populations and age groups, ranging from children around three years old (Smetana, 1981; Smetana & Braeges; 1990) to adolescents (Hollos, Leis, & Turiel, 1986) and adults (Huebner, Lee, & Hauser, 2010), as well as children with autism[124] (Blair, 1996) and children from both urban and rural populations (Song, Smetana, & Kim, 1987), higher and low socioeconomic backgrounds, higher and

---

[122] In practice, this is tested by asking if the transgression would be permissible if an authority figure stated that it was (e.g., Kelly et al., 2007; Nucci & Turiel, 1993; Tisak, Crane-Ross, & Tisak, 2000), if a law were passed that legalized the act (e.g., Huebner et al., 2010; Kelly et al., 2007), or if there were no rules against it (e.g., Nucci & Turiel, 1978; Smetana, 1981).

[123] Earlier versions of the MCD also asked about how *serious* a transgression was, with seriousness serving as a fourth characteristic distinguishing moral and conventional norms. However, seriousness could not be reliably associated with paradigmatic moral judgments, and was dropped from canonical versions of the MCD (Smetana, 1993; Stich, 2018, p. 554, endnote 6; 2019; Tisak & Turiel, 1988; Turiel, 1983). Some studies also ask about the amount of punishment transgressors deserve (e.g., Smetana, 1981). Stich (2019) points to a remark from Smetana that captures the exclusion of seriousness succinctly: "the severity of the transgression is not considered to be a formal criterion for distinguishing moral and conventional rules and transgressions" (Smetana, 1993, p. 117).

[124] Some evidence indications that children with psychopathic tendencies and adult psychopaths do not distinguish moral and conventional norms (Blair, 1995, 1997; Blair et al; 1995), though these results have been challenged by a recent, modified version of the MCT that found that a population of incarcerated psychopaths did draw the MCD (Aharoni, Sinnott-Armstrong, & Kiehl, 2012). With so little data, it remains unclear whether psychopaths distinguish moral and conventional norms.

lower IQ, and better and worse living conditions at home, including homes where children are neglected or abused (Smetana, 1984). The MCD also appears in insular religious communities (Nucci & Turiel, 1993) and in a diverse array of cultures spanning every populated continent (Yau & Smetana, 2003; Hollos, Leis, & Turiel, 1986; Nucci, Camino, & Sapiro, 1996; Nucci, Turiel, & Encarnacion-Gawrych, 1983; Song, Smetana, & Kim, 1987; for reviews, see Helwig, 2006; Killen, McGlothlin, & Lee-Kim, 2002; Nucci, 2001; Tisak, 1995; Turiel, 2002).

These findings seem to provide overwhelming evidence that the MCD is a ubiquitous feature of human cognition. There is little doubt that at least *some* prototypical normative transgressions conform to their respective signature response patterns in many populations. But it is less clear that the tendency to draw these distinctions adequately captures a genuine universal distinction between moral and conventional norms, much less an innate capacity for doing so. Several studies show that the signature response patterns don't emerge in every community or for every norm violation. In one of their own studies, Nucci and Turiel (1993) found that children from a Conservative Jewish[125] community in Chicago judged all of the religious rules they were given to be authority independent (e.g., working on the Sabbath, circumcision, and day of worship), despite these rules having little apparent relation to harm, justice, or rights. Likewise, Nisan (1987) found that Muslim Arab children in a traditional village in Israel treated all violations, including prototypical conventional norms (e.g., coed bathing and addressing teachers by first name) as authority independent (Kelly & Stich, 2007; but see Nisan, 1988; Turiel, Nucci, & Smetana, 1988). Conversely, a population of Colombian children treated both moral and conventional norms as contingent on authority (Ardila-Rey & Killen, 2001).

More recently, Kelly et al. (2007) found that many adult participants did not judge various harmful actions (e.g., spanking, physical abuse in the military) to be authority independent, and many

---

[125] Some references to this study (Kelly & Stich, 2007) describe the children as Orthodox Jews. However, Nucci and Turiel (1993) described the participants in their study as Conservative and distinguished this community from the Orthodox tradition (see Rosenthal, 1978).

judged slavery and corporeal punishment to be less serious transgressions when committed in the distant past than the present, suggesting that for some people, harmful and unjust acts do not generalize with the same strength to other cultures and historical periods. Unfortunately, there are some problems with this particular study. For instance, participants were slightly more likely to judge slavery in ancient Greece and Rome to be less bad than slavery in the southern United States 200 years ago. Yet Kelly et al. stated that slavery was important for both the economic and social systems of Greece and Rome, while they only mentioned the importance of slavery to economic system of the US. Coupled with the greater salience of historical knowledge of the latter, potential self-presentational and demand effects for not denouncing US slavery more thoroughly, and preexisting beliefs about the comparative seriousness of slavery in the ancient world compared to slavery in the southern states, participants may have simply reasoned that the two scenarios had substantive moral differences. Failing to regard slavery in the ancient world as equally morally bad may reflect the judgment that slavery was comparatively less severe than more recent instances of slavery. If so, this would not indicate a failure to generalize; it would simply reflect differences in the content of the moral actions themselves, rather than the cultural and temporal contexts in which they occurred. The same may hold true of their other scenario, involving corporal punishment of sailors. Sailors expecting and tolerating the practice in a culture which permits it may not feel that they are being treated unjustly or having their rights violated, whereas contemporary sailors presumably would object to being physically beaten. The cultural context in which an act occurs can influence how it is experienced, and this can hold true even for committed objectivists. Thus, once again, participants may have simply regarded the actions as having different moral standing, rather than being morally equivalent actions occurring in different contexts.

Although there may be credible alternatives to Kelly et al.'s (2007) account, their findings may still indicate a diminution in moral seriousness towards temporally distant moral acts and a willingness

to treat some prototypically moral acts as authority independent. Fessler et al. (2015) also offer compelling evidence that corroborates Kelly et al.'s findings. They presented transgressions involving harm, justice, and rights violations[126] to each of seven culturally distinct populations, from affluent urban populations in California to small-scale indigenous societies in South America, Fiji, Indonesia, and Melanesia, which exhibited varying degrees of social stratification. Participants were asked to judge how morally bad these actions were if they occurred in a distant time, a distant place, or if an appropriate local authority declared that the action was "not bad." All societies judged actions as less bad when performed in the past, in other cultures, or when approved by authority figures, but they showed dramatic differences in degree, with some (e.g., the Yasawa of Fiji) showing relatively little change in attitude and others (residents of Storozhnitsa, a Ukrainian village) expressing substantial reductions.[127]

Even early findings suggested that the signature response patterns of moral and conventional norms are neither stable nor reliable among young children. Gabennesch (1990) gathered a substantial collection of these findings. To take just one example, half of the second and fourth graders in one sample judged eating with your hands to be wrong independent of authority and to generalize to other societies (Carter & Patterson, 1982; see also Damon, 1977; Komatsu & Galotti, 1986; Lockhart, Abrahams, & Osherson, 1977; Miller & Bersoff, 1988, Tisak & Turiel, 1988). More recently, children in the United States judged disgust-inducing behaviors (e.g., pitting in a glass of water before drinking

---

[126] The seven items used described (adapted from Fessler et al., 2015, p. 3):

(1) a man stealing a stranger's money
(2) a man battering his wife without provocation
(3) a man striking and injuring a friend after the friend unintentionally injured him
(4) a man cheating a stranger in a market transaction
(5) a man knowingly spreading a false rumour that his rival is a thief
(6) the initiator of a fight bribing a witness to lie about who was at fault, resulting in the innocent party being punished
(7) and a man raping an unfamiliar woman

[127] Surprisingly, urban Californians exhibited moderate reductions, which seems inconsistent with claims that Westerners uniformly draw sharp distinctions between moral and conventional norms (but see Piazza & Sousa, 2016 for a methodological critique and Fessler et al., 2016 for a reply).

it and picking your nose) to not only be authority independent, but to be serious transgressions that generalized to other populations (Nichols, 2004). College students did not judge the same acts to generalize to other populations, but likewise judged them to be serious and authority independent transgressions. Likewise, low-SES populations in the United States and Brazil treated victimless acts as serious moral violations, despite these behaviors appearing to be unrelated to harm, justice, and rights, such as cleaning the toilet with your nation's flag, or siblings kissing each other passionately in private (Haidt, Koller, & Dias, 1993).

These observations played a role in the development of Haidt's Moral Foundations Theory, which has gathered substantial evidence that some individuals and cultures moralize content domains other than harm, or the sorts of rights and justice violations that appear in Turiel's MCT, including violations related to purity, loyalty, and authority (Haidt, 2012; Haidt & Joseph, 2004; Graham et al., 2013). In fact, people appear capable of flexibly construing almost any otherwise trivial issue in either moral or nonmoral terms under the right circumstances, such as wearing a sweater-vest (Allidina & Cunningham, 2018; Van Bavel et al., 2012). This suggests that the scope of moralized beliefs may not be exclusively determined by their content, and that some individuals and cultures moralize beliefs that are not justified in terms of harm, justice, and rights. Nisan (1987) proposed one explanation for differences like these, arguing that, while people in the West may tend to cordon off norms associated with welfare, justice, and rights from social conventions, other cultures "may anchor norms in heteronomous commands, thereby diminishing this distinction" (p. 719). In support of this suggestion, Nisan appeals to Shweder, Mahapatra, and Miller (1987), whose interviews in India and the United States suggest that people in the former may not distinguish convention from morality because "the social order is not separated from the natural moral order. In these cultures, customary practices that we regard as conventions are viewed as part of the nature of things" (pp. 719-720). If so, then there may be no principled way for these cultures to distinguish conventional and moral

norms, since the two stem from the same fundamental source. This would explain why conservative Jewish children regarded their religious conventions as authority independent. If so, this would suggest that the MCD is not an innate feature of normative cognition, but is instead a culturally constructed distinction that is contingent on surrounding beliefs about the source and justification of norms. Since these beliefs may vary across cultures, the MCD may only emerge in some cultures.

Even if the signature moral and conventional patterns emerged with respect to some norms across most or even all cultures, this would not decisively establish that Turiel and colleagues had provided a satisfactory account of the moral domain. Turiel relies on an *a priori* conception of morality that defines it in advance as serious, authority independent, generalizable, concerned with harm, justice, and rights, and for Turiel (1983), morality *just is* "analytically independent of systems of social organization that coordinate interactions" (p. 39, as quoted in Quintelier et al., 2013, p. 222). In other words, moral norms are not mere social conventions *by definition*. This preconception about morality has led to a narrow and selectively generated set of transgressions that fit his and other researcher's intuitions about what would represent prototypical moral and conventional norms, but excludes transgressions that might not neatly conform to their judgments. Sure enough, when Huebner et al. (2010) included a wider range of transgressions, they did find that moral items included still displayed the signature moral pattern, but conventional norms formed a continuum ranging from those displaying the signature conventional pattern to those exhibiting one or more qualities associated with the moral domain. This suggests that there may be no sharp dividing line between moral and conventional norms, despite Huebner and colleagues interpreting their findings as at least a partial vindication of the MCD.

Another issue with the *a priori* approach Turiel and colleagues take is that little effort is made to determine whether subjects *themselves* view these issues as moral. Evidence indicating that folk moralization is inconsistent with the MCD puts pressure on Turiel and colleagues to explain why they

are inconsistent. One way to respond to this is to argue that people lack introspective access to their own patterns of judgment, or otherwise fail to explicitly recognize and articulate a reliable distinction between categories of norms despite indirect questions revealing a pattern like the MCD. But absent such evidence, indications of inconsistencies between how people explicitly categorize norms and the MCD puts pressure on Turiel and proponents of the MCD to explain why these inconsistencies exist. For instance, when Wright, Cullum, and Schwab (2008) explicitly asked to categorize a long list of topics and behaviors as moral or nonmoral, participants overwhelmingly agreed that some topics were nonmoral, such as music preferences and exercise, but many social behaviors and many seemingly paradigm instances of moral behavior, many of which Turiel and colleagues would undoubtedly classify as moral or be at pains to insist were not moral beliefs, did not even approach unanimity. For example, the proportions below represent just a handful of the proportion of participants who judged the item in question to be moral:

| | |
|---|---|
| Abortion | 76% |
| Cheating on exams/papers | 69% |
| Honesty | 66% |
| Euthanasia | 36% |
| Incest | 74% |
| Rape | 90% |
| Euthanizing disabled children | 77% |
| (Adapted from Wright, Cullum, & Schwab, 2008, p. 1465) | |

Participants frequently moralized issues that do not unambiguously conform to the MCD, either, including eating pets (51% judged this to be a moral issue), pornography (18%), gay marriage (25%), sexual promiscuity (21%), preferring women/minorities (21%), and believing in God (16%, p. 1465). Wright, Grandjean, and McWhite (2013) found similar results in a second study, only this time participants were asked to judge whether the item in question was moral or fell into one of several other categories (e.g., "social," "personal," "scientific fact," etc.). Once again, there was little

consensus on whether a given item was moral, with the item most people agreed was moral (racial discrimination) achieving only 67% agreement.

Similar studies find significant cultural differences in what people regard as moral or nonmoral. Chinese participants were more likely to include "uncivilized" behavior in their notion of morality, compared to the Western emphasis on harm (Buchtel et al., 2015), and Levine et al., (2021) report similar variation in propensity to categories issues as moral or nonmoral among Christian, Mormon, Jewish, and nonreligious participants. Arguably, many people may differ in which actions they regard as having harmful consequences (or are unjust or violate rights), in which case, variation between participants could be captured by differences in belief about whether an action causes harm, is unjust, or violates someone's rights (Gray, Schein, & Ward, 2014). However, if this were so it would still raise serious methodological problems for the way the MCT is designed. Existing data using the MCT relies on researcher's own *a priori* assumptions about whether a given act is prototypically moral or not, rather than the attitudes of participants themselves. Even if the MCD did reliably emerge among diverse populations, if it is inconsistent with the explicit distinctions people themselves make between moral and nonmoral norms, it is unclear whether this pattern has earned the empirical credibility to supersede folk classification and justifiably claim to refer to the *moral* domain.

The MCD also focuses exclusively on moral *transgressions*, but it is implausible that "the moral domain" consists exclusively of moral violations. People care about positive moral duties, supererogatory acts, moral character, the moral status of institutions, laws, and other abstractions that don't reduce to specific transgressions, and other potentially moralized concerns which may fall within a broader conception of the moral domain the MCD appears to exclude. Thus, even if people did distinguish moral *transgressions* from conventional transgressions, it's not obvious that the same distinction would extend to other qualities and acts that seem to fall within a plausible conception of what the moral domain is supposed to entail.

Given these challenges, it is unclear whether the MCT establishes that people reliably distinguish a domain of norms that we may justifiably regard as *moral*, since it is unclear whether prototypical moral norms uniformly exhibit the signature moral response pattern. On the contrary, it appears many children and adults (a) moralize issues unrelated to harm, justice, and rights (b) often regard prototypically conventional norms as serious, authority independent, and generalizable (c) often judge prototypically moral norms to lack one or more of these qualities (d) do not agree on what counts as a moral or nonmoral norm or issue and (e) moralize more than just transgressions. The MCD fails to pick out a distinctive cluster of characteristics that reliably distinguish moral norms from conventional norms. As a result, the MCD literature does not provide adequate evidence that there are distinctively moral norms characterized by (among other things) universality, which obviates the need for Stanford (2018) and others to account for a purportedly innate capacity for the universality of moral thought.

### E1.2 The MCD is not evidence of realism

I speculate in the main text about various reasons philosophers have not engaged with or conducted research on folk metaethics. I omitted the possibility that research on the moral/conventional distinction (MCD) has already confirmed that ordinary people are moral realists. The intuition that moral norms differ from nonmoral norms will probably seem compelling to most of us. Consider what *telling white lies* and *committing genocide* have in common. Very little, it would seem. But the moment we contrast them with *wearing crocs to a wedding* or *showing up to a meeting in our pajamas*, their relationship is obvious: despite the vast gulf in severity, white lies and genocide are both *moral* issues. Questionable fashion choices are not. One way to put this is that genocide and white lies fall within the *moral domain*, a distinct category of norms that we readily distinguish from nonmoral norms.

As discussed in the previous section, this intuitive distinction is supported by a substantial body of research pioneered by Turiel and others demonstrating that children[128] and adults reliably attribute a distinct cluster of properties to moral norms that they do not attribute to other norms (Huebner et al., 2010; Nucci, 1982; Nucci & Nucci, 1982; Nucci & Turiel, 1979; 1993; Nucci, Turiel, & Encarnacion-Gawrych, 1983; Smetana, 1981; Smetana & Braeges, 1990; Smetana, Kelly, & Twentyman, 1984; Smetana et al., 1999; Turiel, 1979; 1983). This research purportedly shows that people regard moral norms as *(1) authority independent* (the rule cannot be overturned by authorities)*, (2) universal* (i.e. they apply to everyone, including people outside the participant's community), and *(3) justified in terms of harm, justice, and rights*, while conventional norms (such as table manners) display the inverse of these characteristics (i.e., authority dependence, non-generalizable, and not justified by appeal to harm, justice, or rights).

This moral/conventional distinction (MCD) has fueled claims that humans possess an innate, pancultural capacity for distinctively moral cognition (Dwyer, 1999, 2004; Joyce, 2006) characterized by the experience of moral norms as *objective moral truths* (Levy, 2005; Stanford, 2018). This use of *objective* means more or less what I mean by *realism*, in that *realism* is the view that there are *"objective" (stance-independent) moral facts*. Whatever Turiel's own interpretation of his findings, others have taken the MCD as evidence of folk realism. Nucci (2001) appeals to features of the MCD to argue that, from an early age, children "understand that it is objectively wrong to hurt others" (p. 86). Nichols and Folds-Bennett (2003) express a similar view, claiming that "the child's capacity to distinguish morality from convention shows that children regard moral violations as objectively wrong" (p. 23), and Kelly et al. (2007) claim that among the "core ideas that researchers in this tradition have advanced about moral rules" is that they have "an objective, prescriptive force" that is "not dependent on the authority

---

[128] As young as age two or three (Royzman, Leeman, & Baron, 2009; Smetana & Braeges, 1990).

of any individual or institution" (p. 118; for more examples see e.g., Colebrook & Sarkissian, 2018; Heiphetz & Young, 2017; Stanford, 2018).

Evidence of the MCD has also inspired more recent efforts to directly investigate folk realism. Both sets of findings purportedly demonstrate that children are realists, and that despite some increase in antirealism as people mature, adolescents and adults for the most part continue to be realists about the moral domain but not other evaluative domains, such as aesthetic and taste preferences (Beebe et al., 2015; Beebe & Sackris, 2016; Heiphetz & Young, 2017; Nichols & Folds-Bennett, 2003; Schmidt, Gonzalez-Cabrera, & Tomasello, 2017; Wainryb et al., 2004). Both lines of evidence have in turn prompted efforts to explain why this unique mode of cognition evolved, or at least why it is widespread (Dwyer, 1999; 2004; Dwyer, Huebner, & Hauser, 2010; Joyce, 2006, Ch. 4; Stanford, 2018). Jebari and Huebner (2018) summarize a recent example from Stanford (2018):

> Stanford argues that human ultrasociality is possible because we treat moral obligations as part of an externally imposed moral order, which applies equally to all; because the experience of moral motivation feels *objective*, it automatically generates the demand that others be similarly motivated, and in populations of like-minded individuals, this yields correlated interactions that are less likely to be exploited. (p. 22, emphasis mine)

Yet Stanford's claims rely in part on evidence of the universality of the MCD and the claim that it supports folk realism. Unfortunately, the MCD does not support Stanford or anyone else's claims of a pancultural tendency towards moral realism. As a result, there is no need to put forward elaborate accounts of how and why a capacity for distinctively moral cognition evolved, or to explain why we regard moral norms as objective.

Unfortunately, two features of the MCD, *universality* and *authority independence,* are frequently cited as evidence that people are moral realists (e.g., Colebrook & Sarkissian, 2018; Kelly et al., 2007; Nichols, 2004; Nichols & Folds-Bennett, 2003; Nucci, 2001; Stanford, 2018).

Neither attribute entails nor even strongly suggests realism. Recall that realism is the claim that moral truths are made true by certain stance-independent facts about the world, rather than the

attitudes, values, preferences, or beliefs of different individuals, nor consensus among social groups (Shafer-Landau, 2003).[129] As Goodwin and Darley (2008) note, universality does not entail realism because "the question of whether ethical standards should apply to all cultures is a question about the *scope* of ethical standards, and is independent of the question of whether such standards and beliefs are objectively or subjectively true" (p. 1341). Realism concerns the source of moral truth: whether such standards are made true by facts that aren't reducible to any particular standard of evaluation, i.e., they're *stance independent.* In principle, subjectivists, cultural relativists, and noncognitivists could judge that others ought to conform to their values, even if they acknowledge that those values are not objectively true. There is no compelling reason to think antirealists must believe that their moral standards are only locally applicable, and ought not be universalized. They *may* think this, or they *may* not. Whether or not the antirealist does so is a contingent matter and cannot be assumed *a priori.* Contrary to popular misconception, cultural relativism does not require that one must be tolerant of other cultures or refrain from seeking to impose one's moral values on other societies (Bush, 2016).

More generally, antirealist positions do not commit antirealists to any particular normative moral standards. Subjectivism, noncognitivist, error theory, *do not* entail that one's moral standards are only locally applicable, e.g., that they only apply to oneself or members of your culture. Even error theorists, who deny that there are moral facts at all, can and often do believe that everyone ought to abide by the same moral standards, where *ought* is understood to convey their nonrealist moral position. Who one believes their moral standards ought to apply to simply has *nothing to do* with whether those moral norms are stance-independently true. Since universality is consistent with

---

[129] This should not be mistaken for the claim that people's attitudes and values are morally irrelevant. Attitudes and other mental states can play a role in determining the moral status of an act even for objectivism. For instance, if it were an objective moral fact that one ought to treat others as they wish to be treated, how any given person wishes to be treated would depend on their preferences. However, the moral rule itself would not be true or false based on whether people believed or preferred it to be true.

antirealism, belief in the universality of moral norms cannot be used to infer objectivism. I'll cite myself as an example. I'm a moral antirealist. I don't believe there are any stance-independent moral facts. Nevertheless, I think that *nobody* should torture babies for fun, and that *everyone* should (generally) be honest and compassionate. I just don't think there are stance-independent normative facts that provide people with external reasons to comply with these moral standards.

It's also possible that judgments about the universality of a norm indicate inferences about what norms are likely to emerge in other communities. Since conventional norms can be at least somewhat arbitrary[130], there is little reason to expect precisely the same set of conventions to arise everywhere. Yet the universality of suffering and aversion to injustice could lead people to infer that other communities would have an interest in enforcing the same "moral" rules. If so, universality would not indicate objectivism, but a recognition that non-arbitrary norms are likely to arise in similar contexts. For example, if a child affirms that they think it would be wrong for children in distant times or places to hit each other or divide treats unequally, this would not necessarily indicate that they think these actions are *objectively* wrong; it could simply involve the recognition that other children tend to share the same subjective preferences, preferences which could very well be true on account of stance-*dependent* facts about shared psychological characteristics.[131]

For comparison, we would expect people in other societies to share similar food preferences, e.g., a preference for meat and sweet foods over dirt and feces, yet this would not demonstrate we are "gastronomic realists" that believe facts about which food is good or bad are true independent of our

---

[130] There may be non-arbitrary elements of social conventions, e.g., style of dress may be both a matter of convention and promote comfort or survival in particular climates.

[131] Children may also be reluctant to state that a moral rule would not apply in other times and places due to a desire to appear to be "good," well-behaved children. If so, demand effects may inflate the apparent strength of the MCD. Haidt et al. (1993) found intriguing hints of this possibility, observing that:

"In pilot testing, we found that many children judged violations of social conventions, such as eating with one's hands, to be universally and unalterably wrong, across many wordings of the probe questions. We feared that some children thought they were being tested and were seeking to demonstrate that they were 'good' children by condemning all violations in the strongest possible terms." (p. 618)

stance or preference toward food (Loeb, 2003). In other words, merely expecting or presuming that other people would have similar standards, values, rules, or preferences does not entail that you think those rules reflect stance-independent normative or evaluative facts. In short, inferences about the universality of norms or evaluative attitudes are consistent with both realist and antirealist stances; they are *not* identical with realism and nor even good evidence of realism.

If universality does not establish objectivity, what about authority independence? This is a far better candidate, as it suggests that moral norms cannot be arbitrarily dispensed with on a whim, which could imply that moral norms are derived from some external source. Yet authority independence is also not great of realism because believing that moral norms can or cannot be changed by authorities are both consistent with realism *and* antirealism.

First, realism is compatible with believing (at least some) moral norms are authority-dependent in certain respects. For instance, a realist could believe that certain authority figures are in a position to make determinations about whether an action is permissible or permissible, e.g., people could believe that they have an objective moral obligation to abide by the decrees of religious figures, the rulings of judges, or the moral rules set by the elders of the community. In such cases, these people would believe in a form of lower-order authority-dependence, whereby authority figures are endowed with the ability to determine or change moral rules.[132] If so, they could also believe there is a higher-order objective fact of the matter that authorities have this ability (and perhaps objective facts about who is and isn't an authority). However, questions about whether authorities can change rules do not disambiguate a metaethical interpretation of authority-independence from the lower-order, normative form of authority independence I've described. As such, responses to questions about authority independence do not clearly and ambiguously prompt metaethical interpretations, but could instead

---

[132] Perhaps with some constraints, e.g., they can make moral rules about who can get married, but not whether an incestuous marriage is acceptable.

prompt normative interpretations. If so, questions about authority independence may not be valid measures of higher-order (metaethical) conceptions of authority independence that could potentially serve as evidence of folk realism.

At the same time, antirealists may not endorse authority-dependence. One of the primary alternatives to realism is *relativism*, which treats the truth of moral claims as contingent on the standards of cultures (*cultural relativism*) or individual subjective values (*subjectivism*) (Bush, 2016). Both realism and relativism are compatible with denying that moral norms are dependent on authorities. For subjectivists, whether an action is moral depends on individual attitudes, not the judgments of authorities. Unless a subjectivist regards an authority figure as being in a position to do so, that figure simply lacks the ability to alter moral rules by fiat. Granted, it is *possible* for a subjectivist to outsource their moral position to authority figures by believing e.g., "My personal moral position is that I ought to abide by the moral judgments of authorities," but they *need not* think this. And most people probably *don't* think this. So it'd be odd to presume that someone who denies authority independence isn't a realist, which is what one would have to do if one is to take agreement with authority independence to be evidence of realism.

Likewise, there is no reason why cultural relativists must think moral rules are authority dependent. For cultural relativists, moral rules are true or false relative to the standards of different cultures. It may simply not be part of a culture's standards that authority figures can alter moral rules on a whim. Instead, they could believe that their moral standards cannot be changed, or that if they can be changed, that individuals lack the authority to do so. Perhaps changes to moral rules require cultural consensus, or novel exegesis of sacred texts, or a democratic referendum. It could be anything. And even if a person feels that *nothing* can change the moral rules, this still does not entail realism. An antirealist is under no obligation to regard their moral position as negotiable or subject to change based on social consensus or the public decree of authorities. One of the most persistent and egregious

mistakes researchers make in inferring realism from authority-independence is the failure to recognize that one's first-order moral beliefs may be rigid and non-negotiable *regardless* of whether that person is a realist or antirealist. I submit myself as an example: I am an antirealist. I don't believe there are any moral facts. And yet my response to questions used to establish the belief that moral norms are authority-independent are exactly the same as the realist: should people in other communities be able to change the moral rules so that hitting people or stealing their belongings is okay? No. If a teacher or authority figure says it would be okay to break these rules, or that these rules no longer exist, would that be okay? No. Do I think everyone, everywhere, should follow the same moral rules? Yes. I'm not unusual in this regard. Most moral antirealists have strong moral standards and values, just like realists do. There's no reason to think they're especially likely to be indifferent to people committing atrocities overseas or to think it would be morally permissible to kill their grandmother and bake her into a pie if Congress passed a bill declaring that they could.

Relativists could also believe morality is authority-dependent, but that the figure *in the situation the participant was presented with* lacked the requisite authority to change the rule. Thus, many studies using the moral-conventional task couldn't detect genuine authority-dependence even if it were present. Finally, there is no reason to think that noncognitivists must be committed to the notion that authorities can change moral rules. The belief that moral norms express nonpropositional attitudes simply does not entail authority-dependence.

In short, realism and antirealism are both conceptually consistent with authority-dependence and authority-independence, and their consistency with each is neither strained nor implausible. If there is an association between authority-independence and realism, it is at best only a contingent one, i.e., it could be that antirealism correlates with authority-independence, even if neither conceptually entails the other or should be presumed on *a priori* grounds to suggest the other. As a result, measures of authority-independence cannot distinguish realists from antirealists without independent empirical

evidence of a contingent connection of some form. This is because, if the belief that moral norms were authority independent indicated a commitment to realism, we should be able to distinguish moral realists from moral antirealists by how they respond to the MCT. Not only is there no evidence of such a connection, what little evidence there is suggests this is *not* the case. Nichols (2004) found that realists and antirealists distinguished between moral and conventional norms in the same way, which is hard to reconcile with claims that the MCD is evidence of folk realism.[133] Even if there is a correlation between belief in authority-independence and realism, it is implausible that they form such a tight connection that measures of the former can serve as proxies for the latter. The persistence of the belief that the MCD is evidence of realism is, in light of this, rather puzzling. Many researchers may have simply taken universality or authority independence to conceptually entail realism. This is definitely *not* the case. Still others may have presumed a contingent connection between features of the MCD and realism. Yet there is no obvious reason why universality or authority-independence should entail or even be an especially strong indication of a commitment to realism. I'm an antirealist, and I endorse both universalism and authority-independence: I think murder is *bad* and that there should be laws against it. I think it's bad and should be illegal regardless of whether it occurs in my country, a distant country, or an alien planet. And I don't think authority figures should have carte blanche to override rules against assaulting other people. Speaking on behalf of antirealists more generally, *almost all of us think this.*[134] The MCD is just as transparent and natural a distinction for me to draw as it is to most members of WEIRD populations, even for those of us who are cognizant of

---

[133] This is, of course, assuming that the measures used to classify participants as realists or antirealists were themselves valid. Perhaps they were not. This may be cold comfort for anyone treating the MCD as evidence of realism since methodological reasons for doubting the validity of direct measures of folk metaethical belief might also apply to the MCT.
[134] This *ought* to be obvious to anyone who spends a minute reflecting on what it would mean if only realists endorsed universality and authority-independence. Are those who make this error operating under the assumption that antirealists like me are *indifferent* to moral rules in faraway places? Or that we think that if judges and law enforcement declared it okay to chop our neighbors up with a hatchet that it would be fine to do so? It is so absurd on the face of it that I can only imagine that they haven't really adequately reflected on the implications of the MCD being diagnostic of realism. More generally, critics of antirealism often seem to underestimate the resources antirealists have available for accommodating conventional moral beliefs and attitudes.

being antirealists. Recognizing the MCD is simply and straightforwardly *not* an indicator of realism. It is time we retire this misconception.

**SCHRÖDINGER'S CATEGORIES:**
**THE INDETERMINACY OF**
**FOLK METAETHICS**

**A Dissertation**
*Presented to the Faculty of the Graduate School*
*of Cornell University*
*in Partial Fulfillment of the Requirements for the Degree of*
*Doctor of Philosophy*

by

Lance Spencer Bush
May 2023

# SUPPLEMENTS

# TABLE OF CONTENTS

# SUPPLEMENT TO CHAPTER 1

## S1.0 Introduction

This section provides a more detailed explanation of the central arguments presented in the main text, and clarifies certain terms, concepts, and claims that would require too detailed a discussion to appear in the main text. I also address a variety of miscellaneous concerns that may arise when reading the main text.

## S1.1 Clarifying the central argument

My central argument is that folk metaethics is *mostly* indeterminate with respect to moral realism and antirealism.[1] Specifically, I argue that *most* ordinary people do not have philosophical stances towards the truth status of first-order moral claims[2] and that *most* ordinary language does not contain implicit philosophical commitments to any particular account of the truth status of first-order moral claims.[3] A first-order moral claim is a claim about what is morally right or wrong, good or bad, and so on (Olson, 2014).[4] Second-order moral claims are claims about morality (including first-order moral claims). In denying that people have determinate metaethical stances and commitments, I am *not* claiming that people have no determinate first-order moral stances or commitments. I am claiming that they have no determinate second-order moral stances or commitments specifically with respect to whether first-order moral claims are stance-independently true or not.

I say *most* because there are some instances in which ordinary people do hold a determinate stance towards realism or antirealism, and there are some instances of moral judgment and discourse

---

[1] There is no simple way to characterize what this dispute is about, due to inconsistencies in the labels, terms, and distinctions.

[2] At least not ones that correspond to the traditional categories recognized by philosophers.

[3] Most research on (1) focuses on whether people endorse *objectivism*, *relativism*, *noncognitivism*, etc. or use moral language in a way best fits one of these accounts.

[4] Olson (2014) characterizes first-order moral claims as entailing that "some agent morally ought to do or not do some action; that there are moral reasons for some agents to do or not do some action; that some action is morally permissible, that some institution, character trait, or what have you, is morally good or bad, and the like." (p. 11)

that best fit one or another of competing metaethical analyses. Thus, I acknowledge that there are portions of folk metaethical discourse that *are* determinate with respect to moral realism and antirealism (see Gill, 2009). My position could be more accurately characterized as indeterminacy with a splash of *pluralism*[5], the hypothesis that elements of both *realism* and *antirealism* do determinately characterize some instances of folk moral thought and language. However, insofar as this is the case, it is unlikely that ordinary moral thought and language share a *uniform* commitment to a *shared* metaethical framework (see Gill, 2009). It also remains an open question how much, and to what extent, ordinary moral thought and language is characterized by determinate metaethical stances and commitments.

*Why* do I think ordinary people sometimes have (or at least occasionally express) determinate metaethical stances or commitments? On occasion, a participant in one of my studies *does* provide a decent description of realism or antirealism (or some derivative of the two, e.g., relativism), suggesting that on occasion people that may lack significant philosophical training nevertheless pick up on and exhibit some competence with the relevant concepts. And some people may be exposed to just enough philosophy at the pub, at a meditation retreat, or at a church or temple that they cross the threshold into having a determinate stance or commitment. I just don't think this happens very often. When it does, such apparent glimmers of determinacy are maybe superficial, reflecting little more than a vague familiarity or a shallow capacity for echoing some of the relevant terms and concepts without a deeper understanding. This is what David Moss (personal communication) observed when interviewing people about their metaethical views: they might seem, at first, to exhibit a determinate stance. But

---

[5] Gill (2009) refers to determinate but variable folk metaethical stances or commitments as *variability*. Since most researchers have opted for the term *pluralism*, I will make use of the latter. Personally, I think *variability* is a more apt term. Pluralism is often associated with an explicit commitment to multiple meanings, whereas variability strikes me as a more neutral, descriptive term.

scratch the surface, and the veneer of determinacy crumbles to reveal uncertainty, confusion, and a fumbling lack of facility with the relevant terms and concepts.

These fleeting instances of determinate variability qualify my case for indeterminacy, but do little to undermine the substance of my argument or the implications it has for philosophy and psychology.[6] In addition, traditional accounts differ in that they attempt to describe the *central* or *primary* function of ordinary language (Gill, 2009). I do not believe the argumentative and other social goals of moral language capture the *primary* or *central* use of ordinary moral thought and language, but are instead parasitic on such usage, which is *itself* indeterminate.

This might nevertheless appear to be a concession, and to some extent it is. I do not insist that *nobody ever* thinks or speaks in a determinate way. Rather, I claim only that for the most part, paradigmatic instances of moral thought and language are not primarily in the business of determinately expressing realist or antirealist stances or commitments. While *some* people likely have some stable and consistent stance or commitment to particular metaethical accounts, and while *some* moral utterances seem to clearly fit realism or antirealism, such determinacy is the exception, not the rule. And insofar as such determinacy does exist, there is no reason to believe it will be *uniform*. Rather, it will be highly variable and, in many cases, internally inconsistent (Colebrook, 2021; Gill, 2008; 2009; Loeb, 2008).

Even with this concession, I suspect my argument will strike many people as wildly implausible. Even those who express positions similar to mine are confident that elements of realism and antirealism clearly figure into folk metaethics. For instance, Loeb (2008) claims that "Mackie was surely correct in thinking that some sort of commitment to both objectivity and prescriptivity is built

---

[6] The notion that folk metaethics is constituted by some degree of both indeterminacy and variability was first articulated by Gill (2008; 2009), which he dubbed the Indeterminacy-Variability, or IV thesis. Notably, Gill explicitly presents the IV thesis as an *empirical* account of folk metaethics, that stands or falls on the evidence. My project is, at its heart, an effort to evaluate the IV thesis in light of the evidence currently available.

into our moral thought and talk" (p. 361). But, as I hope to show, the language of objectivity, relativity, and expressivism are *general* (that is, we use language to express relative and nonrelative propositional claims and to express nonpropositional attitude *in general*) and metanormative terminology sometimes piggybacks on moral discourse to serve social and argumentative goals (Mercier & Sperber, 2011). They aren't so much *built into* the semantics of moral language as they are external semantic drifters capable of fusing with moral language in particular conversational contexts to serve specific, local social goals, without being incorporated into the deeper commitments that characterize moral language *itself*.

My argument could also be misinterpreted as the claim that ordinary moral claims don't have any content or meaning. After all, If I deny that ordinary people are determinately committed to cognitivism, this might imply that people are not making truth claims at all when they make moral claims. Conversely, if I am denying that people are determinately committed to noncognitivism, I might be taken to deny that there is any expressive or imperatival content to ordinary discourse. Again, this is not the case. Indeterminacy about folk metaethics is *not* the claim that when a person says that "murder is wrong" that they are neither making a propositional claim nor only expressing a nonpropositional attitude. It is the claim that *neither* account necessarily captures the internal states or beliefs of that person (their stances) and that there is no evidence we might appeal to that would show one or the other of these accounts to do a better job of making sense of what that person is trying to do.

This is because the explananda we are trying to account for can be more or less equally well-accommodated by *either* realism or antirealism. It is not necessarily the case that neither account offers an adequate explanation, it could turn out that *both* do. Thus, I do not think realism and antirealism both equally fail to capture folk metaethics; we could just as readily say that both equally succeed. It would be better to say that I believe that, when the battle between them is settled, it will turn out to

be a draw: both can account for linguistic outputs about as well as one another, such that neither can claim to be the decisive victor.

**S1.2 Secondary criticisms of folk metaethics research**

Even if the central argument is mistaken, I have several secondary lines of argument that are independently defensible. First, existing research on metaethics is so methodologically flawed that it cannot be used to support determinate conclusions about folk stances and commitments, yet it typically *has* been interpreted to support determinate accounts. Most of this research suggests *metaethical pluralism*. While I acknowledge that there is good reason to believe there are identifiable (if rare) instances of both realism and antirealism among ordinary people, most of this apparent variability is likely an artifact of experimental design. This is due in part to the flaws and limitations in existing studies. As I argue here and in Bush and Moss (2020), that it is also due to the fact that such studies are not designed to detect indeterminacy and would appear to serve as evidence of variability *even if* the indeterminacy thesis is true.

If people do have determinate stances or commitments, it is unclear whether we currently possess any viable methods for discovering what they are, since folk philosophical research is limited by inherent difficulties in presenting questions in a way that ordinary people reliably interpret as researchers intend (Bush & Moss, 2020, Kauppinen, 2007). Even metaethicists that support empirical research on folk metaethics have balked at the use of surveys, anticipating that such methods face potentially insurmountable hurdles (Kauppinen, 2007; Moss, 2017; Moss & Bush, 2020). If these difficulties can be overcome, it will require alternative or more sophisticated methods than researchers have employed so far.

Finally, even if people do have measurable and determinate beliefs about metaethics, overreliance on the categories and distinctions of interest to contemporary analytic philosophy still represents a narrow and stultifying picture of folk philosophy that misses important features of ordinary thought, language, and behavior. At best, current research on folk philosophy should be broadened to include considerations outside the limited scope of the peculiar interests of academic philosophers. Why focus so much on whether people think moral claims are stance-dependent or stance-independent? And why not conduct bottom-up descriptive research that could potentially uncover patterns in folk metaethics that have little or nothing to do with traditional topics in metaethics? The way ordinary people think about the nature of morality may exhibit a host of fascinating and practically relevant patterns that aren't captured by their stance towards obscure philosophical positions.

These are not the only arguments I will present in this dissertation, but they provide a narrative center around which tertiary arguments will tend to gravitate. My general critique can be summarized as follows:

1.  Many researchers presume ordinary people have determinate philosophical *stances*: they endorse philosophical positions that roughly correspond to traditional philosophical distinctions.

2.  Many researchers presume ordinary people have determinate philosophical *commitments*: they speak and think in ways that fit traditional philosophical distinctions.

3.  Many of these distinctions do not figure into the way ordinary people speak or think (i.e., there are no determinate folk stances and commitments).

4.  This *indeterminacy* likely applies to folk moral realism and antirealism

5.  Even if I am mistaken, folk metaethics research is too flawed to support any determinate account of folk metaethics.

6.  Correcting methodological flaws with these studies is not adequate. Inherent difficulties in ensuring intended interpretations place our ability to study some features of folk philosophy outside the scope of what can be studied using the tools of conventional social scientific

methods. If people do have determinate metaethical stances and commitments, existing methods are insufficient to measure them.

7. Efforts to refine our tools in order to overcome these hurdles may result from a misplaced reliance on top-down attempts to coerce folk philosophy into traditional philosophical categories. Progress might be better served by redirecting efforts towards a more bottom-up, descriptive approach to folk philosophy.

**S1.3 Why I do not use the term "experimental philosophy"**

Some readers may wonder why I use the term *folk philosophy* rather than experimental philosophy. The main reason is to avoid engagement with metaphilosophical disputes over *intuitions*. However, there are a handful of other reasons I've opted to avoid using the term.

**S1.3.1 Experimental philosophy is best characterized as a social movement rather than a field with a specific subject matter**

*Experimental philosophy* is an interdisciplinary approach to exploring traditionally philosophical questions using methods typically employed in psychology (Knobe, 2016; Knobe et al., 2012; Knobe & Nichols, 2008; 2017). In practice, the term *experimental philosophy* typically describes empirical research conducted by philosophers, although psychologists and neuroscientists sometimes engage in similar research (Diaz, 2019; Theriault et al., 2020). Since there is considerable convergence in the assumptions, methods, and aims of both self-identified experimental philosophers and researchers studying folk philosophy, it might seem reasonable to describe anyone conducting research on folk philosophy as engaging in experimental philosophy, regardless of their academic credentials or the labels they identify with.

Yet in practice this doesn't appear to be the case. Instead, *experimental philosophy* tends to describe a social movement within philosophy that some people identify with and use to describe their work (and other people's work), rather than a term universally used to describe a specific kind of research. Much of the research I discuss was conducted by psychologists who would not refer to themselves as experimental philosophers and would not describe their research as experimental

philosophy, while there are other researchers using the same methods and studying the same topics who *would* call themselves experimental philosophers and *would* describe their work as experimental philosophy. Since I am only concerned with attempts to evaluate the stances and commitments of ordinary people, and have no interest in disputes about labels, I will refer to all such efforts as research on *folk philosophy*.

**S1.3.2 Experimental philosophy typically focuses on philosophical *intuitions*, which may differ from *stances* and *commitments***

I also avoid the term "experimental philosophy" because there is a difference between my characterization of research on *folk philosophy* and experimental philosophy: namely, that I emphasize the study of *stances* and *commitments* rather than *intuitions*. Conventional research in personality and social psychology explores an indiscriminate array of psychological states and mechanisms, often with no or at best weak conceptual distinctions: beliefs, attitudes, values, memories, perceptions, phenomenal states, etc. as well as behaviors and unconscious psychological processes. Psychology thus seeks to capture any measurable aspect of human thought and behavior of interest to researchers.

Experimental philosophy, on the other hand, is most closely associated with the much more narrow study of philosophical *intuitions*.[7] This characterization is so ubiquitous it is barely you might have to search for a description of experimental philosophy that *doesn't* associate it with the study of intuitions (which isn't to say it can't be done). Indeed, many accounts of experimental philosophy would *define* it as the study of philosophical intuitions or the psychological processes that produce philosophical intuitions. For instance, Knobe & Nichols (2017) state that "most research in experimental philosophy makes use of a collection of closely connected methods that in some way

---

[7] This is not to say experimental philosophers don't frequently discuss features of human psychology distinct from intuitions (Knobe, 2016). For some examples of the range and diversity of methods, approaches, and psychological states studied by experimental philosophers, see Fischer & Curtis (2019).

involve the study of *intuitions*." This characterization is typical (e.g., Alexander, 2012; Knobe et al, 2012; Sosa, 2007; Stich & Tobia, 2016).

Although it would be possible to construe what I have in mind by *stances* and *commitments* as "intuitions," this would be needlessly confusing. There are whole literatures dedicated to questions about what intuitions are, what their epistemic status is, and what role (if any) they play in philosophy.[8] I am not interested in entangling myself in these disputes because my argument doesn't turn on what intuitions are or what role they play (or should play) in philosophy. In addition, some critics have argued that that philosophy itself does generally rely on intuitions, that their use isn't central to contemporary philosophy (e.g., Cappelen, 2012; Deutsch, 2015; cf. Baz, 2015; 2017), and that it is a mistake for experimental philosophers to characterize their work as the study of philosophical *intuitions* as such (Cappelen 2014; Horvath & Koch, 2021; Machery, 2017). Adequately engaging with this topic would require a digression into metaphilosophy that would deviate so far from my objectives that it would have questionable relevance at best.

### S1.3.3 "Intuition" is an ambiguous and unhelpful term with no clear meaning

I also avoid the term *intuition* because there are ways of construing *intuitions* that are not subject to my critique. This contrast will become clear once I specify what I mean by philosophical *stances* and *commitments*:

(a) *Philosophical stances*: The philosophical beliefs endorsed by ordinary people.

(b) *Philosophical commitments:* The philosophical positions implicit in the way ordinary people speak, think, and act (independent of their belief in or awareness of these positions).

My conception of philosophical *stances* and *commitments* roughly maps onto the distinction Sinnott-Armstrong (2009) draws between *internal* and *external* descriptions of moral language:

---

[8] e.g., Baz (2015; 2017); Booth & Rowbottom (2014); Cappelen (2012, 2014); Chalmers (2014); Chudnoff (2013); Cohnitz & Häggqvist (2010); Deutsch (2009; 2010; 2015) Goldman (2007); Gopnik & Schwitzgebel (1998); Horvath & Koch (2021); Machery (2017); Pust (2017); Sosa, (2009); Talbot (2010).

There are two ways to describe moral language. An *internal project seeks to capture the psychological processes or representations that actually occur when people use moral language.* However, contemporary realists and expressivists are not trying to do that. When Jackson and Pettit use networks of truisms or when Gibbard cites hyperstates, they surely know that these theoretical constructions do not reflect actual psychological entities or events. Instead, they want their theories to be *externally adequate in capturing the outputs of our linguistic systems without necessarily reflecting the internal workings of that system.* In this respect, their project is more like Chomskian grammar, which uses constructs without claiming psychological reality. (p. 237, emphasis mine)

My conception of stances and commitments differs in that it encompasses *all* folk philosophy, not just morality. Second, the accounts Sinnott-Armstrong describes typically focus on developing externally adequate accounts of *language* or internally accurate accounts of the psychological states *associated with linguistic acts*. In principle, stances and commitments need not necessarily manifest in or be exclusively captured by *beliefs* and *linguistic practices*, but could also encompass nonlinguistic features of cognition and behavior (e.g., phenomenology, behavior, and so on). For instance, researchers might ask people if they have experiences (or *phenomenology)* that would be more consistent with realism or antirealism. Indeed, Zijlstra (2021) recently conducted a study evaluating folk metaethical phenomenology.[9] Nevertheless, my primary focus will be on linguistic commitments, and unless otherwise specified (or unless I forget) I will typically be describing linguistic commitments.

**S1.4 What are philosophical stances and commitments?**

***Philosophical stances***

A philosophical stance is a belief in the truth of a philosophical proposition, e.g., the belief that "there are stance-independent moral facts" is *true* or *false* (or *neither*[10]). As such, they represent an *internal*

---

[9] Zijlstra reports that 77.5% of participants said that moral disagreements "feel like factual disagreements" while only 22.5% stated that they feel like "matters of taste" (p. 8). This is not good evidence that most people are moral realists. Factual disagreements are consistent with relativism, constructivism, and relation-designating accounts (such as ideal observer theory). More generally, it's just unclear so shallow a measure could serve as robust evidence for a metaphysical thesis about the nature of morality.

[10] It could be "neither" if, for instance, a person believes that the concept of a stance-independent normative fact is unintelligible. It is therefore possible to have a determinate metaethical stance without necessarily regarding a particular claim as true or false: one could deny that the claim is sufficiently meaningful to be evaluated as true or false. This may

description of ordinary people's psychological states. Such beliefs do not require familiarity with a specific terminology, i.e., a person need not encode stances via specific linguistic representations, e.g., that "*'free will' is 'compatible' with 'determinism'*". For example, a person may believe that there are stance-independent facts about what people should and should not do. However, they may lack the vocabulary or conceptual sophistication to articulate this belief, e.g., they may not use the term "moral realism" or some cognate term in another language to refer to this belief. They must simply hold some belief in some propositional claim that is isomorphic with some meaning specified by the philosophical term or distinction of interest. A philosophical stance is a *belief* in the truth of a philosophical position a person either could articulate, or if they lacked the ability to articulate their belief, is the kind of belief they could correctly say, "Yes, that's what I thought all along!" if it were explained to them and they understood the explanation.

### Philosophical commitments

In contrast, philosophical *commitments* are not explicit beliefs in the truth or falsehood of propositions and don't have to map onto particular psychological states. Instead, to say that people are *committed* to a particular philosophical account is to say that this account provides an adequate description of the meaning of their speech or interpretation of their behavior *and* that this description is superior to any alternative.[11] A philosophical commitment does *not* depend on what people believe or how they would describe their thoughts and actions. For instance, a person could consistently think or act in a way best described by saying they are committed to *moral realism*, given a particular conception of what *moral realism* entails, even if that person has never considered moral realism or claims to believe it is

---

seem strange, but imagine there was an apparently serious dispute over whether "all square circles are angry." I don't think this is true or false. I think it's just gibberish. Rejecting philosophical disputes as predicated on fundamental misconceptions that render the whole dispute confused are not unheard of in philosophy, so positions like mine are not without precedent. See, for instance, Baz (2017).

[11] This latter condition is included to preclude the possibility of multiple, equally-adequate accounts. Under such conditions folk commitments would be indeterminate since there would be no principled reason to prefer one account over another.

false. People could even have the relevant conception of moral realism in mind, but could be mistaken about their own commitments or be unwilling to acknowledge their commitments. Although the latter possibility will vary given the philosophical topic in question, it should not be controversial to point out that people can be confused about their own beliefs to such a degree that they misidentify their own commitments, fail to understand the stipulated labels others use to refer to those commitments, or be unwilling for whatever reason to align their explicit position with commitments they themselves recognize that they have.

To further illustrate, suppose that a commitment to *free will* entails judging that *morally competent agents that commit a moral violation deserve to be punished, but agents who are not morally competent do not deserve to be punished*. Suppose you develop a sophisticated account of the conditions under which an agent is morally competent or incompetent, refer to this as an account of *free will*, and then go out and investigate how ordinary people speak. You discover that a particular person consistently judges that certain people deserve to be punished and others deserve not to be punished, and their pattern of judgments closely corresponds to your notion of *moral competence*. If so, then this person's judgments fit an external description of a "commitment to free will," regardless of the psychological states involved in this person's judgments or whether that person explicitly acknowledges or denies believing in "free will." To frame this in conventional philosophical terms, such a person is committed to a *compatibilist* account of free will[12], even if they have no particular beliefs at all about the existence or nonexistence of free will, or even explicitly deny compatibilism.[13]

*Commitment* could be interpreted as placing some epistemic burden on people to align their explicit beliefs with their implicit commitments, but inconsistency between stances and commitments

---

[12] e.g., Dennett (1984).

[13] Though the latter might constitute one output inconsistent with a description of their outputs as *compatibilist*. Taken as a whole, their outputs may still fit better with compatibilism overall. To say that a theory is adequate need not require that it perfectly accommodate all available data; such theories can plausibly claim adequacy despite some tolerable inconsistency or lacunae (see Gill, 2009; Sinnott-Armstrong, 2009).

does not necessarily reflect error or carry any other normative implications. For instance, a person could endorse moral antirealism but speak and act in a way that best fits some form of realism because it is useful to do so (Kalderon, 2005). This person's linguistic commitments would best fit a *realist* description, even if that person held an antirealist stance. Or someone could deny that moral claims can be true or false, but speak or act in a way that commits them to cognitivism. In other words, *commitments* refer to consistent patterns in how a person speaks, thinks, or acts that fit some philosophical account. Although these accounts are typically denoted by a particular terminological label, e.g., *realism*, *noncognitivism*, etc., the person who exhibits a commitment to these positions need not use that label themselves or have any explicit knowledge of that label, and may even *deny* that they exhibit the relevant sort of commitment. Commitments, then, are better described as accounts of patterns of speech, text, and behavior rather than descriptions of particular mental states (such as beliefs) or psychological processes.

### Distinguishing stances and commitments from intuitions

Although my descriptions of *stances* and *commitments* are somewhat underdeveloped, it is still important to offer some account of them because claims about one may not apply to the other, and because they are distinct from some conceptions of *intuitions*. Philosophical intuitions are sometimes described as beliefs (Lewis, 1983, p. x; Pust, 2017). In such cases, "intuitions" may refer to what I mean by *stances*. Yet some philosophers also use the term "intuition" to refer to *dispositions* to believe certain propositions are true, in which case such intuitions are not beliefs, and are thus not *stances*. For instance, van Inwagen (1997) claims that in at least some cases, intuitions may be "the tendencies that make certain beliefs attractive to us, that 'move' us in the direction of accepting certain propositions without taking us all the way to acceptance" (p. 309, as quoted in Pust, 2017). In other words, an intuition may simply be an inclination towards holding a belief, but not an instance of holding that belief. A utilitarian might feel a strong attraction to the notion that it is wrong to kill someone even if

doing so would save dozens of lives. But in spite of this attraction to the wrongness of killing, they may nevertheless judge that killing someone is morally required if it would maximize utility. Such a person may acknowledge that they have the *intuition* that it is wrong to kill one person to save a greater number of people even if they do not believe it is wrong (and thus don't hold the *stance* that it is wrong).

This strikes me as a perfectly sensible way to talk about intuitions. People (or at least philosophers) really do sometimes say things like, "Although I have the intuition that it's bad to kill one man to save five lives, I still think it is morally permissible to do so" or, "Although I have the intuition that some people deserve to suffer, I resist this impulse because I recognize it isn't consistent with my moral principles." More generally, philosophers often explicitly maintain that while they find a proposition intuitive, they reject it because it is inconsistent with their theoretical commitments (Climenhaga, 2018). We can distinguish between the inclination to hold a belief and holding that belief, and philosophers do appear to use "intuition" to refer to defeasible inclinations to hold beliefs, since they frequently treat intuitions as evidence for a philosophical position, but not necessarily *decisive* evidence, and that as a result, the intuition that a given claim is true does not necessarily entail the belief that it is true.[14,15]

One potential objection to my argument for indeterminacy can appeal to this notion of intuitions as *inclinations to believe* and argue that I am misinterpreting the purpose of (at least some) relevant research on folk philosophy. Perhaps this research is simply intended to investigate which philosophical positions people are intuitively disposed to endorse. Evidence that a person is disinclined to push the fat man off the bridge in the trolley dilemma does not necessarily tell us about

---

[14] Intuitions may also be characterized as *suis generis* states with distinctive phenomenological qualities (Pust, 2017). Such accounts may or may not overlap with my characterization of stances and commitments but do not seem relevant to the arguments presented here.

[15] However, it is also possible that *some* intuitions are incorrigible. In that case, intuitions may capture inclinations to hold beliefs that are for some reason incapable of being overridden.

their philosophical stances or commitments, but perhaps *it isn't intended to.* Researchers could use findings about intuitive dispositions to support or oppose philosophical positions. Greene has done exactly this by arguing that the psychological processes involved in deontological intuitions are less reliable than the processes involved in utilitarian intuitions (Greene, 2003; 2008; 2014a; 2014b). Or perhaps the aim of collecting data about intuitive dispositions could be to provide insight into the psychological processes involved in folk judgment and cognition. If so, the philosophical aspect of these studies may be little more than window dressing for psychology as usual. Josh Knobe claims that this is the primary purpose of most experimental philosophy. According to Knobe (2016):

> The majority of experimental philosophy papers are doing *cognitive science.* As such, they are doing precisely the sorts of things one would expect cognitive science papers to do. They are revealing surprising new effects and then offering explanations [sic] those effects in terms of certain underlying cognitive processes. (p. 39)

Indeterminacy is only concerned with what people believe and what philosophical accounts best explain how they speak. Empirical discoveries about what people are *inclined* to believe are important and psychologically interesting, but are irrelevant to indeterminacy about features of folk philosophy. This is because it is possible for people to have an inclination to believe philosophical accounts even if they don't believe those accounts or speak and act in ways that don't reveal a distinctive commitment to a philosophical account consistent with their intuitive inclinations. If research on folk metaethics was only concerned with intuitive dispositions, then my argument for indeterminacy would be misplaced, since I'd be arguing that research on folk metaethics fails to show something that it isn't intended to show in the first place. Researchers studying folk philosophy could claim that their research is only intended to measure ordinary people's intuitive dispositions rather than their stances and commitments. If so, such research is unrelated to efforts to describe the stances and commitments and is not subject to my case for indeterminacy.

**S1.5 Folk metaethics is about stances and commitments**

Plenty of research may fit this description, but this is not the target of my critique. With respect to folk metaethics, numerous researchers have implied or explicitly stated that they are attempting to capture mental states, psychological processes, or linguistic practices that correspond to my description of stances or commitments. Given that the distinction between the two is not recognized in the literature, it is not always clear whether researchers have interpreted their findings as evidence for one or the other. Even so, most researchers have interpreted their findings as evidence of what people believe or how they speak, rather than how they are *disposed* to think. A handful of representative remarks will hopefully suffice for demonstrating that researchers have interpreted their findings as evidence of folk stances or commitments *not* intuitive dispositions. In their seminal article on folk metaethics, Goodwin and Darley (2008) state that:

> In this paper, we develop a method that distinguishes ethical objectivists (i.e., *individuals who take their ethical beliefs to express true facts about the world*) from ethical subjectivists (i.e., *individuals who take their ethical beliefs to be mind-dependent, and to express nothing more than facts about human psychology*). (pp. 1357-1358, emphasis mine)

Given that their participants are ordinary people, and they are described as individuals that "take their ethical beliefs to express" either "true facts about the world" or "nothing more than facts about human psychology," they would appear to be capturing folk *stances*, since G&D appear to be presenting an internalist account of what people take themselves to mean.

Zijlstra (2019) presents a set of measures intended to measure "folk moral objectivism," and attempts to develop a scale that attempts to map established metaethical distinctions to corresponding psychological constructs. Likewise, Collier-Spruel et al. (2019) constructed a scale that attempts to capture the degree to which ordinary people subscribe to moral relativism and likewise construe belief in relativism as a distinct psychological construct. Finally, Wright claims that:

> [...] meta-ethical pluralism exists—and that the empirical scholarship showing that people are both realists and anti-realists cannot be simply dismissed on the basis of being philosophically

inadequate, because even when we increase the level of clarity and rigor, the pluralism clearly remains. (p. 144)

None of these studies appear to be evaluating the metaethical distinctions people are merely *inclined* to believe. Rather they unambiguously attempt to capture the stances or commitments ordinary people *already held prior to participating in the study*.[16] Across numerous studies exploring folk metaethics, participants are described as realists or antirealists, not merely people disposed to endorse one of these positions or are merely inclined to endorse a particular metaethical account or that *would* find them intuitive *if* they were to reflect on the matter.

These remarks capture the general tenor of research on folk metaethics. These studies do not seem especially concerned with only capturing intuitive dispositions, nor are they merely using questions about philosophy to explore general psychological processes. Rather, they primarily aim to describe ordinary thought and language. But just what is ordinary thought and language, and who are ordinary people?

## S1.6 What are ordinary people?

Implying that "the average family has 2.4 children" is a literal description of actual families may induce no more than the ghost of a chuckle, but it is still recognizably a joke. We recognize that statistical abstractions don't pick out individual instances, nor are they intended to, yet they still provide us with valuable information about the world. Just as there are no actual cases of the proud parents of two ordinary children and the tottering torso of their headless 40% of a sibling, there are also no *actual* instances of ordinary people, headless or otherwise.

This might sound like a strange claim to make for a paper purporting to discuss ordinary people, but it should not be. An *ordinary person* is an abstract, idealized individual that lacks formal

---

[16] Note that it is often unclear whether researchers intend to describe stances or commitments in particular. This is not surprising given that this distinction is not in common use and is not necessarily a fault with the articles referenced here.

training in philosophy and who has not engaged in deviant levels of introspection about how they or members of their community think and speak. Ordinary people are thus little more than fictional receptacles for *ordinary thought* and *ordinary languag*e. Ordinary thought and language, in turn, are the *pretheoretical* stances and commitments of ordinary people. That is, they are idealized descriptions of how people would think and speak in a perfect atheoretical vacuum.[17]

But of course, nobody actually lives in a perfect atheoretical vacuum. While it would be convenient to define *ordinary people* as "nonphilosophers" or "anyone that lacks significant education in philosophy," philosophers are not the only people that have ever stopped to think about the great profundities of life. Recurring tropes in movies, books on pop philosophy, and every student essay since the dawn of time testify to the intrusion of philosophy into everyday thought. Over 35% of Americans have college degrees (United States Census Bureau, 2022), and many of them have taken a course or two in philosophy or encountered the trolley problem. And with most people belonging to some religion, political tribe, or social movement replete with dictums and dogmas and slogans, few have managed to escape exposure to philosophical ideas.

Given all this exposure to ambient philosophy, we might worry that most people are not very ordinary. Are Mormons ordinary? What about vegans, students in debate clubs, or people that go to Burning Man every year? What about people that have read Jordan Peterson's *12 Rules for Life* (2018)? Surely there's nothing ordinary about how clean their rooms are. And that's just it: no actual people are ordinary. This is why it only makes sense to construe the "ordinary person" as a statistical composite, a useful fiction intended to capture how all the people who aren't publishing in academic philosophy think and speak about the topics academic philosophers think and speak about.

---

[17] For my purposes, *the folk* is interchangeable with *ordinary people*, while *folk thought* and *folk language* are interchangeable with ordinary thought and ordinary language, respectively.

One concern with this construal is that ordinary thought and language is far too heterogeneous. This is because we may be tempted to describe ordinary thought as the typical beliefs, attitudes, and psychological processes involved in the kinds of judgments relevant to philosophy, and ordinary language as what people typically mean when they say things like "Alex knows where it is" or "Sam thinks she is a terrible person." Surely there must be enough overlap in what people think and mean for there to be some typical set of stances and commitments.

Undoubtedly, people do think in similar ways and mean much the same thing when they talk. Yet it would be question-begging to presume that, *with respect to a given philosophical distinction*, that there *is* or *must* be some stance or commitments shared by all (or most) ordinary people. One must furnish evidence of this fact, since to do otherwise is to simply presume that there is a uniform and determinate way people think and speak that decisively favors a particular philosophical account without any evidence that this is the case and when this is the very position I am rejecting.

Although widespread heterogeneity in meaning may play havoc with the desire for tidy, uniform accounts of folk philosophy, ruling out pluralism is not necessary for characterizing ordinary thought and language. It could turn out that folk philosophy is far messier than we supposed, and that pluralism characterizes some aspects of folk philosophy. With respect to metaethics, perhaps some people speak or think like realists and others like antirealists, and neither is more central or characteristic of ordinary folk metaethics (Gill, 2009). While there is no reason why this couldn't be true in principle, there may be good reasons to be skeptical of metaethical pluralism in practice (Johansson & Olson, 2015; Sinnott-Armstrong, 2009). Nevertheless, *metaethical pluralism* is a very real possibility that threatens indeterminacy insofar as it represents an alternative explanation for existing empirical data on the folk metaethics, and some have interpreted existing evidence as convincingly establishing pluralism (Davis, 2021; Hopster, 2019; Pölzler, 2017; Pölzler & Wright, 2020a; 2020b; Wright, Grandjean, & McWhite, 2013).

The potential conflict between *folk indeterminacy* and *folk pluralism* is a difficult needle to thread and a challenge I address to some extent in the main text. For now, it is relevant only insofar as it raises challenges for specifying what is meant by ordinary people, thought, and language. If there is no *uniform* folk conception of traditionally philosophical topics, some shared set of stances and commitments that distinguish the folk from philosophers, then in what sense is there *a* folk view? That is, if there are no *typical* stances and commitments that characterize ordinary thought and language, then in what respect are there any ordinary people?

One possibility is to reject attempts to distinguish ordinary people by the ostensibly shared *content* of their stances and commitments, and to instead distinguish them by some other standard, e.g., by some sociological, psychological, or epistemic difference between ordinary people and whoever we are distinguishing them from (e.g., *philosophers*). For instance, perhaps one way of identifying the stances and commitments of ordinary people might appeal to my initial description of ordinary thought and language: that it is in some sense *pretheoretical*. This is questionable, since it is unclear how pretheoretical ordinary people are.

A glib response to this challenge is that this isn't *my* problem. If there is no satisfying way to specify what ordinary thought and language is because there is no ordinary thought and language, then research on folk philosophy is in even more dire a situation than my objections would suggest, since we would now be in doubt as to whether there even are ordinary people whose stances and commitments would characterize folk philosophy.

A less glib response is to return to my initial definition of an ordinary person as a useful fiction. In the real world, there are no ordinary people, but there are certainly people whose way of thinking and manner of speech have been *less* subject to the corrupting influence of academic philosophy, and we must simply do our best to extrapolate from samples of these people to form a picture of ordinary thought and language. And in practice, we must call on *actual people* to participate in our studies. If

there are no ordinary people, then who are these people, and why are they appropriate subjects of study? Again, the glib reply would be "That's not *my* problem." If there are no appropriate methods for studying folk philosophy, then the content of folk philosophy is indeterminable. Perhaps someone will present a compelling case that the whole notion of studying folk philosophy is an unintelligible error of fantastic proportions. If so, that would only make my case even stronger. This seems overly pessimistic even to me. So, for the sake of charity, let us attempt to rescue folk philosophy from the ashes of some future dissertation.

Some people are at least less exposed to academic philosophy than people who have degrees in philosophy or identify as philosophers or publish philosophical work in academic journals. Even if many of these people do not have completely pretheoretical views, nonphilosophers exist along a spectrum from the utterly unreflective to the enlightened autodidact, and it is likely far more people resemble the former than the latter. However heterogeneous they are, however exposed to theory they may be, we can still place people along a continuum of those more or less familiar with the explicit conceptual arguments and distinctions discussed in academic philosophy, or who have developed similarly sophisticated accounts on their own.

In practice, this will require us to offer some operationalization of *ordinary people* for the purposes of conducting studies. This may mean only including participants that do not have degrees in philosophy, do not identify as philosophers, or who claim to be unfamiliar with the topic of study.[18] While they may not be ordinary people in a deeply satisfying respect, drawing on such samples should suffice. And why shouldn't it? We don't conclude that there are no relevant distinctions between how musicians and non-musicians think about music simply because non-musicians have different

---

[18] The latter operationalization reminds us just why the notion of an ordinary *person* is a fiction. In principle, a person could be very familiar with one philosophical topic but completely ignorant of another. Such a person is ordinary with respect to one issue but not another. In what respect, then, are they an ordinary person in some general sense? Ideally, we would judge a person's perspective on various topics piecemeal, rather than treating everyone as being ordinary or not *simpliciter*.

preferences from one another. And it would be absurd to conclude that there is no substantive distinction between musicians and non-musicians by pointing out that most people half-heartedly played a musical instrument in grade school or once drunkenly sang "Bad Romance" at a karaoke bar. In short, we can still distinguish non-musicians from musicians according to reasonable operational guidelines even if there are few people with no exposure to music and even if non-musicians differ markedly from one another in their attitudes about and experiences of music.

And so we come full circle. Earlier, I dismissed the notion that we could describe ordinary people as "nonphilosophers." Perhaps someone will raise good objections to this characterization. But if nothing else, there is a meaningful distinction between people who have had significant exposure to academic philosophy and those who haven't, and we can make meaningful discoveries about the differences between these two groups *even if* some of "ordinary people" have casual conversations to stumble onto homebrewed notions of representationalism or panpsychism or some other philosophical position. For the purposes of delineating ordinary people from philosophers, this distinction should be adequate: ordinary people are *people without significant exposure to terms, methods, and concepts distinctive to academic philosophy*.

It is fortunate I'm not obliged to present a less contentious account of *ordinary people*, since I face a steeper challenge. If people really do have all this exposure to philosophy, then why propose that folk philosophy is (at least in some cases) indeterminate? Why not suspect varied but determinate philosophical stances and commitments? That is, why not embrace *folk pluralism* instead of *folk indeterminacy*? This is a good question because most studies on metaethics at least superficially suggest folk pluralism. There are at least three ways to interpret this evidence:

(1) *Accept it at face value*. Pluralism seems to be true because it is true (e.g., Davis, 2021; Hopster, 2019; Pölzler, 2017; Pölzler & Wright, 2020a; 2020b, Wright, Grandjean, & McWhite, 2013).

(2) *Argue for an error theory.* There is systematic error in one or more patterns of response. Once accounted for, ordinary people share uniform and determinate stances/commitments (Beebe, 2020, Sarkissian et al., 2011).

(3) *Argue for indeterminacy* (or agnosticism). Apparent variability is an artifact of experimental design and does not reflect genuine pluralism in ordinary thought and language (Bush & Moss, 2020).

The first is the most straightforward interpretation and seems to be widely accepted among researchers. Given the strength of the evidence for (1), anyone who wants to present a case for (2) or (3) has their work cut out. Not surprisingly, I will present a case for (3). However, there are good arguments for all three views and, at present, there is no decisive evidence in favor of any of these perspectives. I also recognize that *some* people without philosophical training have determinate and varied stances and commitments. Just as philosophers clearly exhibit a variety of perspectives, at least some nonphilosophers have thought enough about the nature of morality to have an identifiable perspective. Yet it is unclear whether such determinate stances and commitments are common enough to serve as evidence for the contents of ordinary thought and language, or if such people are the very aberrations such data would ideally exclude. I would be content to merely demonstrate that such people are a minority among their peers, and that just as *most* people don't endorse a particular interpretation of quantum physics, nor do they speak in a way that best fits any particular, ordinary people likewise lack firm and determinate stances about metaethics and free will.

**S1.7 What is ordinary language and thought?**

***Ordinary language***

By *ordinary language*, I simply mean language as it is used by ordinary people. I eschew taking any substantive philosophical stance on language for the purposes of this paper. However, I do have operating assumptions (admittedly underdeveloped) about the nature of language. As a result, some of the arguments, assumptions, and claims made here may, to someone with a different conception of how language works, appear to be in tension with philosophical positions, e.g., semantic externalism

(Kallestrup, 2013; Lassiter, 2008; Wikforss, 2008). This assumption may work its way into my case for folk indeterminacy in ways that are not apparent. If so, hopefully flagging this possibility will render such concerns salient or motivate an enterprising specialist in philosophy of language to consider whether I'd need to add any caveats or qualifiers, or defend any particular philosophical assumptions that slipped into my work.

Note that "the way language is used by ordinary people" should *not* be taken to imply that there is a single way people use moral language. That would be unhelpfully vague. But I also suspect it would be inaccurate at any level of deeper specification beyond the basic and universal features of human languages (whatever those are). I suspect language can and does vary wildly within and between populations. For instance, researchers studying the Pirahã report such significant differences between their language (e.g., Frank et al., 2008; Everett, 1983; 1986; 2005; 2008; 2009; 2012) and most other languages that these alleged differences have become a matter of considerable controversy (Nevins, Pesetsky, & Rodrigues, 2009), with some critics responding with incredulity (Bower, 2005). The Pirahã are a small, secluded society. Their language has evolved in isolation for so long, and there are so few existing language groups related to it, that it may have diverged dramatically from other languages. I cannot meaningfully weigh in on these possibilities, since I lack the requisite training and knowledge. Nevertheless, the Pirahã language at the very least provides a glimpse into the *possibility* of significant linguistic variation, a possibility I am more than open to: I confess an enthusiasm and hope that claims about wildly divergent linguistic features among the Pirahã and other human societies are vindicated.

**Ordinary thought**

*Ordinary thought* is a bit harder to pin down than ordinary language. Roughly, it is simply the thoughts ordinary people have. That is, in the absence of significant exposure to academic philosophy, ordinary thought captures the psychological processes and mental states that characterize the way people tend to think. Like ordinary language, there need not be *one* way ordinary people think. Indeed, there are

undoubtedly numerous psychological differences between people. What distinguishes ordinary thought from philosophical thought is that the former reflects *any* psychological processes or ways of thinking that are not caused by or distinctive to academic philosophy. Nonphilosophers undoubtedly think about the nature of the world, the existence of God, art, morality, and many other topics that interest philosophers. Yet academic philosophy is replete with a host of distinctive terms, concepts, categories, distinctions, and norms that distinguish those who engage in it from others. In addition, academic philosophers tend to study a shared canon of thinkers, and thus not only share methods and ways of thinking, but substantive and highly overlapping knowledge of the writing and thoughts of a particular array of thinkers. Finally, academic training in philosophy *can* occur in relative isolation, but in practice often (if not typically) involves at least some (and often a great deal) of social interaction with other people who study philosophy. Those who receive formal training or enter into philosophical community via conferences, degree programs, online discussions, and so are inducted into a particular philosophical community that further reinforces the distinctive modes of thinking characteristic of such communities. Taken together, the cumulative effect of philosophical thinking and socialization may result in ways of thinking distinctive to the study of academic philosophy.

**S1.8 What is metaethics?**

*Folk metaethics* is a subfield of folk philosophy dedicated to the study of ordinary people's *metaethical* stances and commitments. But what is *metaethics*? Moral philosophers often distinguish three main areas of moral philosophy: *applied ethics*, *normative ethics*, and *metaethics* (Kagan, 1997; Wolff, 2018). As its name suggests, *applied ethics* deals with the moral evaluation of *specific* practical moral issues that are relevant to our public institutions and personal lives, e.g., the ethics of abortion, capital punishment, how we treat animals and the environment, etc.

     *Normative ethics* likewise deals with concrete questions about what is right and wrong, but seeks to develop a more general account of what makes actions morally right or wrong, which character

traits are morally good or bad, and so on (Kagan, 1997). For instance, *consequentialists* maintain that the fundamental locus of moral concern is with the consequences of our actions, and that what makes an action morally good or bad depends on the *outcome* of that action. Deontologists, in contrast, argue that actions are right or wrong in virtue of their conformity to *duties*, and that the fundamental locus of moral concern rests with actions themselves. For instance, lying may not be wrong because it tends to produce negative outcomes, but because we have a duty to abstain from lying.

*Metaethics* differs from both applied and normative ethics in that it does not attempt to address substantive questions about the moral status of principles, actions, or character traits, but instead addresses *fundamental questions about the nature of morality* (van Roojen, 2015). Metaethical questions address a broad range of issues related to the meaning of moral language, the metaphysical status of moral facts, whether and how we can acquire moral knowledge, and the relation between moral and nonmoral facts (Sayre-McCord, 2012). For instance, metaethics centers on questions such as:

(1) Are there moral facts? If so, what kinds of facts are they?

(2) Can we acquire moral knowledge? If so, how?

(3) When people state that an action is morally wrong, are they making a propositional claim or only expressing a nonpropositional attitude?

(4) What is the relationship between moral judgment and motivation?

Roughly speaking, applied ethics addresses *concrete* and *specific* moral questions, normative ethics addresses *concrete* but *general* moral questions, and metaethics deals with *abstract* and *general* questions about the nature of morality.

### S1.9 What is folk metaethics?

Given these distinctions, we may now ask what *folk metaethics* is. Folk metaethics is not concerned with directly addressing metaethical questions, i.e., it does not ask whether there are moral facts, or how we might acquire moral knowledge. Instead, it addresses how ordinary people *think* about these

questions and how they use moral language in everyday interactions. Folk metaethics is thus a *descriptive* enterprise whose central purpose is to catalog ordinary people's metaethical *stances* and *commitments*. For instance, instead of asking whether there are moral facts, folk metaethics may ask whether people *believe* there are moral facts, and if so, what *kinds* of facts they think moral facts are (i.e., do they think these facts are relative or nonrelative).

Folk metaethics need not concern itself with explicit folk metaethical beliefs, but could also assess the nature of ordinary moral discourse. Much as languages possess a grammatical structure that native speakers may be oblivious to, ordinary moral thought and language may contain implicit commitments to particular metaethical accounts even if ordinary people are unaware of these commitments and are incapable of explicitly reporting them (Sinnott-Armstrong, 2009). In other words, people may employ moral language in a way that reliably conforms to patterns best described by a realist or antirealist framework, even if such usage is not introspectively accessible and even if the speaker has no explicit philosophical stance in much the way Nisbett and Wilson have argued that we may generally lack introspective access to the underlying psychological processes associated with our judgments and behavior (Nisbett & Wilson, 1977). In such cases, researchers may wish to determine whether ordinary moral language fits a particular semantic account better than any alternatives, e.g., that the meaning of sentences like "murder is wrong" best fits a cognitivist or noncognitivist analysis, independent of people's metaethical stances.

Folk metaethics is not intended to prescribe behavior, recommend how we ought to think about moral issues, or directly address questions in metaethics or ethics in general. Given its descriptive nature, we might be tempted to characterize folk metaethics as a thoroughly empirical discipline. While this may seem tempting, it seems to me there is still an important role for armchair philosophers to play in discussing the status of folk metaethics even if they do not directly conduct empirical research. After all, the same could be said of the role of philosophers with respect to physics

or biology. Insofar as there are plausible grounds for believing that philosophers of science have something to offer other scientific endeavors, so too might metaethicists, philosophers of language, and philosophers of psychology have something to offer to questions about the nature of folk metaethics, even if many of its central claims are empirical. In short, while we might be tempted to hand folk metaethics entirely over to science, we are justified in resisting so long as we believe philosophy has some relevant auxiliary role to play in the sciences.

**S1.10 What are moral realism and moral antirealism?**

This is not a treatise on the nuances of all the various ways one might frame moral realism and antirealism. I would still like to offer some additional discussion about how I use these terms. There is considerable inconsistency and lack of clarity in much of the academic literature on metaethics on precisely how to describe moral realism and antirealism. This is regrettable, but it is an unsurprising feature of the literature that is typical of many disputes in academic philosophy. Philosophers appear to be unwilling or unable to agree on and share a precise set of terms and to commit to using them consistently.

Nevertheless, there are a few general features that moral realism and antirealism share in common across different descriptions and accounts. First, virtually all accounts of moral realism describe the position as, at the very least, the claim that *there are at least some moral truths*. However, there are at least two ways this claim is immediately complicated. First, moral realism is presented as a cluster of claims, one of which is a semantic claim about the meaning of moral claims. We may call this a *semantic thesis*, and this conception of moral realism holds that *moral claims express propositions about what is morally right or wrong* and *at least one of these claims is true*. I find this to be a strange way to describe moral realism, since it seems to make the existence of moral facts contingent on ordinary moral language. While there may be particular accounts of the philosophy of language where this wouldn't be a

problem, it is strange to frame moral realism in such a way that it *requires* substantive philosophical commitments to particular, contestable positions in the philosophy of language.

Suppose, for instance, that very persuasive moral noncognitivists convince all people on earth of moral noncognitivism, and after thousands of years, our languages evolve in such a way that people continue to use what they regard as moral language, but they use it to convey nonpropositional attitudes. Thus, they are no longer ever intending to make propositional claims about what is morally right or wrong. Would we insist that these people are not making moral claims at all? That's one way we could react to this situation. Another response would be to say that they are using moral language, but that they are not using it in a noncognitivistic way. We might say, in this future world, that as a matter of descriptive fact, moral discourse is noncognitivistic: it does not involve assertions about what is morally right or wrong. If we accepted this, would it follow that therefore there are no moral facts? I don't see why. Our external reality doesn't conform to how any given population of people happen to speak.

I am not suggesting that we would have to agree that these people are engaged in moral discourse. What I am claiming is that it is a mistake to insist that realism *requires* you to deny that noncognitivism is the correct descriptive account of folk metaethics. It is a mistake, in other words, to embed substantive philosophical views about the nature of language, or specific descriptive theses about how nonphilosophers speak or think, into the concept of realism itself. I am not familiar with this point being expressed and defended in contemporary metaethics with any frequency (though that doesn't mean it hasn't been). The only instance I know of where a philosopher explicitly argues that moral realism doesn't require a semantic claim is Kahane (2013) in an article aptly titled *Must metaethical realism make a semantic claim?* Kahane's answer is "no." This is strange, and hints at what may be a pervasive problem with the way metaethical positions are framed.

A second problem with the way moral realism is described is that it is sometimes unclear whether it is merely the claim that there are moral truths (regardless of what makes them true) or whether only certain kinds of truths qualify (e.g., stance-independent truths). Absent the latter condition, moral realism would merely consist of the claim that there is at least one moral truth. One concern with this characterization is that it is consistent with stance-dependent truths. On this view, versions of relativism and relation-designating accounts such as ideal observer theory would count as realist positions. We *could* draw the dividing lines in this way, but it strikes me as unsatisfying to include subjectivists with realists. Unfortunately, one of the most popular online resources for philosophy, the Stanford Encyclopedia of Philosophy, is unhelpful regarding this issue. The author for the entry on moral realism, Sayre-McCord (2015), begins by describing realism as follows:

> Moral realists are those who think that, in these respects, things should be taken at face value—moral claims do purport to report facts and are true if they get the facts right. Moreover, they hold, at least some moral claims actually are true.

This is deeply unhelpful. No mention is made of whether the facts in question must be stance-independent. I am genuinely unsure whether Sayre-McCord would characterize realism as the claim that at least some moral claims actually are *stance-independently* true or not. I would like to register my official request that Sayre-McCord resolve this ambiguity by clarifying the entry to address this concern. Note below that Sinnott-Armstrong (2009) is explicit: "I do not count subjectivism or cultural relativism as a kind of realism" (p. 236). If this weren't an open question, though, it would be odd to explicitly say this. Thus, the very fact that Sinnott-Armstrong makes an active effort to say so pragmatically implies that this is something others might dispute or construe differently.

If all this seems overly pedantic to the nonphilosopher, I have bad news: this brief commentary merely scratches the surface. It would be more accurate to describe *moral realism* as a term that references a bundle of claims that frequently cohere, and for which there is no agreement (or even much discussion) about which, if any, of the features involved are necessary for a position to "count"

as a realist position. Take Sinnott-Armstrong's (2009) description of realism. According to Sinnott-Armstrong, "Moral realists [...] claim a bundle of theses on many levels. In particular, the complete package for moral realism contains *at least* these five theses" (p. 235, emphasis mine). Note that Sinnott-Armstrong isn't even limiting realism to the five theses he presents! One can almost imagine the phantom voice of the realist providing yet another feature of their view, only to say "...But wait! There's more!" Here are the five features Sinnott-Armstrong attributes to realism:

(1) Metaphysical Thesis: There are some objective moral facts.

(2) Semantic Thesis: Moral statements are true if and only if they correspond to objective moral facts.

(3) Alethic Thesis: Some (positive) moral statements are true.

(4) Epistemic Thesis: We can and often do know some objective moral facts.

(5) Pragmatic Thesis: Moral statements (try to) describe objective moral facts. (p. 235)

Sinnott-Armstrong adds that this doesn't even exhaust the bundle of features the realist might endorse, adding that they might also assert that "moral statements express beliefs" (p. 236). We're left, then, with moral realism consisting of at least five claims, and possibly more. Yet even this description will not suffice, since some realists would reject one or more of these features while still insisting they're a realist. We've already seen that Kahane (2013) denies that moral realism must make a semantic claim. This is an especially fascinating exclusion, given the focus of Sinnott-Armstrong's discussion. After describing the various commitments of realism, he says that he "will focus on the semantic theses and, to some extent, the pragmatic theses" (p. 236; here he is referring to both realism and expressivism, a term that has started to supplant "noncognitivism"—yet another inconsistency in the terms used in metaethics). It is fascinating because the semantic thesis may be seen as one of the central tenets of realism, and yet some realists dispense with it entirely. Another equally strong candidate for a central pillar of realism is the metaphysical thesis. One of the most influential and beloved moral realists of the 20th century, Derek Parfit, explicitly denied that realism needs to make any substantive

metaphysical claims, as does Scanlon, another prominent realist (Veluwenkamp, 2017). It's not clear that realists must necessarily be committed to *any* of the central theses. A realist might, for instance, claim that there are (or could be) stance-independent moral facts, but that we have no epistemic access to them. Unfortunately, there is simply no established consensus on what a "moral realist" has to be.

Unfortunately, semantic claims seem to be so entrenched a feature of the way realism is framed that it is difficult to extricate oneself from ways of describing realism that cash it out in largely semantic terms. Consider how Sayre-McCord (2015) frames opposition to moral realism:

> As a result, those who reject moral realism are usefully divided into (i) those who think moral claims do not purport to report facts in light of which they are true or false (noncognitivists) and (ii) those who think that moral claims do carry this purport but deny that any moral claims are actually true (error theorists).

As a matter of historical description, this is probably accurate: most antirealists do fall into one of these categories. However, these are not the only categories available to antirealists in principle.[19] I'm a moral antirealist, yet I do not fit into either of these categories. If someone were to ask if I think moral claims "purport to report facts in light of which they are true or false" how could I possibly respond? *Which* moral claims? *All* of them? The question presupposes that there is a single, uniform, categorical, and determinate fact of the matter about whether *all* moral claims purport to report facts or not. If I had to fill out a form answering this question, I'd have to skip the question or write "N/A" in the margins.[20] I don't think there is a uniform and determinate account of folk metaethics. I think Gill (2009) is correct that such claims presume the uniformity and determinacy of folk moral discourse. And I think this assumption is false. So the division between cognitivists and noncognitivists rests on

---

[19] Hopefully the SEP is updated to include positions that fall outside the scope of narrow, semantic-focused accounts of realism and antirealism (especially those that presume uniformity and determinacy; see Gill, 2009), so that positions like mine (along with Gill and Loeb) can be recognized and included in the taxonomy of available metaethical positions.
[20] I have, in fact, confronted just this problem when attempting to answer questions for the PhilPapers survey: for many of the questions, I cannot provide a categorical answer to a question about which side of an argument I agree with because I reject the entire framework on which these distinctions are predicated.

what I take to be a false dichotomy: I *cannot* be classified according to this distinction, any more than I could tell you whether unicorns like or dislike pineapple on pizza.

## S1.11 What is indeterminacy?

My use of the term *indeterminacy* can be traced to Gill's (2009) use of the term. Gill does not provide a highly technical or detailed description of what he means by *indeterminacy*, nor do I. As such, the term remains somewhat underdeveloped in the context of metaethics. Gill does provide some explanation of what he means:

> The Indeterminacy Thesis holds that some parts of ordinary moral discourse give us no reason to prefer an analysis that involves one meta-ethical commitment over an analysis that involves the commitment that has traditionally been taken to be its meta-ethical competitor. (p. 216)

This represents Gill's initial sketch, but he goes on to provide a more thorough explanation of what he means:

> According to the Indeterminacy Thesis (which is the "I" of the IV Thesis), many parts of our moral thought and language provide no good answers to the questions that were central to much of 20th century meta-ethics, vindicating neither relativism nor absolutism, neither internalism nor externalism, etc. The Indeterminacy Thesis holds that the relationship between some instances of ordinary moral discourse and these meta-ethical debates is analogous to the relationship between ordinary arithmetic and debates in the philosophy of mathematics. (p. 218)

Gill then elaborates on this example:

> There is no fact of the matter as to whether ordinary mathematic usage is better explained by a Platonist or anti-Platonist conception of number. The way people use numbers in everyday math simply does not contain answers to the questions that animate philosophy of mathematics. That is not to say that the question of what numbers are isn't philosophically important. But it's an ontological question on which conceptual analysis of ordinary arithmetic gains very limited purchase. (p. 218)

Finally, Gill proposes that, just as in the plausible case of indeterminacy about folk mathematical Platonism, ordinary people may likewise have no determinate metaethical stances:

Similarly, there may be no fact of the matter as to whether parts of ordinary moral discourse are better explained by, say, absolutism or relativism. That is not to say that the question of whether we ought to hold that moral reasons are absolute or relative isn't important. Such a question, however, may be one to which conceptual analysis of ordinary moral discourse may not provide a determinate answer (even if moral metaphysics or prescriptive ethics may). (p. 218)

Gill's conception of indeterminacy is not confined to the dispute between realism and antirealism. This is true of my conception of indeterminacy as well. I focus exclusively on realism and antirealism to limit the scope of my project and to provide a single narrative. However, indeterminacy could apply to other metaethical distinctions (and to non-metaethical distinctions as well).

However, one way my conception of indeterminacy differs from Gill's is that I have broadened the scope to include not only indeterminacy with respect to commitments, but indeterminacy with respect to stances, which explicitly refer to ordinary people's mental states, e.g., their *beliefs*. Gill's focus on commitments is a sensible one. As Gill sees it, much of 20th century metaethics was focused on what Gill calls *descriptive metaethics*, which was tasked with providing "the best analysis of the ordinary uses of moral terms" (p. 215, footnote 1). Unfortunately, much of the way descriptive metaethics is framed, even by those engaged in it, is unclear. Much of it certainly *seems* like it's intended to go beyond an external account of the linguistic outputs, independent of the psychological states of the speakers. Yet philosophers central to the debate don't seem to think this is the case. Take, for instance, Sinnott-Armstrong's (2009) distinction between internal and external approaches to moral language. In an article directly responding to Gill's (2009) article on indeterminacy and variability, Sinnott-Armstrong (2009) draws a distinction between an internal and external project:

> There are two ways to describe moral language. An internal project seeks to capture the psychological processes or representations that actually occur when people use moral language. However, contemporary realists and expressivists are not trying to do that. When Jackson and Pettit use networks of truisms or when Gibbard cites hyperstates, *they surely know that these theoretical constructions do not reflect actual psychological entities or events. Instead, they want their theories to be externally adequate in capturing the outputs of our linguistic systems without necessarily reflecting*

*the internal workings of that system. In this respect, their project is more like Chomskian grammar, which uses constructs without claiming psychological reality.* (p. 237, emphasis mine)

Critically, the external project *is not concerned with describing psychological states*. Yet it is *this* approach that Sinnott-Armstrong takes regards as *the* approach that philosophers have taken. After drawing the distinction, he concludes: "Overall, then, I take moral realism and expressivism to be trying to *externally* describe the semantics of all standard moral language" (p. 237, emphasis mine). In other words, Sinnott-Armstrong takes descriptive metaethics to be *exclusively concerned with the external, non-psychological project*.

Any outsider who has a look at the 20th century work that purportedly isn't intended to describe psychological reality could readily reach a different conclusion. Much of the language used by the philosophers purportedly engaged in the external project at least looks, to an outsider, like a claim about ordinary psychology. Consider some of examples Sinnott-Armstrong's own examples:

> "[W]e seem to *think* moral questions have correct answers; that the correct answers are made correct by objective moral facts" (Smith, 1994, p. 6, as quoted in Sinnott-Armstrong, 2009, p. 238, emphasis mine)

> "The ordinary user of moral language *means to say* something about whatever it is he characterizes morally, for example a possible action, as it is in itself, or would be if it were realized, and not about, or even simply expressive of his or anyone else's relation to it." (Mackie, 1977, p. 33, as quoted in Sinnott-Armstrong, 2009, p. 238, emphasis mine)

It is difficult *not* to interpret claims about what ordinary people "think" and "mean to say" as attempts to describe psychological reality. If these remarks aren't intended to do so, their authors were, at the very least, using incredibly misleading language. Perhaps, understood in its proper context, this would be made clear, but one might be forgiven for suggesting that if philosophers don't intend to describe people's psychological states, that it might be best to avoid using paradigmatic psychological terms. This same use of terms with established colloquial psychological interpretations appears in Gill's examples of descriptive metaethics as well.

"Stevensen, for instance, says that his work is concerned to analyze 'the judgments of the ordinary man as he finishes reading the morning's newspaper." (Stevenson, p. v, as quoted in Gill, 2009, p. 216).

Mackie (1977), too, is quoted conveying his concern with "ordinary thought," and again to his concern with what "the ordinary user of moral language *means to say*" (pp. 31-33, as quoted in Gill, 2009, p. 216). Gill also cites Brink (1989), who claims that his views more closely reflect "commonsense moral thinking" but adds that this is "perhaps a little misleading" (p. 37, as quoted in Gill, 2009, p. 216). Note on the same page, Brink also states that "If moral judgments merely purported to state facts, it is claimed, they could not fulfill the action-guiding function they do. To fulfill this function, *moral judgments must concern or express affective, fundamentally noncognitive features of people's psychology*" (p. 37, emphasis mine).[21]

If *this* doesn't qualify as a claim about the *psychological reality* of moral claims, what *would* suffice? I'm not suggesting Sinnott-Armstrong is mistaken in supposing that philosophers are primarily concerned with external descriptions of the outputs of our linguistic practices rather than capturing the psychological states associated with moral discourse. I am drawing attention to the fact that if this is the case, philosophers have a track record of embarrassingly misleading remarks that clearly suggest otherwise. I would be sincerely unsurprised if this is the case. However, I suspect it isn't, and that we are instead dealing with descriptive pluralism: some philosophers are engaged in an external project, others in an internal project, and others are engaged in both. Some (myself included) may even question the legitimacy of the distinction.

---

[21] There are many other examples, as well. For instance, Nichols (2004) quotes Darwall (1998), who states that "Ethical thought and feeling have 'objective purport.' *From the inside*, they apparently aspire to truth or correctness and presuppose that there is something of which they can be true or false" (p. 24, as quoted in Nichols, 2004, p. 7). This reference to how moral thought and feeling are *from the inside* seems to be describing an aspect of moral phenomenology, and it is utterly implausible that this isn't about psychological states. Either some descriptive metaethics *just is* about psychological states, or the philosophers engaged in descriptive metaethics are (i) shockingly confused or (ii) completely misleading. I know of no other way to make sense of what seems like deliberate and explicit psychological ascriptions.

Whether or not descriptive metaethics has historically concerned itself with the *psychology* of folk metaethics, Gill's description of indeterminacy does not explicitly include a concern with philosophical stances. This omission is critical to the empirical study of folk metaethics, since it would boggle the mind to imagine that what is quite clearly research on the *psychology* of metaethics *isn't about claiming psychological reality*. When Goodwin and Darley conducted their seminal research on whether people are realists or antirealists, I take it that, unlike Gibbard, they would *not* claim that their findings are merely theoretical constructions that "surely…do not reflect actual psychological entities or events." Likewise for most other research on folk metaethics. Collier-Spruel et al. (2019) are not citing "networks of truisms" or "hyperstates," they are describing ordinary people's *actual psychological states*.

Unfortunately, the lack of cross-talk between philosophy and psychology has obscured what it is we're supposed to be indeterminists *about*: an externally adequate description of the linguistic outputs of ordinary speakers? Or ordinary people's *beliefs* about realism and antirealism? Or could it be something else entirely, e.g., "intuitions"? I've opted not to presume that our interests must be confined to any particular characterization of what philosophers are attempting to describe or what researchers are attempting to measure. We can assess each separately, hence my emphasis on both philosophical *stances* and *commitments*. Regrettably, this distinction does not appear in research on folk metaethics, nor has there been any substantive effort to clarify what *exactly* researchers are attempting to measure. Many articles appear to be addressing stances and not merely commitments, but when it comes to the matter of measurement, it remains unclear whether researchers take themselves to be engaged in the same empirical enterprise, perhaps in part because some studies are conducted by philosophers and others are conducted by psychologists: while both are studying *folk philosophy*, each discipline brings its own presuppositions along for the ride. Such language is silent or at best unclear about the degree to which it intends to capture any particular psychological states of the people engaged in ordinary discourse.

Nichols (2004) often speaks of *commitments* to moral objectivity (though not necessarily using the term to mean the same thing I do). However, several remarks suggest a psychology element to folk metaethics as well: "Many undergraduates seem explicitly to disavow moral objectivism at least for some standard moral violations" (p. 8). In describing the first study in the article, Nichols references "judgments about moral objectivity" and states that the experiment "was designed to explore participants' views about the objectivity of morals, conventions, and ordinary facts" (p. 9). This seems to be at odds with Sinnott-Armstrong's emphasis on the external project, since it is hard not to see this as an at least partially internal project that is explicitly intended to capture psychological reality. Most studies on folk metaethics likewise employ language that suggests an attempt at describing ordinary psychology, and thus likewise does not seem intended to merely assess the linguistic outputs via an external project. Yet some researchers still refer to "intuitions." Pölzler & Wright (2020b) describing folk metaethics as a study of *intuitions*:

> In the last 15 years an increasing number of psychologists have begun to study folk intuitions about the existence of objective moral truths. Their results suggest that rather than being realists, ordinary people intuitively tend towards "metaethical pluralism" [...] (p. 54)

Yet they describe metaethical pluralism as the claim that ordinary people "regard moral realism as true with regard to some moral sentences or circumstances and anti-realism as true with regard to other moral sentences or circumstances)" (p. 54), and they describe their own research on metaethics as "psychological research on folk moral realism" (p. 55). This would be incomprehensible if it were not intended to capture psychological reality. Taken together then, it is clear that folk metaethics research does not appear to be exclusively concerned with the external project alone, and is thus not only concerned with commitments. It is also concerned with stances, insofar as stances represent beliefs or other psychological states about metaethics.

This leaves us with *two* potential forms of indeterminacy: indeterminacy with respect to commitments, and indeterminacy with respect to stances. I defend both, though I do little in the main

text to emphasize either over the other, or to focus heavily on the distinction. Indeterminacy could also apply to different metaethical distinctions, though my emphasis is only on realism and antirealism. Unfortunately, this only tells us what indeterminacy applies (or could apply) to, but doesn't tell us what indeterminacy itself amounts to.

I do not have anything deep or technical in mind by *indeterminacy* with respect to a given folk philosophical stance or commitment. In its simplest form, indeterminacy with respect to a given distinction is the view that *there is no fact of the matter* about which of a given set of distinctions is correct. For instance, we could endorse indeterminacy for an earlier question about whether unicorns *like* or *dislike* pineapple on pizza. Since unicorns do not exist, there is no fact of the matter about whether they like or dislike pineapple on pizza. Our answer would be "neither." Yet if this option is, by stipulation, unavailable to us, then there is no way to answer the question, or at least no way to answer the question that conforms to the presumptions stipulated by the questioner. Yet *why* a given claim is indeterminate could vary. With respect to the gastronomic standards of unicorns, the reason there is no determinate answer is because unicorns do not exist. Yet suppose there were unicorns, but we had no way of knowing what their preferences were. These unicorns could live in an alternate universe. Perhaps a few travelers from this *uni*verse visited our own, told us that there were unicorns with strong opinions towards pineapple on pizza in their own world, then returned home without telling us what those opinions were. Then the bridge between worlds collapsed, such that it was no longer possible to discover what unicorns think about pineapple on pizza. In such a case, there may be a fact of the matter about whether unicorns like or dislike pineapple on pizza, but we'd have no epistemic access to such facts. Would the question of whether they like or dislike pineapple on pizza be *indeterminate*? In a certain sense, it would: we could at the very least say that it is *indeterminable for us*. Yet this isn't the same thing as there being no fact of the matter.

The same could apply to questions about folk metaethics. It *could* be that ordinary people have determinate metaethical stances or commitments, but that there is no way for us to discover what those stances or commitments are. Perhaps we are dealing with an empirical question but the only tools used to address the question are insufficient for the task.

Yet inveterate optimists may hold out hope that while we are unable to figure out whether ordinary people are moral realists or antirealists *now*, that this is simply because we lack the proper methods. With the right tools in hand, we could settle the matter. If so, then we'd simply lack the tools to determine whether people are realists or antirealists *now*, but we could acquire such tools in the future.

Taking stock of all these considerations, it seems *indeterminacy* could be cashed out in a variety of ways, and could perhaps be roughly plotted along a continuum. In its most minimal form, we might say that a certain issue is at the very least indeterminate given the available arguments and data, i.e., what we might call *local indeterminacy*: something is indeterminate in a given informational context. This is at best a very weak form of indeterminacy, if it qualifies as a form of indeterminacy at all. Many questions may lack an immediately determinate answer, but could easily be answered with little difficulty. If I head to the fridge to see if I have any milk left, whether there is milk in the fridge or not is *indeterminate*, but this is immediately resolved the moment I open the door and peer inside. In other situations, we may have no determinate way to know the answer, but can be confident there will be *some* determinate answer. For instance, suppose we were watching a football game. We may wonder who will win, even if the answer is "nobody" because it's a draw or the game is halted by an alien invasion. Some questions may have no determinate answer until we put in a little effort or simply wait, but these seem like poor candidates for a substantive form of indeterminacy.

We might also imagine a kind of *methodological indeterminacy*. It may be that certain *methods* are incapable of decisively resolving a dispute in a way that furnishes us with a determinate answer. It could be that scientific methods are not capable of providing us with a determinate answer about

whether God exists, murder is wrong, or 2+2=4. Knowledge of such claims might instead only be obtained by e.g., *a priori* reasoning or divine revelation. Conversely, philosophical methods may be incapable of resolving certain empirical disputes. Conceptual analysis or the method of cases isn't going to enable us to diagnose diseases or solve murder cases. It could be that the methods used by analytic philosophers cannot resolve disputes about the meaning of ordinary moral claims or questions about the content of ordinary moral thought, but such questions *could* result in a determinate answer using some other method (e.g., empirical research). Yet the typical methods existing researchers have employed may likewise be unable to resolve questions about folk metaethics, e.g., surveys, but some other method could, e.g., advanced neuroscientific methods that allow us to scan people's brains in ways that provide insights that surveys can't, or perhaps a well-designed approach to interviewing people or studying their behavior could reveal metaethical stances or commitments. It's also not clear any of these possibilities should qualify as genuine instances of indeterminacy rather than interesting challenges or limitations that don't quite pass the threshold for capturing actual indeterminacy.

We begin to move closer to the realm of unambiguous indeterminacy when we consider the possibility of an issue for which there may be some fact of the matter, but there is no feasible way for us to resolve it. For instance, there are many facts about historical events that we couldn't answer and for which two or more competing explanations are equally consistent. For instance, did a particular ichthyosaur (let's call her "Gwendolyn") swim to the left or to the right on August 3rd, at 2:19:46 PM GMT 114 million years ago? In these circumstances, there would be some determinate fact about the direction Gwendolyn swam, but we'd have no epistemic access to it. We might call this *epistemic indeterminacy*: no explanation is better than another, not because there is no fact of the matter, but because we have no available means for knowing what that fact is. Of course, someone could point out that it's *metaphysically possible* or at least *logically possible* to settle the question of where Gwendolyn went.

There would be some fact of the matter, and even if we don't know what that fact is, maybe we could find out. Yet in the absence of such implausible possibilities, we'd still be unable to know which account is correct, so we'd be left with at the very least a practical degree of indeterminacy. Yet so long as it remains possible, given what we know about the universe, then it at least remains not merely a logical possibility but a nomological possibility that we could settle the matter.

We can dial up the indeterminacy further. Perhaps there is some fact of the matter, but there is no way, given the laws of physics, for us to find out, e.g., because time travel turns out to be impossible. In this case, we may encounter a kind of *nomological indeterminacy*, something for which there is some determinate fact that we lack epistemic access *and* couldn't obtain epistemic access to because of the physical constraints of the universe we're in. In this case, we'd still be dealing with an epistemic access problem, so we'd still be in the realm of a kind of inability to know which explanation is correct, even though there is a correct explanation.

Finally, we may be unable to determine which of two competing accounts is correct because *there is no fact of the matter*. Yet once we cross the threshold from a mere lack of epistemic access to there being no discoverable fact even in principle, there's still the question of *why* there's no fact of the matter: is there no fact of the matter due to contingent causal-historical events? If so, then it was still metaphysically and logically possible for there to be a determinate answer, there just happens not to be one. For instance, we could ask whether Caesar's pet cat was named Severus or Claudius. If Caesar did not have a pet cat, then there'd be no fact of the matter. Yet it was both possible given the laws of physics for Caesar to own a cat, and there were no logical reasons why he couldn't have owned one.

Or there could be no fact of the matter due to the laws of physics, e.g., if the Copenhagen account is correct, we could face scenarios such as the titular *Schrödinger's cat*; such situations involve linking the outcome of some macroscopic event (e.g., whether a cat is killed or spared) to a subatomic

particle (e.g., a photon) that is in a state of quantum superposition where one of two events could occur: if the first outcome occurs, the cat dies. If the second outcome occurs, it lives. Is the cat alive or dead? Until the superposition collapses via interaction (i.e., until it's "observed"), it exists in an indeterminate state: the cat is *both* alive *and* dead.

Finally, we could imagine scenarios where the laws of physics would be irrelevant, and there'd be no fact of the matter because a question about which of two or more possibilities is correct could not be answered in principle. This could be because we request a determinate answer to a question that is unintelligible or underspecified or otherwise framed in such a way that there just isn't any determinate answer to it, e.g., "Do all zorps florp or do all zorps flarp?" Without clarity on what this means, there'd be no way of providing a determinate answer. Or it could artificially restrict possibilities to ones that are not logically possible, e.g., "Do all squares have exactly two sides, or exactly six sides?" While the correct answer is determinately "neither," one might frame questions in ways that make the correct answer unavailable, and prohibit one from providing "neither" or "none of the above" as a possible response. Such mistakes don't have to restrict possibilities to a given domain, but could involve category mistakes. For instance, there's no determinate answer to whether prime numbers prefer chocolate or vanilla ice cream. Prime numbers don't have food preferences, and thus cannot prefer one flavor over another. This leaves us with a rough continuum of levels of indeterminacy:

(i)     *Local quasi-indeterminacy*

(ii)    *Methodological quasi-indeterminacy*

(iii)   *Epistemic quasi-indeterminacy*

(iv)    *Incidental indeterminacy*

(iv)    *Nomological indeterminacy*

(v)     *Logical indeterminacy*

The first three categories may be thought of as quasi-indeterminacy: each involves a situation in which there is some fact of the matter that could be known, but (for whatever reason) isn't. The latter three

forms of indeterminacy all reflect instances in which there is no fact of the matter. This is at best a crude and cursory list that likely excludes a variety of important considerations and may exaggerate the degree to which these distinctions are categorical. I'm no metaphysician. Still, we could ask which of the following best captures what I am proposing when claiming that metaethical indeterminacy is a plausible account of metaethics. While the title may be based on *Schrödinger's cat*, I do not think that ordinary people lack determinate metaethical standards due to the physical laws of the universe. *Schrödinger's cat* is simply a metaphor. Rather, I take *incidental indeterminacy* to best capture the way in which ordinary people's metaethical standards are indeterminate: it is logically possible for ordinary people to have determinate metaethical standards (indeed, some *do* have them), and it's consistent with the laws of physics for them to have determinate metaethical standards. They just don't tend to have them for mundane reasons, including *ex hypothesi* that (a) a commitment to realism or antirealism hasn't been a feature of any natural languages, (b) it isn't an innate feature of our evolved psychology, and (c) it isn't a typical feature of our enculturation or everyday experiences, and as a result most people haven't thought about realism and antirealism and reached any kind of determinate conclusions. There is no reason in principle why any of these conditions couldn't have been met. I just don't think they were. This puts metaethical indeterminacy in a comfortable goldilocks zone between our merely not knowing what people's metaethical stances or commitments are, and it being literally impossible for people to have determinate metaethical stances. While we might be tempted to claim that ordinary people, in virtue of people ordinary, *can't* have determinate metaethical stances or commitments, this would be a mistake. Engaging in academic philosophy isn't the only way for someone to have determinate metaethical stances or commitments. Evolution could have selected for beings who spoke or thought like realists or antirealists. Or it could be that the moral discourse in a possible language could be best explained by realism or antirealism. Or it could be a feature of a particular culture or religion that moral facts "don't depend on our subjective values" without this rising to the level of a

substantive philosophical thesis, but just something people take for granted in a community and understand well enough to express. This latter possibility may be true of existing populations researchers have yet to study. It does not strike me as unbelievable that an insular religious community could be composed of people who would both reliably affirm realism but lack any meaningful contact with academic philosophy. The plausibility of there being such populations is one of the reasons I've qualified the indeterminacy hypothesis with an explicit recognition for local determinacy in some cases. Researchers have yet to reach sufficiently diverse populations to know one way or another how they think about metaethics.

Metaethical indeterminacy is therefore the thesis that most ordinary people happen not to have any determinate metaethical stances or commitments with respect to realism and antirealism. It is *not* the claim that we have determinate stances or commitments but we haven't discovered (or can't discover) what they are. Yet it is also not the claim that they *couldn't* have determinate metaethical stances or commitments (whether for nomological or logical reasons). Rather, it is the intermediate claim that they simply *don't* have any particular metaethical stances or commitments.

### S1.12 Types of pluralism

I don't discuss metaethical pluralism (or *variability*) much in the main text. Here, I will outline different forms that pluralism can take. The two most basic forms of pluralism are *interpersonal* and *intrapersonal* pluralism, and these are the two discussed most frequently in the empirical literature (note that all forms of pluralism can apply to stances or commitments, unless otherwise specified). My emphasis will be on pluralism with respect to realism and antirealism, but note that pluralism could apply to other metaethical distinctions as well as non-metaethical distinctions.

**Interpersonal variation**

Interpersonal variation is stable variation *between* participants, e.g., one person may be more disposed to endorse realism, while another is more disposed to endorse antirealism. For instance, suppose we

presented Alex and Sam with the same set of ten moral issues and measured whether they took a realist or antirealist stance towards each of those issues. Alex chose a realist response for 8 out of 10 cases, while Sam chose a realist response for just 2 out of 10 cases. If these 10 moral issues represent their views towards morality as a whole, we could conclude that Alex takes a realist stance towards most moral issues, but Sam takes an antirealist stance towards most moral issues. This would reflect *interpersonal variation*: variation *between* people. We might think of interpersonal variation in conventional psychological terms by conceiving of it as a measure of *individual differences*, whereby some individuals exhibit a greater or lesser tendency towards one or another of different metaethical stances or commitments.

**Intrapersonal pluralism**

Intrapersonal pluralism, in contrast, refers to the adoption of different metaethical stances or commitments *for the same participant*. For instance, suppose Alex has a realist stance towards murder and stealing, but an antirealist stance towards abortion and euthanasia (a common pattern in the literature). Interpersonal and intrapersonal pluralism aren't mutually exclusive: there can be both stable patterns of variation *between* people and *within* a particular person's metaethical standards. Alex and Sam, in the examples above, exhibit stable patterns of variation in the degree to which they favor realism over antirealism, yet neither *exclusively* favors realism or antirealism about all issues. Both instead exhibit some pluralism towards the moral domain, expressing both realist and antirealist stances towards at least some moral issues. Intrapersonal pluralism does *not* require having a balanced tendency to judge moral issues as a realist or antirealist, such that one is a realist about half and an antirealist about the other half. So long as a person exhibits any variation at all across moral issues, they exhibit some degree of intrapersonal pluralism.

The most common form of intrapersonal pluralism is *content pluralism*. Content pluralism refers to having different metaethical stances or commitments towards different moral issues. A person

might endorse or speak like a realist when talking about murder or torture, but endorse or speak like an antirealist when talking about abortion or euthanasia. It is also possible for people to endorse content pluralism towards moral subdomains; for instance, a person could have the same metaethical stance or commitment towards all moral issues related to harm or fairness, but a different metaethical stance or commitment towards moral issues in another domain, such as loyalty, respect, or purity (See Davis, 2021). We could refer to this as a subtype of content pluralism, *domain content pluralism*.

Finally, someone could have different metaethical stances or commitments towards different abstract moral principles, which we could think of as *principle pluralism*. Perhaps they are a realist about the "*ought implies can*" principle, but an antirealist about the doctrine of double effect. This strikes me as an unusually sophisticated form of pluralism that I doubt anyone would exhibit, but it would still reflect a kind of content pluralism, only towards general moral principles rather than specific moral issues.

**Other forms of pluralism**

Most research on folk metaethics focuses on interpersonal and intrapersonal pluralism. However, these are not the only forms of pluralism, and each admits of further subdivisions. Interpersonal variation is an instance of a broader category of variation *across* moral perspectives rather than *within* them. Likewise, most intrapersonal pluralism focuses on *content pluralism*, i.e., expressing different metaethical standards towards different moral issues. However, intrapersonal variation can vary along other dimensions as well. Here, I describe some of the more exotic and understudied forms of metaethical pluralism.

**Intergroup pluralism**

Intergroup pluralism is simply the population-level counterpart to interpersonal pluralism. Some *populations* could be more inclined towards realism or antirealism relative to other populations. Perhaps, for example, Mormons are more inclined towards realism, but atheists are more inclined towards

antirealism. Intergroup pluralism could occur between *any* populations: religious groups, cultures, and so on. We can even imagine stable variation in the metaethical stances and commitments of different species, e.g., we could encounter an alien species that favors realism, and another which favors antirealism, a possibility we could describe as *interspecies pluralism*. To my knowledge, no serious effort has been made to explicitly describe or empirically assess any form of intergroup pluralism, but some research could be described in this way, and I *would* describe it this way, so there is at least a little research on intergroup pluralism. In particular, Beebe et al. (2015) conducted cross-cultural research on folk metaethics in the United States, Poland, Ecuador, and China, though they didn't find substantial differences across populations. Even so, this reflects an attempt at assessing *cultural intergroup pluralism*. Sarkissian et al. (2011) likewise find a similar response pattern among students in the United States and Singapore.

Beebe & Sackris (2016) studied changes in the overall proportion of folk realism and antirealism across the lifespan, identifying a period of reduced realism in people's late teens and twenties. This reflects *intergroup age pluralism*, i.e., variation in metaethical stances or commitments as a function of one's age. Such variation could reflect a stable developmental pathway that emerges in different societies and could even be a species-typical trait. It could be, in other words, that people (or members of some populations) tend to begin life favoring realism, exhibit a decline early in life, then return towards a higher level of realism as they approach their thirties which persists for the remainder of their lives. However, it's also possible that some generations are more disposed towards realism than others. Perhaps people born in the 1990s are more inclined towards antirealism than other populations. I don't find this to be a plausible explanation for Beebe and Sackris's findings. Since there are no especially good reasons to think the patterns found in their study are merely a byproduct of the age cohorts they happened to study, we could imagine broader generational trends that *could* result in

stable intergroup variation. For instance, it could be that people born in medieval Europe were more inclined towards realism, but that people in 21st century Europe are more inclined towards antirealism.

**Context pluralism**

In the article originally outlining metaethical indeterminacy, Gill (2009) describes a form of pluralism that has received little empirical attention: *context* pluralism. Gill describes context pluralism as the view that:

> [T]here are some contexts in which moral terms are used in a manner that is best analyzed as involving one commitment and other contexts in which moral terms are used in a manner that is best analysed as involving the commitment that has traditionally been taken to be the former's meta-ethical competitor. (p. 222)

Gill only seems to be describing pluralism about commitments, but we could expand the notion to include stances as well. Broadened in this way, context pluralism occurs when a person exhibits one set of metaethical stances or commitments in one context but a different set of metaethical stances or commitments towards in another context. For instance, an academic philosopher could speak or think like an antirealist when adopting the theoretical lens of a philosopher, e.g., while writing a philosophy article, teaching a class, or discussing philosophy with colleagues, but speak or think like a moral realist outside of such contexts. This is similar to Gill's own example, described elsewhere (Gill, 2008), which suggests pluralism between "personal and professional settings" (Sinnott-Armstrong, 2009, p. 248; for Gill's example, see Gill, 2009, p. 229). Gill also mentions a handful of other possibilities:

> And there also may be some people who use moral terms relativistically in certain situations— say, when discussing the moral status of individuals in distant times or places or when conversing with other people who themselves use moral terms in a predominantly relativistic way—and who use moral terms objectively in other situations—say, when assessing the laws, policies, or customs of their own country or when conversing with other people who themselves use moral terms in a predominantly objectivist way. (p. 391)

Adopting different metaethical stances or commitments when speaking about different populations represents an unusual form of pluralism: one where one's standards don't vary based on the context

the thinker/speaker is in themselves, but rather vary based on who they're speaking about, i.e. a type of pluralism similar in some respects to the distinction between agent and appraiser relativism (Quintelier, De Smet, & Fessler, 2014). Gill also describes a form of context pluralism where a person mirrors or adopts the commitments of the group they're speaking with, which we might think of as a type of *social adaptation pluralism* or *local linguistic convention pluralism*. Yet there are still more ways a person could adopt different metaethical stances or commitments in different contexts. A person could shift between different cultural paradigms, adopting one set of metaethical standards or commitments when operating within one cultural lens, but adopting a different set of stances or commitments when speaking and thinking through a different cultural lens. A person who has experienced or lived in different cultures could shift between the two. Perhaps, for instance, someone from a highly religious society moved to a secular society. When speaking to and interacting with relativists, they speak (and perhaps even think) like a moral realist, but they speak and think like a moral antirealist in their everyday lives. There are many possibilities. I am not aware of any studies that explicitly explore the possibility of context pluralism, though Sinnott-Armstrong (2009) provides some reasons to be skeptical of context pluralism (see pp. 248-250).

**Linguistic pluralism**

There are yet more possible forms of pluralism. For instance, *linguistic pluralism* is a possibility, though one confined to commitments rather than stances. Linguistic pluralism would, unsurprisingly, consist of variation between languages. It could be, for instance, that when speaking English, people are committed to realism, but that when speaking some other language, they're committed to antirealism. I haven't seen this possibility discussed anywhere. I could take this absence as a pretext for raising a more general complaint about academic philosophy, and I think I will. Far too much academic philosophy is conducted in a small subset of the world's languages, and concerns the work of people writing in those languages, primarily English. People studying and publishing in philosophy do not

come anywhere close to a representative sampling of the world's languages. As such, whatever inferences philosophers make about "our" linguistic commitments may be predicated on so impoverished and unrepresentative a sampling of the actual (much less possible) ways people could speak may not be justified. *Even if*, for instance, we could decisively show that native English speakers, along with people who speak German, French, Italian, Spanish, and Portuguese, spoke in ways that committed them to realism, it does not follow that other languages also commit speakers to realism. Like psychologists, philosophers may err in presuming that if the people around them exhibit a particular philosophical commitment that *everyone* shares that same commitment.

**Other types of pluralism**

There are even more complicated forms of pluralism. One could be a *nested pluralist*, which might roughly map onto Wong's (1995, 2006) notion of pluralistic relativism. This would involve having a particular overarching metanormative standard towards a particular domain, but allow for local variation in metanormative standards within a domain. For instance, one might be a realist about our obligation to show respect for the dead, but an antirealist about the particular ways in which we must comply with this demand. For instance, you could be a relativist about whether we cremate or bury the dead. This wouldn't be identical to the claim that there are different ways of complying with the same moral standard. Rather, the view could be that claims about whether we should bury or cremate the dead carry implicit indexicals that are true or false relative to the practices they are relativized to. On this view, you'd have a type of antirealism (specifically, a form of cultural relativism) nested *within* a type of realism.

**Pluralism in other domains**

All of these types of pluralism could apply to other normative domains as well, such as epistemic, prudential, or aesthetic norms. They could also apply to various types of claims of non-normative (or *descriptive*) claims, e.g., claims about historical events or scientific or religious claims. Although many

studies include claims about social conventions, aesthetics, and descriptive claims, these are typically included as controls or foils that allow for cross-domain comparison; the focus is always on morality.

**Semantic pluralism**

Sinnott-Armstrong (2009) also proposes a type of terminological pluralism he refers to as "Semantic term variantism" (p. 240).[22] This form of pluralism consists of using some terms in ways that are best explained by one metaethical account, but uses a different set of terms in a way best explained by a different metaethical account. For instance, it could be that when people use deontic terms such as "permissible" or "must," that they are committed to realism, while when they use evaluative terms such as "good" or "bad," that they are committed to antirealism. Sinnott-Armstrong cites a handful of articles that offer different semantic analyses of different moral terms, e.g., Edwards (1955) suggests different ways of analyzing "ought" and "good," and Gert (2005) provides different accounts of "wrong" and "ought." However, Sinnott-Armstrong (2009) is most interested in the semantic variation proposed in the following passage:

> I propose that we distinguish between valuations (typically recorded by such forms as 'x is good', 'bad', 'beautiful', 'ugly', 'ignoble', 'brave', 'just', 'mischievous', 'malicious', 'worthy', 'honest', 'corrupt', 'disgusting', 'amusing', 'diverting', 'boring', etc.—no restrictions at all on the category of x) and directive or deliberative (or practical) judgements (e.g. 'I must y', 'I ought to y', 'it would be best, all things considered, for me to y', etc.). (Wiggins, 1987, p. 95, as quoted in Sinnott-Armstrong, 2009, p. 240)

I won't rehash Sinnott-Armstrong's discussion or the objections he raises (see pp. 240-242). This passage apparently offers a different analysis of practical judgments and valuations. Setting this issue aside, we may point to another type of semantic pluralism: it could be that one adopts different metaethical stances or commitments with respect to claims about *thick* moral terms (e.g., *cruel*, *kind*,

---

[22] For some reason, Sinnott-Armstrong defines semantic term variantism as the view that "some moral words should be understood in the way expressivists claim but other moral terms should be understood in the way realists claim" (p. 240). I see no reason to limit the notion of semantic pluralism to the distinction between cognitivism and noncognitivism, but instead would open it to all possible metaethical distinctions.

*unscrupulous*, *brave*, etc.) and *thin* moral terms (e.g., *right*, *wrong*, *good*, *bad*; Väyrynen, 2021). For instance, someone might express a form of realism towards thick moral claims, believing that there is a stance-independent fact about whether people are cruel or kind, brave or cowardly, and so on, but speak in a way that best fits antirealism when talking about what's morally good or bad or what we should or shouldn't do.

## Tense pluralism

Sinnott-Armstrong briefly considers (in order to dismiss) *tense pluralism* (p. 239). Tense pluralism would involve variation in our metaethical commitments when speaking in different tenses, e.g., past, present or future tense, or when speaking in first, second, or third person. Sinnott-Armstrong describes tense pluralism as "indefensible," and I'm inclined to agree that it's not a likely candidate for describing how anyone actually speaks.

## Atypical or unprincipled pluralism

There may also be atypical or unprincipled forms of pluralism. That is, there may be instances where people express different metaethical standards, but they aren't governed by any principled patterns or regularities. While there would presumably be some fact of the matter about *why* a person would speak or think like a realist or an antirealist in some cases but not others, these cases could be so atypical as to fall outside the scope of conventional categories; e.g., it could be that a person tends to speak more like a realist earlier in the day, but gradually transitions towards speaking like an antirealist as the day goes on or as they become more tired. Perhaps such a person is more prone towards realism when they're under less cognitive load, or when they exert more effort. Or someone might arbitrarily adopt different metaethical standards on a whim. While not plausible, we should be open to unexpected ways in which metaethical stances or commitments could vary.

**Meta-metaethical pluralism**

Note that all of these examples involve descriptions of the way people speak or think. Yet none explicitly address whether people are aware of pluralism in their stances or commitments. In principle, ordinary people could have stances or commitments towards metaethical pluralism itself, in which case they'd exhibit a kind of meta-metaethical pluralism: a metaethical stance or commitment *about* metaethical pluralism. For instance, someone could become aware of such pluralism and explicitly endorse it. That is, they could express the view that metaethical pluralism is the correct account of the nature of moral truth. They could endorse stance-independence about the moral status of murder, but stance-dependence about the moral status of abortion. In principle, people could also speak in ways that commit them to metaethical pluralism. Neither of these possibilities strike me as plausible accounts of how any significant number of ordinary people are disposed to think or speak, but they nevertheless reflect possibilities in principle.

**Strategic pluralism**

The last form of pluralism I wish to discuss is the one I consider to be the most important and interesting possibility: *strategic pluralism*. Strategic pluralism occurs whenever people adopt different metaethical stances or speak in ways that seem to best fit particular metaethical analyses in different contexts in order to achieve some personal goal, e.g., a social goal such as persuading others or signaling positive character traits.

Consider Senator Phil. Phil doesn't have any genuine position on moral realism or antirealism. He has never studied philosophy, and doesn't much care for it. He's more into gambling, whiskey, and expensive vacations. Phil is a clever politician, however. He deftly employs just the right intonation and vocabulary to make his political rivals look as corrupt and odious as possible, but paints his allies as noble servants of the common good. Without realizing it, Phil has come to systematically employ language that would best fit a realist or antirealist analysis depending on whether language alluding to

one or the other position would best suit his argumentative goals. For instance, Phil might employ realist-sounding language when denouncing his rivals, invoking the notion that our rights "aren't a mere matter of convention or arbitrary value, but reflect higher principles that transcend time and space." Yet when defending policies favored by Phil's own political party, Phil might claim that our moral standards are constructed through a legitimate political process, and that so long as we agree to a given set of standards, that standard is as valid as any other: "We live in a democratic nation and are governed by the consent of the majority. As elected officials, it is the prerogative of my colleagues to take the initiative in crafting social policy they believe is in the best interests of our nation, and that best reflects the will of the people." This language may best fit a constructivist approach to morality, which Phil may be exploiting to frame his party's actions as less morally objectionable. In this case, Phil might employ such language because it is *effective*, not because he consistently speaks or thinks like an antirealist, and Phil could effectively employ this language without realizing that he's speaking in a way that fits particular antirealist metaethical positions in the academic literature.

Phil's rhetoric exhibits a kind of intrapersonal context pluralism with respect to metaethical commitments. Notably, Phil may not employ this language outside an argumentative context; such discourse can be confined exclusively or at least primarily to particular argumentative contexts, but does not reflect the way Phil speaks to his family or in casual, apolitical settings.

Notably, Phil's use of realist and antirealist language can be predicted on the basis of what would serve his argumentative or social goals. As such, the principles governing his deployment of particular metaethical language aren't governed by an arbitrary or incidental shift in metaethical commitments across contexts, nor a principled stance towards different contexts or situations. That is, it's not as though Phil genuinely holds to a substantive philosophical position whereby his colleagues really are subject to antirealist standards, but his enemies aren't. Phil need not believe what

he says, though it's also possible that Phil's metaethical stance shifts along with shifts in his rhetoric.[23]

However, there is evidence that self-deception can make people more persuasive (Schwardmann & Van der Weele, 2019), so it is a possibility that people's metaethical stances shift alongside their outward commitments.

Phil's behavior illustrates how in principle a person could strategically adopt different metaethical stances to achieve social goals. In this case, his goal is to depict one's rivals as terrible people and attempt to justify the actions of one's allies. Such contexts need not be to win arguments, but could also serve to signal desirable qualities. For instance, someone might speak like a relativist around college friends in order to signal tolerance but speak like a realist at church to appease family and avoid the deacon's glares.

Strategic pluralism may provide a plausible account of apparent instances of metaethical pluralism in everyday moral discourse: such occurrences may reflect a shallow, superficial commitment to different metaethical positions that manifest in contexts where expressing an apparent commitment to a particular metaethical perspective would serve particular social goals. There are at least four bodies of literature that support the plausibility of strategic metaethical pluralism. However, note that it is *not* the purpose of this section to justify or argue for strategic pluralism, so I will simply provide a sketch.

First, people already associate moral relativism with tolerance and realism with a rigid and closed attitude towards others (Collier-Spruel et al., 2019; Goodwin & Darley, 2012; Wainryb et al., 2004). If so, this provides fertile ground for such associations to influence moral judgment and

---

[23] Strategic self-deception might seem implausible, but our capacity for self-deception could enable people to sincerely alter their metaethical beliefs without realizing that they are unconsciously driven to do so by a cynical desire to achieve certain practical goals. While self-deception may have costs, it may also have advantages (Chance & Norton, 2015; Mijović-Prelec & Prelec, 2010; Moomal & Henzi, 2000). Some studies show that self-deception can reduce cognitive load (Jian et al., 2019). Since increased cognitive load is an indicator of lying, reducing cognitive load via self-deception can help conceal one's deceptions (Jian et al., 2019; Trivers, 2011). More generally, deceiving others (or at least persuading them to do what you want) could be more readily facilitated if one believes one's own lies. Yet strategic pluralism need not involve self-deception; it could merely reflect a kind of pragmatic inconsistency. It would only constitute self-deception if a person would recognize strategically varying their metaethical beliefs as an accurate reflection of their genuine higher-level perspective on the matter.

discourse. After all, if people *didn't* associate realism and antirealism with any considerations that had normative implications, this would undermine much of the rhetorical impact of expressing a commitment to one or the other. Relatedly, Fisher et al. (2017) report that people who engaged in cooperative exchanges were less likely to endorse moral realism than people engaged in competitive exchanges. This suggests that one's objectives in a particular social situation could influence their metaethical stances or commitments, illustrating that metaethical stances may be able to vary by context though this does not directly demonstrate that they do so in ways that facilitate one's objectives.

Strategic pluralism is similar in some ways to the proposal put forward by Wright (2018) that metaethical pluralism may be best explained by serving a pragmatic function. According to Wright, metaethical pluralism serves to "assist in our ongoing individual and collective navigation of normative space by creating and maintaining a civil space for discourse" (p. 140). According to Wright, metaethical pluralism plays a psychosocial role in "modulating the level of permissible choice and dialogue about moral issues, both within and between socio-cultural groups" (p. 140; see also Wright, Grandjean, & McWhite, 2013, and Wright, McWhite, & Grandjean, 2014). Wright argues that when we take a realist stance towards a moral issue, we remove it from "the realm of legitimate personal/social negotiation" (p. 140). This means that it becomes unacceptable to hold an opposing view about the moral issue in question, and efforts to inhibit people from engaging in the action may be warranted. Conversely, adopting an antirealist stance towards a moral issue allows one to both signal its moral importance (rather than it *merely* being a personal preference or social convention) while simultaneously allowing one to signal respect, tolerance, and receptivity to dialog and negotiation about the issue in question, which renders efforts to inhibit the action in question less acceptable (see Wright, pp. 140-141). Wright reports two studies that support her hypothesis about the pragmatic role of metaethical pluralism. First, Wright argues that people attributed greater internal motivation to

opposing moral actions when they expressed a cognitivist rather than noncognitivist stance towards the moral issue, a difference which is consistent with moral judgments not simply being an incoherent jumble but potentially tracking normatively relevant judgments about oneself and others.[24]

Second, there is some evidence that people engage in motivated moral reasoning (Ditto, Pizarro, & Tannenbaum, 2009). Motivated reasoning is a cognitive bias that causes us to process information and reach conclusions based in part on the degree to which we desire to believe those conclusions, and not merely on good reasons to believe those conclusions are true (Kunda, 1990). This is unsurprising; if anything, it would be more surprising if motivated reasoning *didn't* extend to moral judgment and reasoning. Our moral beliefs and standards represent some of our most firm convictions (Skitka et al., 2005) and are central to our identity (Heiphetz et al., 2018; Heiphetz, Strohminger, & Young, 2017). To the extent that motivated reasoning can prompt us to deviate from an exclusive concern with engaging in truth-optimizing judgment and discourse, this provides a foundation on which the case for strategic pluralism can be made. To the extent that motivated reasoning can be understood as serving a useful sociofunctional role, we may recognize it as a feature and not a bug.[25] The particular role it may serve with respect to strategic pluralism is in facilitating people to reason and argue in ways that promote their social goals in a given situation by selectively reasoning towards conclusions that support their arguments. People don't keep perfect track of everything they've said in the past, and don't mechanically espouse moral stances or present moral arguments in ways that are logically restricted only to statements consistent with all their previous statements. That would be absurd; we're not robots that will abruptly get an error message if what

---

[24] This is my best attempt at briefly summarizing the rationale behind these studies. The degree to which these findings support Wright's proposed pragmatic hypothesis is not entirely clear to me but strikes me as somewhat tentative. This is no strike against these studies; there's nothing wrong with tentative and indirect findings in support of a hypothesis.

[25] All else being equal, explanations of our cognitive processes that can explain why the way they function is adaptive rather than deleterious should be favored. This is *not* to say that we should presume any and all features of human cognition are adaptations, and that the only question is why they're useful. I am not advocating a dogmatic or naive panadaptationism (Dupré, 2016; Koonin, 2009). It is merely an explanatory virtue of an explanation for a particular trait if, all else being equal, we can make sense of why it could have been favored by natural selection rather than inexplicably maladaptive.

we're saying in a discussion isn't consistent with something we said years ago, or even a few moments ago. In other words, it may not be useful for people to be rigid and consistent across situations when it comes to presenting arguments. We can imagine Senator Phil reasoning in ways that prompt him to *sincerely* believe an argument he presents on the senate floor and to *sincerely* believe an argument he makes a week later, even though a close inspection of these arguments would reveal that one seems to imply a commitment to realism and the other to antirealism.

Third, there is already considerable evidence that our moral standards and the way we frame our moral positions can signal socially desirable traits, and that people take advantage of this to pursue social objectives such as increased status. For instance, Carnes et al. (2022) draw on a host of findings which indicate that people employ outward displays of moral judgment and behavior as "diagnostic traits like trustworthiness, impartiality, and cooperativeness" (p. 2). According to Carnes and colleagues, this includes moral *condemnation* (Hok et al., 2020), *punishment* (Gordon & Lea, 2016; Jordan & Rand, 2020; Kurzban, DeScioli, & O'Brien, 2007), *compensation* (Dhaliwal, Patil, & Cushman, 2021), and *behavior* (Choshen-Hillel, Shaw, & Caruso, 2015; Everett et al., 2018; Shaw, Choshen-Hillel, & Caruso, 2016).[26] Indeed, one recent study found that people are so motivated by a desire to improve their status that they will punish people under ambiguous circumstances where the participant is uncertain whether the punishment is warranted (Jordan & Kteily, 2022). Research on moral grandstanding also demonstrates that a motivation to engage in *moral grandstanding* (exploiting moral discourse to improve one's status) is associated with moral conflict and a disposition to seek status, (Grubbs et al., 2019). Tosi and Warmke (2016; 2020) argue that grandstanding is ubiquitous, and that modern institutions and technology have only served to amplify a preexisting tendency for people to strategically exploit moral expressions in order to achieve social objectives, e.g., improving their status.

---

[26] They included a single reference to each. I have added a handful of additional references.

Finally, there is a more foundational account of how human reasoning may have evolved that lends additional credence not only to strategic pluralism, but to the prediction that we should *expect* to observe strategic pluralism given a plausible account of the selective pressures that may have shaped our capacity for reasoning more generally (e.g., outside a metaethical context or moral context more generally). Our eyes, ears, and other sensory capacities all presumably evolved to provide us with *accurate* information about the external world; they enable us to spot predators, prey, and potential allies, find food, and avoid environmental hazards. Yet humans have an advantage over nonhuman organisms: we can engage in far more sophisticated reasoning. Mercier & Sperber (2011) note that our initial assumption may be that reasoning evolved to "improve knowledge and make better decisions" (p. 57). Yet decades of research in psychology and behavioral economics reveal that we are vulnerable to a suite of biases that distort our picture of reality and lead us to make bad decisions (Caverni, Fabre, & Gonzalez, 1990). Perhaps natural selection cobbled together a marginally functional set of psychological mechanisms that by some miracle enable us to fill the *cognitive niche* (Pinker, 2010). Why, for instance, are we so ludicrously vulnerable to motivated reasoning (Galef, 2021; Kunda, 1990) and confirmation bias (or "myside bias"; Mercier & Sperber, 2017; Nickerson, 1998; Peters, 2020)?

Sperber and Mercier (2011) offer an alternative explanation for why we'd be so prone to erroneous reasoning and poor judgment. They suggest that the primary function of reasoning is *argumentative*, and that it serves principally to "devise and evaluate arguments intended to persuade" (p. 57). The *argumentative theory of reasoning* (ATR) strikes me as a compelling and tidy explanation for why we reason in the way that we do, redescribing confirmation bias, motivated reasoning, and other seemingly faulty elements of our reasoning as features—rather than bugs—that facilitate our social goals. I suspect that *this* insight is the key observation that lends plausibility to strategic metaethical pluralism. There is nothing deep, insightful, or distinctive about the notion that people may exploit metaethical or seemingly-metaethical terms and concepts when arguing or engaging with one another.

Such pluralism is a downstream byproduct of a more general tendency for people to reason in ways optimized for persuading others. This emphasis on persuasion synergizes well with the ubiquity of moral grandstanding and the role moral discourse plays in signaling our desirable traits. We don't just attempt to persuade others of our moral position. We argue that our position is noble and righteous, and that we are noble and righteous because we stand behind the right cause. At the same time, people who disagree with us aren't merely mistaken: they're terrible, horrible, no good, very bad people who we should condemn and dissociate with. Such moral posturing can serve not merely to signal our desirable traits, but to signal our tribal allegiances, forge alliances, galvanize our side of a dispute, mitigate punishment for ourselves or our allies, portray our rivals in the worst possible light, attempt to sway public opinion against rivals and competing ideologies, and so on.

To the extent that the language of realism and antirealism could facilitate these argumentative goals, we should actively expect people to shift the way they speak, and even the way they think, from one argumentative context to another, employing language appropriate to the situation. Insofar as a commitment to realism would make us appear better and our rivals appear worse, that's the language we should expect people to employ. Realist-sounding language could be used to signal our conviction and confidence, but it could also be used to portray our rivals as bigoted, inflexible, and intolerant. The language of relativism could be used to present ourselves as tolerant and open-minded, but it could also be used to depict our rivals as wishy-washy, pathetic, indifferent towards suffering or evil, or even welcoming of depravity.

Critically, strategic pluralism not only predicts metaethical pluralism, it also predicts when and how it occurs, and provides a tidy rationale for *why* it occurs that doesn't require us to conclude that ordinary people are simply stupid, confused, or irrational, as some have suggested (Colebrook, 2021). This isn't to say there aren't instances of stupidity, confusion, or irrationality, but simply that these aren't the *only* reasons why people might employ realist and antirealist language in different contexts.

Strategic pluralism also illustrates an important feature of folk philosophy: while people may be inconsistent with respect to traditional philosophy categories and distinctions, they are *not* inconsistent with respect to *their goals*. A great deal of academic philosophy operates on the assumption that a commitment to one or another of competing philosophical theories is implicit in the way that we speak and think, and that these commitments are logically consistent, as though our linguistic practices are structured in a way that conforms to the values and ideals of academic philosophers. *There are no good reasons to think people would speak or think this way.* Logical consistency may be a desirable feature of a philosophical theory, but it is not necessarily a desirable restriction to place on the way we speak and think. It would limit our argumentative advantages and curtail our rhetorical excesses if we spoke and thought in ways constrained by a desire to use terms in a precise, truth-tracking way optimized for clarity and logical consistency. Philosophers may acknowledge that of course people deploy language to suit their agendas by twisting, bending, and contorting our shared terms and concepts as needed. Yet, they might insist, such uses are parasitic on the everyday meaning of those terms.

Strategic pluralism allows us to make sense of why people would express, or appear to express, different metaethical commitments in different contexts. Yet we should not take such usage to necessarily reflect genuine shifts in metaethical stances, nor should we read too much into using seemingly realist or antirealist language. While metaethical terms and concepts may have penetrated particular linguistic communities, we should be extremely cautious about generalizing from any particular set of anecdotes or data to other populations, including past societies whose terms and concepts may differ even when they superficially resemble our own. Ironically, many of the real or apparent instances of ordinary people deploying metaethical language may result from such terms and concepts trickling into popular consciousness. For instance, terms like "moral realism" and "moral

relativism" only took off after 1940.[27] In the ensuing years, these terms may have trickled out of academic publications and made their way into our common discourse. More generally, terms and concepts bandied around in academic contexts may work their way into popular consciousness in ways that are poorly understood and reflect little more than superficial imitation. This could lead to the *appearance* of widespread metaethical stances and commitments where none truly exist. Such language is simply a convenient bag of terminological tools adapted to contemporary argumentative contexts.

**S1.13 Inconsistencies in metaethics terminology**

Some readers may be concerned that I use a variety of terms that differ both from the terms they are personally familiar with, and that deviate from the terminology used in the articles I describe. For instance, an article may purport to measure beliefs about "objectivism," but I will describe it as a measure of "realism." And academics who study metaethics will often use the term *mind*-independent, not *stance*-independent. Others may note that I opt for specific terms, such as *stance*-independent, in lieu of more common terms like *mind*-independent. While even if I clarify that I mean the same thing by "realism" as the authors of the study in question mean by "objectivism", or that *stance*-independent means the same thing as *mind*-independent, it is still reasonable to ask why I opt for specific terminological choices, and why those choices differ from the terminology used in the literature. There are three reasons:

(1)  I chose specific terms and stuck with them for the sake of internal consistency

(2)  I chose terms that I believe best reflect the relevant distinctions

---

[27] See Google Ngram Viewer, which shows results for these terms from 1800 to 2019 for relativism (using the search "moral realism," Google Ngram Viewer, 2019a) and realism (using the search "moral relativism," google Ngram Viewer, 2019b).

(3) There is no standard or canonical terminology in use in research on folk metaethics or metaethics more generally, so I am in no way bucking conventions or eschewing some standard set of terms.

With respect to (1), the rationale for consistently using the same terms should be clear enough that I see no reason to provide an extensive justification for doing so. Using a variety of technical terms interchangeably would serve no purpose other than to confuse people.

With respect to (2), it would be extraordinarily tedious even for someone as pedantic as I am to explain the rationale behind every terminological choice. Some choices are fairly simple, and there are many reasons to favor one set of terms over another. For instance, I favor *realism* over *objectivism* because "objective" has far too many colloquial usages, e.g., "unbiased," "capable of being measured or quantified," and so on. Objectivism has no natural, established lexical counterpart the way realism does (i.e., *antirealism*). I don't know of any philosophers employing clunky terms like a-objectivists or anti-objectivists. Some researchers have opted for the somewhat less awkward "non-objectivist," (e.g., Goodwin & Darley, 2008) but this is not a term in common use among contemporary philosophers. Realism and antirealism, on the other hand, *are* well-established terms frequently used in the literature. Finally, many researchers, and perhaps some ordinary people familiar with some familiarity with metaethics, may contrast "objective" with "subjective." Yet subjectivism is *not* the negation of objectivism, where objectivism is understood to mean "stance-independent." There are two orthogonal distinctions at play here (Joyce, 2015):

(1) The distinction between stance-independence (*realism*) and the denial of stance-independence (*antirealism*).

(2) The distinction between moral claims having an indexical element (*relativism*) and moral claims having no indexical element (*nonrelativism*)

"Objectivism" is typically used in folk metaethics research to refer to *stance-independence*, i.e., as a stand-in for "realism" in (1). The problem with contrasting objectivism with relativism/subjectivism should now be obvious: *relativism and subjectivism are orthogonal distinctions*. In principle, you could endorse stance-

independence *and* the indexicality of moral claims, or you could endorse stance-dependence and non-indexicality. Joyce likewise draws attention to the distinction, observing that although relativism treats moral claims as having "an essential indexical element, such that the truth of any such claims requires relativization to some individual or group," this does *not* necessarily entail that these claims are stance-dependent, i.e., that they are *made true* by the standards or values of those groups:

> In all cases, it *may be* that what determines the difference in the relevant contexts is something "mind-dependent"—in which case it would be anti-realist relativism—but it need not be; perhaps what determines the relevant difference is an entirely mind-independent affair, making for an objectivist (and potentially realist) relativism. (Consider: *Tallness* is a relative notion—John is a tall man but a short pro basketball player—but it is not the case that "thinking makes it so.") (Joyce, 2015, emphasis original)

Endorsing relativism and stance-independence with respect to morality is a somewhat awkward position conceptually, but represents a legitimate logical possibility. One might, for instance, endorse a teleological account of morality whereby what is morally good or bad is contingent on one's nature as an organism. And one might believe that different species have different natures, and thus that different moral standards apply to them. One might then maintain that moral claims made by, e.g., humans and some extraterrestrial civilization can be true or false relative to those species, yet still maintain that the moral claims made by these species index different (but still stance-independent) moral standards. I've never heard of anyone advocating for this position, but it is possible.

The other alternative, where one rejects relativism but embraces stance-dependence, *does* appear in the literature on metaethics, primarily in the form of ideal observer theory (e.g., Firth, 1952). One might believe that we should do whatever an ideally rational and fully-informed moral agent would do. If so, moral claims wouldn't be true or false relative to different standards, but they would still depend on an agent's standards, even if it is a hypothetical agent. Ideal observer theory represents one version of what Joyce (2015) describes as *relation-designating accounts*.[28]

---

[28] Joyce attributes this term and distinction to Stevenson (1963, p. 74).

[...] the non-objectivist need not be a relativist. Suppose the moral facts depend on the attitudes or opinions of a particular group or individual (e.g., "*X* is good" means "Caesar approves of *X*," or "The Supreme Court rules in favor of *X*," etc.), and thus moral truth is an entirely mind-dependent affair. Since, in this case, all speakers' moral utterances are made true or false by *the same* mental activity, then this is not strictly speaking a version of relativism, but is, rather, a *relation-designating* account of moral terms [...]

In case there is any doubt that Joyce and I are on (roughly) the same page here, Joyce is explicit about the orthogonality of the distinction, concluding that:

> In short, the *non-objectivism vs. objectivism* and the *relativism vs. absolutism* polarities are orthogonal to each other, and it is the former pair that is usually taken to matter when it comes to characterizing anti-realism. Moral relativism is sometimes thought of as a version of anti-realism, but (short of stipulating usage) there is no basis for this classification; it is better to say that some versions of relativism may be anti-realist and others may be realist. (Joyce, 2015, emphasis original)

Note that even *these* distinctions are inadequate for framing the range of possible views, not because they aren't exhaustive, but because each begins with a particular position, then frames all opposing views as the negation of that position *as though* that were the only or the most natural dichotomy on offer. Only it isn't, because one could start from some other position and then frame the negation of *that* position as a dichotomy. In other words, these distinctions are inadequate not because they are conceptually flawed or imprecise, but because they attempt to force a dispute that admits a variety of qualitatively distinct positions onto a single continuum as though all possible conformed to a single, canonical dichotomy, when this simply isn't the case. In short, the relevant distinctions *aren't true dichotomies*.

Note, for instance, that the negation of stance-independence *is not* stance-dependence. Someone could believe that moral claims entail neither stance-independent nor stance-dependent claims, because they aren't propositional (noncognitivism) or because they endorse pluralism, indeterminacy, or incoherentism. It could also capture the rejection that moral claims are true or false in an indexical or non-indexical way, again, because one could accept noncognitivism or some other

alternative position. Note that one could, in principle, start from the vantage point of describing stance-dependent positions, then frame all opposing positions as non-stance-dependent, even though this would include versions of both realism and antirealism. Likewise, one could start from the idea that all moral claims are nonindexical, then describe all contrary positions as "anti-nonindexical," even though this includes both accounts that indexicalize moral claims (relativism) and positions that don't (noncognitivism).

Any attempt to propose a dichotomized nomenclature will never be adequate because philosophers always have the option of rejecting some putative dichotomy, especially when that dichotomy presupposes a substantive and contested philosophical claim. For instance, both "objectivism" and "subjectivism" are cognitivist positions in that they treat moral claims as propositional. Yet a prominent philosophical position, *noncognitivism*, denies that moral claims are propositional. As such, it would make no sense to treat the only possible positions available to participants as "objectivism" and "relativism" for the simple reason that these *aren't mutually exhaustive possibilities*. The fact that the vast majority of research on folk metaethics foists a false dichotomy on participants and to frame and interpret all results in such terms is a serious oversight.

Instead, it represents just one of several possible antirealist positions. Indeed, *subjectivism* is just *one type* of relativism. There are actually multiple types of relativism, and no consistent pattern of use among philosophers. Philosophers will sometimes on occasion use "relativism" and "subjectivism" interchangeably. It's a mess, and I see no reason to perpetuate sloppy terminological inconsistencies by participating in their use. Unfortunately, even researchers studying folk metaethics have fallen into the trap of treating objectivism and subjectivism as the only possibilities worth evaluating, and may have made this mistake by gravitating towards the intuitively appealing contrast of "objectivism" and "subjectivism." In doing so, they have effectively ignored the possibility of error theory and noncognitivism, and in some cases have mistakenly presumed that subjectivists deny that there are

moral facts (e.g., Theriault et al., 2017). Even when researchers are aware of the distinction between realism and antirealism, they often use "objectivism" as a stand-in for realism, only to contrast it with subjectivism, *as if* the only alternative to objectivism were subjectivism. For example, Goodwin and Darley's (2008) abstract begins by asking: "How do lay individuals think about the objectivity of their ethical beliefs? Do they regard them as factual and objective, or as more subjective and opinion-based, and what might predict such differences?" (p. 1339). This creates the misleading impression of a mutually exhaustive dichotomy, i.e., one must either be an objectivist or a subjectivist. This dichotomy is reflected in their measures, and it took nearly a decade for researchers to begin reliably introducing additional response options for error theory and noncognitivism (e.g., Beebe, 2015).

Of course, *realism* and *antirealism* aren't perfect terms either. People will occasionally read too much into the terms and make mistaken inferences about what they mean. For instance, people will sometimes assume that all antirealists think morality "isn't real," then draw wild conclusions about what antirealists must think based on this misapprehension. The most common may be the insistence that antirealists necessarily deny that there are moral facts of any kind. After all, they're against morality being "real." Yet not all moral realists deny there are moral facts. Moral antirealists only deny that there are stance-independent moral facts. Some people also assume that all antirealists are amoralists or don't have moral standards or attitudes, or at least can't consistently have them if they are genuinely committed to antirealism. This is also untrue. You do not have to think there are stance-independent *gastronomic facts* that dictate which food we should and shouldn't eat in order to have attitudes about what food is good or bad.

With any technical jargon that employs terms with colloquial counterparts, you will inevitably pay a price in confusion and misunderstanding. The alternative would be to coin fully novel terms, yet the price of coining new terms can be even greater. Inventing completely new terminology can confuse everyone, create the misleading impression that you're describing something different from everyone

else, and you can even give the impression that you're arrogant or a crackpot. Insisting on using a bunch of new terms is also cognitively demanding for anyone who hasn't habituated themselves to novel jargon. Anyone who has encountered a paper that coins a slew of neologisms or exhibits an unhealthy obsession with acronyms will be familiar with the irritating task of repeatedly going back to some earlier point in the paper where some term is defined or some acronym is spelled out because you keep forgetting what the terms mean.

It also takes a certain degree of hubris to imagine that you are the terminological Highlander, and that you, and you alone, will be The One to identify the perfect terminology that everyone will start using in perpetuity because it's so perfectly clear that nobody could possibly misunderstand it. People will *always* find inventive ways to completely misconstrue what you're saying and come up with baffling interpretations of things that you say. There may be occasions where new terminology is warranted, but the distinction between realism and antirealism doesn't call for new terminology. The best way forward will, in practice, often be to strike a balance between cleaving existing terminological distinctions and staking one's ground on a particular set of terms and distinctions, even if they are not the most common or well-established. That is what I've done in opting for realism and antirealism.

Finally, there is (3). You might think everyone settled on the same terminology when it comes to the study of folk metaethics, or metaethics in general. Many people may presume that philosophers would agree on a consistent terminology. I'm not going to present a treatise on the sociology and psychology of academic philosophers. Suffice it to say that for a variety of reasons, philosophers passionately refuse to settle on a standardized set of terms, even within a particular subdiscipline. Metaethicists can't even agree on whether to call it "metaethics" or "meta-ethics"! Philosophy is caught in an eternal spiral of disputes and meta-disputes and meta-meta-disputes over its own terms and concepts, with philosophers coining new terms, drawing novel distinctions between old ones, grouping and splitting existing concepts, drawing previously unrecognized connections between

familiar concepts, and repurposing old terms with urbane abandon. They treat the tools of their trade—words and concepts—like eccentric tinkerers, assembling and disassembling them in a frenzied desire to present each other with the philosophical equivalent of the latest gadgets and gizmos. These aren't the sort of people who will readily settle on a shared lexicon.

Unfortunately, this inconsistency has been echoed in research on folk metaethics. There seems to be little or no effort to employ a consistent set of terms. In some cases, researchers use the same terms to refer to different concepts, different terms to refer to the same concepts, or contrast subtly different concepts with one another, but engage with and cite previous research that contrasts similar (but distinct) concepts. Unfortunately, the only exception to this is the use of "objectivism," which is a poor choice of term for the reasons I highlight above, and for additional reasons discussed below. Here are a few examples that highlight the incredible lack of consistency in the terms contrasted with "objectivism" (and, on occasion, with terms used alongside or in place of objectivism) in **Table S1.1.**[29]

**Table S1.1**

*Terms and distinctions in folk metaethics research*

| Article | Realism/similar concepts | Antirealism/similar concepts |
|---|---|---|
| Nichols & Folds-Bennett (2003) | nonrelativism, objectivism | anti-objectivism, relativism, response-dependence |
| Nichols (2004) | objectivism, realism | anti-objectivism, nonobjectivism, relativism |
| Wainryb et al. (2004) | absolutism, objectivism, realism | relativism, subjectivism |
| Goodwin & Darley (2008) | objectivism | non-objectivism, subjectivism |
| Sarkissian et al. (2011) | absolutism, objectivism | relativism |
| Goodwin & Darley (2012) | objectivism | subjectivism |
| Heiphetz et al. (2013) | objectivism | relativism |

[29] This list is not exhaustive. All instances of terms like "objective" were converted into their -ism counterpart. Note also that not all articles conflate or fail to distinguish the relevant terms and concepts.

| | | |
|---|---|---|
| Rai & Holyoak (2013) | absolutism, objectivism, universalism | relativism, subjectivism |
| Wright, Grandjean, & McWhite (2013) | absolutism, objectivism | relativism |
| Young & Durwin (2013a) | objectivism, realism | antirealism, subjectivism |
| Beebe (2014) | objectivism, universalism | relativism, subjectivism |
| Beebe (2015) | objectivism, realism | relativism, subjectivism |
| Beebe et al. (2015) | objectivism | non-objectivism, non-realism relativism, subjectivism |
| Beebe & Sackris (2016) | objectivism, realism | relativism, subjectivism |
| Khoo & Knobe (2016) | objectivism | relativism |
| Moss (2017) | objectivism | relativism, subjectivism, anti-realism |
| Pölzler (2017) | objectivism, realism | anti-realism, relativism, response dependence, subjectivism |
| Fisher et al. (2017) | objectivism | subjectivism |
| Heiphetz & Young (2017) | objectivism | relativism |
| Schmidt, Gonzalez-Cabrera, & Tomasello (2017) | objectivism | relativism, subjectivism |
| Theriault et al. (2017) | objectivism | subjectivism |
| Collier-Spruel et al. (2019) | objectivism | relativism |
| Pölzler (2018b) | objectivism, observer-independent realism | anti-realism, relativism, subjectivism |
| Yilmaz & Bahçekapili (2018) | objectivism, absolutism | subjectivism |
| Pölzler & Wright (2019) | objectivism, realism | non-objectivism, nonobjectivism, relativism |
| Rose & Nichols (2019) | absolutism, universalism* | relativism |
| Viciana, Hannikainen, & Torres (2019) | objectivism | non-objectivism, relativism |
| Zijlstra (2019) | objectivism | relativism, subjectivism |
| Ayars & Nichols (2020) | universalism* | relativism, subjectivism |

| | | |
|---|---|---|
| Beebe (2020) | mind-independence, objectivism, nonrelativism, realism | mind-dependence, nonobjectivism, relativism, subjectivism |
| Pölzler & Wright (2020a) | objectivism | anti-objectivism, subjectivism |
| Pölzler & Wright (2020b) | objectivism, realism | anti-realism, relativism, subjectivism |
| Theriault et al. (2020) | objectivism | subjectivism |
| Yilmaz et al. (2020) | objectivism, absolutism | subjectivism |
| Davis (2021) | objectivism, realism | antirealism, relativism, subjectivism |
| Sousa et al. (2021) | objectivism | non-objectivism, subjectivism |
| Viciana, Hannikainen, & Rodríguez-Arias (2021) | objectivism, absolutism | relativism, subjectivism, relativism, nihilism |
| Wagner, Pölzler, Wright, (2021) | objectivism | non-objectivism, subjectivism |
| Zijlstra (2021) | objectivism, realism | antirealism, subjectivism, relativism, non-objectivism |

*Note.* Ayars and Nichols (2020) and Rose and Nichols (2020) explicitly distinguish universalism from objectivism and realism, and thus *do not* conflate universalism with either. Nevertheless, they contrast universalism with relativism. This is a different contrast than the contrast between objectivism and relativism.

Notice that while *objectivism* appears to be used fairly consistently, it is not consistently matched with any particular term. When it is, it's typically *relativism* or *subjectivism*. This inconsistency vindicates my objection to the use of "objectivism" as the primary term for the notion that there are stance-independent moral facts: it has no consistent, natural, mutually exhaustive negation term. Variants of anti- and non- objectivism never caught on, and probably never will. Instead, people consistently use "subjectivism" or "relativism" as natural contrasts. Yet, as I've argued above, relativism and subjectivism concern a dichotomy that is orthogonal to disputes about stance-independence, and are technically compatible with it. As such, objectivism (understood as stance-independence) and relativism/subjectivism aren't merely a false dichotomy because they aren't mutually exhaustive, they *aren't even at odds with one another because they represent stances towards fundamentally different issues*: objectivism

is about stance-independence, relativism/subjectivism are about indexicality. The entire literature has operationalized a pair of putatively conflicting constructs that don't actually conflict with one another. *Even if* researchers narrowed their focus to exclusively stance-dependent concepts of relativism/subjectivism, this still wouldn't be adequate, because, if "objectivism" refers to stance-independence (as it appears to in the literature), the negation is the rejection of stance-independence, *not* stance-dependence: stance-independence vs. stance-independence is a false dichotomy that presupposes cognitivism. Participants frequently choose noncognitivist options when they are presented (e.g., Beebe, 2015; Davis, 2021). The one time researchers do opt to employ consistent terminology, they do so in a way that conflates different distinctions, yielding a literature riddled with muddled, inconsistent, and conceptually confused terminology that has negatively impacted the way folk metaethical constructs have been operationalized.

Researchers often conflate or mischaracterize metaethical distinctions, or mislabel the concepts they describe. On occasion, they'll employ different terms to refer to the same concept in the same article, or cite previous articles that used one set of terms, but describe the content of that article using a different set of terms, e.g., they'll describe an article assessing "objectivism" as an article assessing "realism," often without acknowledging the terminological discrepancies. Here a handful of examples:

### Wainryb et al. (2004)

Objectivism and absolutism are used interchangeably. They state that "[...] children progress to a position of absolutism or objectivism [...]" (Wainryb et al., 2004, p. 688). However, they also present their version of the disagreement paradigm as a contrast between relativism and "nonrelativism" (p. 692). Pölzler and Wright (2020b) describe Wainryb et al.'s research by stating that "According to Wainryb et al., "only one belief is right" responses express realist intuitions, and "both beliefs are

right" responses express anti-realist intuitions" (p. 56), but add in a footnote that Wainryb et al. use the terms "objectivism" and "relativism" (p. 56, footnote 3).

### Sarkissian et al. (2011)

They ask: "Do people believe in objective moral truth, or do they accept some form of moral relativism?" (p. 483). This supports my contention that researchers frame objectivism and relativism/subjectivism as a dichotomy. They also use the term "absolutism" interchangeably with "objectivism": "Results of many studies have thus far suggested that people reject relativism about morality, and believe instead in some type of absolute moral

truth." (p. 483).

### Rai & Holyoak (2013)

Interestingly, Rai and Holyoak (2013) use the term "absolutism" to characterize realism. They state that "The philosophical position of *moral absolutism* holds that some moral beliefs are objectively true, and reflect facts that are independent of any social group's specific preferences." (p. 995). Critically, they then state that "*On the other end of the spectrum*, the philosophical position of moral relativism holds that the truth or falsity of moral beliefs are products of our traditions and cultural histories, rather than objective statements based on logic, or facts about the state of the world independent of our own opinions or perspectives." (p. 995, emphasis mine). As I have demonstrated, there is no spectrum. These positions are conceptually orthogonal. Note, as well, that they frame these as the *only* possibilities by suggesting there's a single spectrum, and that relativism anchors the other end of that spectrum.

### Wright, Grandjean, & McWhite (2013)

They state that "Objectivism holds that the moral domain, like the scientific domain, is grounded in universal and fundamental facts that exist (largely) independently of people's beliefs, preferences,

attitudes, norms, or conventions." (p. 1). Objectivism, understood as stance-independence, does *not* entail that the moral facts in question are "universal."

### Young & Durwin (2013)

Uncharacteristically, this article employs the same terms I do: realism and antirealism. They describe moral realism as the view that "objective moral facts exist" (p. 302). This is accurate as far as it goes. However, they mischaracterize moral antirealism. They state that "Moral *antirealists* deny the existence of moral facts, maintaining that there are no real answers to moral questions [...]" (p. 302). This is simply not true. Relativists explicitly acknowledge that there are moral facts, and moral antirealists do not necessarily claim that there are no "real" answers to moral questions. This use of "real" seems to presume that only realist conceptions of answers to moral questions are "real," which is not something an antirealist is obliged to accept. In their attempt to clarify the antirealist position, they unfortunately only muddle the distinction further. They add that "Importantly, moral antirealists do not deny the existence and importance of moral values; antirealists simply assert that moral values reflect the beliefs of a person or a culture, rather than immutable facts that exist independent of human psychology." (p. 302). This is an unconventional distinction between "moral facts" and "moral values." Moral relativists typically do think there are moral facts, they just think those facts have indexical truth conditions. Worse still, in stating that "antirealists" assert that "moral values reflect the beliefs of a person or culture," they conflate antirealism with relativism/subjectivism, implying that all antirealists endorse some form of relativism. So they simultaneously mischaracterize antirealism by implying all antirealists are relativists/subjectivists, *and* mischaracterize relativism and subjectivism by implying that such views deny that there are moral facts.

### Pölzler and Wright (2020b)

They state that "Researchers have typically assumed adequate definitions of moral realism and anti-realism," yet, they add, "[...] these definitions have then not been properly operationalized." (p. 56). I

agree that researchers often capture the appropriate distinction (e.g., Goodwin and Darley, 2008), but fail to operationalize the intended distinction. However, I don't agree that researchers have typically assumed adequate definitions. As this discussion should hopefully illustrate, there are considerable inadequacies in the way researchers have framed metaethical distinctions, not just in how they've operationalized them.

Given the terminological inconsistencies both across and within studies, the inadequacy of the few terms that are used consistently (e.g., "objectivism"), frequent mislabeling and mischaracterization of and mislabeling of the relevant terms and concepts, and the misoperationalization of terms that are accurately labeled and characterized, I think I'm on pretty safe footing in declaring *realism* and *antirealism* to be a better choice for how to frame research on folk metaethics.

## SUPPLEMENT TO CHAPTER 2

### S2.1 The empirical aspirations of metaethics

The central task of metaethics is to address questions about the nature of morality: *Are there moral facts? If so, what makes them true? And how can we discover what these moral facts are?* Whatever the answer to these questions may be, we won't find them in telescopes or test tubes. So how are we supposed to answer them? Any attempt to answer such questions must begin with what we mean by *moral facts*, and the meaning of moral terms and concepts in general. This does not mean that the existence of moral facts is contingent on how we speak or think (Kahane, 2013; Loeb, 2008). But it does mean that whether a given term describes something that exists or has certain properties will depend on what that term refers to.

So what do moral terms and concepts refer to? *Morality* is not an obscure technical term. Moral thought and discourse are a part of our everyday experience. Moral considerations arise not only in war, political disputes, and the courtroom, but emerge even in the most mundane aspects of our lives, from gossip about celebrity scandals to the decisions we make in the grocery store. And moral disputes can occur in any social context, from philosophy departments, to dive bars, to kindergarten classrooms. Presumably, answers to questions about what morality refers to should begin with an assessment of what these situations have in common, and what ordinary people engaged in moral discussions are referring to when they use moral terms and concepts in these circumstances.

Philosophers have done just this by engaging in *descriptive metaethics* (Gill, 2009). The goal of descriptive metaethics is to provide an account of ordinary moral thought and language, i.e., an account of *folk metaethics*. While there is ongoing debate about the philosophical relevance of folk metaethics, there is little dispute that it has at least some relevance to philosophical disputes; the only question is *how much*. Some philosophers have argued that the central questions of metaethics, such as

the metaphysical status of moral truth claims, turn on discoveries about folk metaethics (Joyce, 2001; Loeb, 2008; Mackie, 1990). Others offer a more modest appraisal of the role folk metaethics plays in adjudicating philosophical disputes (e.g., Kahane, 2013; Sinclair, 2012). For instance, it is typically taken to count in favor of a philosophical position that it accords with ordinary thought and practice (Fuqua, 2021; Greco, 2014; Joyce, 2021; Lycan, 2019; Sinclair, 2012; Yasenchuk, 1997), e.g., if most ordinary are committed to or endorse moral realism or moral realism *seems true* (Huemer, 2007) this serves as some defeasible evidence in favor of realism and foists the burden of proof on antirealists to show why realism is mistaken.[30] In practice, moral realists often present their position as the default view precisely because they believe ordinary people are moral realists and that ordinary language best fits a realist analysis. For instance, Smith (1994) states that "we seem to think moral questions have correct answers; that the correct answers are made correct by objective moral facts" (p. 6), while McNaughton (1988) states that:

> The realist insists on an obvious, but crucial, methodological point: there is a presumption that things are the way we experience them as being […]. Moral value is presented to us as something independent of our beliefs or feelings about it; something which may require careful thought or attention to be discovered. There is a presumption, therefore, that there is a moral reality to which we can be genuinely sensitive. (p. 40, as quoted in Pölzler & Wright, 2020b, p. 54)

These authors appear to presume that, just as most of us experience trees and tables as "real," we likewise experience morality as "real," i.e., that the commonsense conception of morality is realist in nature, and our experience of the world is one in which moral facts seem to be a part of the world around us, *not* merely an expression of our emotions or subjective standards.

---

[30] If this seems odd, consider how the burden of proof would operate were the apparent truth far more evident. Suppose a group of people were having a discussion when, to all appearances, an elephant stomped into the room. Of the dozen people in the room, all but one noticed the elephant, came to believe there was, in fact, *an elephant in the room* and addressed the elephant in the room. But suppose the twelfth person declared that they saw no such elephant. In such circumstances, it would be plausible for everyone else to expect the person who would not address the elephant in the room to provide some argument or justification as to why, despite all appearances to the contrary, there was, in fact, no elephant. Just the same, if most people speak and think as moral realists, and on reflection, it *seems to them* like realism is true, then it seems reasonable to ask anyone who denies this is the case to explain why, appearances to the contrary, moral realism is false.

Even critics of moral realism appeal to features of ordinary thought and language to sustain their objections. Noncognitivists, error theorists, and relativists all maintain that certain features of how we think and speak about morality are best captured by noncognitivism such as their conative or affective characteristics (van Roojen, 2018). The presumed relevance of folk metaethics is so entrenched in the way metaethicists engage with one another that metaethical positions often incorporate semantic theses into the substantive content of their positions (Kahane, 2013; Sinnott-Armstrong,2009). This is as true of antirealists as it is of realists. For instance, it is especially clear in the case of *moral error theory*. An essential component of error theory *just is* a semantic thesis about the meaning of moral claims. All versions of error theory hold that moral claims are implicitly committed to one or more false presuppositions and that, as a result, all moral claims are false. For instance, Richard Joyce (2001), defends a version of error theory with two central theses:

(1) *Conceptual thesis*: Moral claims aim to describe (or *refer*) to particular facts or properties that purportedly exist

(2) *Substantive thesis*: The facts or properties moral claims refer to do not exist (Tully, 2014)

As a result of (1) and (2), all moral claims are false. In other words, when people make moral claims (e.g., "murder is wrong") they are attempting to describe the world…they just fail to do so. Joyce likens error theory to how we think about historical references to *witches*. People who use the term *witch* are attempting to attribute certain properties to people, e.g., that they *consort with demons* or *cast spells*. Yet no people have these abilities, so all such claims fail to refer to any properties people actually have. As such, attempts to describe someone as a witch are false. Likewise, insofar as moral claims attempt to describe properties that don't exist, moral claims are similarly false.

Error theory depends just as much on (1) as it does on (2). Claims about *witches* would only be false if they really did attempt to describe people that consorted with demons or cast spells. But they wouldn't be false if they referred to an entirely different set of properties that did describe some feature

of reality, e.g., *old widows that live alone* or *unlicensed apothecaries*. Since there really are people who fit these descriptions, there really would be witches. This shows that the practical relevance of error theory depends on whether it accurately captures the meaning of some *actual* discourse. It is trivially true that *if* moral claims referred to nonexistent properties, then moral claims would be false. But do they? Answering this question would require assessing moral language as it is spoken by some population. This would require us to specify *whose* moral claims are captured by the conceptual thesis. And error theorists are typically offering an account of what *ordinary people* mean when they make moral claims:

> *The ordinary user* of moral language means to say something about whatever it is he characterizes morally, for example a possible action, as it is in itself, or would be if it were realized, and not about, or even simply expressive of his or anyone else's relation to it. (Mackie, 1977, p. 33, as quoted in Sinnott-Armstrong, 2009, p. 238, emphasis mine)

The same generally holds for other metaethical accounts that defend some form of realism or antirealism: most such accounts attempt to demonstrate that their brand of realism or antirealism offers the best account of ordinary moral thought and language.

Given that all of these accounts purportedly describe how people are disposed to *think*, *speak*, and *act*, they would seem amenable to empirical analysis. I am not the first to suggest this is the case. Loeb says that when it comes to questions about folk metaethics, "the matter to be investigated consists largely of *empirical* questions" (p. 798, emphasis original). Striking examples of the empirical aspirations of metaethics can even be found in the way prominent metaethicists characterize the field. In the Stanford Encyclopedia of Philosophy, Sayre-McCord (2012) describes metaethics as "the attempt to understand the metaphysical, epistemological, semantic, and psychological, *presuppositions and commitments of moral thought, talk, and practice*" (emphasis mine). Sayre-McCord may quibble over precisely just whose moral thought, talk, and practice he has in mind, but there is little doubt he is referring to ordinary moral thought and language, with the reasonable exclusion of incompetent or

idiosyncratic speakers.[31] And what are moral *thought*, *talk*, and *practice* if not phenomena that could be

(and already are!) studied by social and cognitive science?[32]

     That folk metaethics consists largely of empirical claims does not mean that philosophers must

step aside and make way for scientists, nor that every metaethical position must appeal to empirical

data. Some philosophers argue that metaphysical questions central to metaethics don't depend on

empirical facts about folk metaethics at all (e.g., *are there non-natural moral properties?* see Kahane, 2013).

And philosophers can (and sometimes do) defend invitations to speak or think about morality in a

certain way, regardless of how ordinary people speak or think (e.g., fictionalism, see Joyce, 2001;

Kalderon, 2005). Yet even in these cases, such discussions would be hard to interpret as discussions

about *morality* without first establishing what moral thought and language is *about*. After all, exhorting

us to explicitly adopt a *different* way of thinking and speaking about morality only makes sense if we

*already* thought and spoke about morality in some other way. And the force of arguments for

---

[31] I consider the latter exclusions completely acceptable, and my impression is that this is generally uncontroversial. Kauppinen (2007), for instance, states that "It should be obvious that when philosophers appeal to 'us' in making their claims, the extension is limited to those who are competent with the concept in question" (p. 102). This hardly seems worth defending, but *if* we did not permit such exclusions, it would hardly count in favor of determinacy, and if anything would only bolster my position. Hence, it is an allowance that if anything slightly favors uniformity and determinacy.

[32] This definition is incomplete, and Sayre-McCord goes on to discuss metaethical puzzles that can be addressed without direct appeal to the empirical sciences. Notice, however, that after providing the definition above, Sayre-McCord goes on to say "[...] As such, it [metaethics] counts within its domain a broad range of questions and puzzles…" *As such*? Sayre-McCord hints that these non-empirical metaethical questions are in some way related to empirical questions about the "presuppositions and commitments of moral thought, talk, and practice." But *whose* moral thought, talk, and practice is he referring to? He does not specify, but elsewhere, Sayre-McCord (2009) cryptically maintains that such questions concern what "*'we'* are doing in thinking and talking about what *'we'* characterize as morality" (p. 934). He continues:

"The 'we' is not so capacious as to include all who use the terms 'right', 'wrong', 'moral' and 'immoral', yet it is supposed to include those who speak languages other than English (in cases where they have terms that are properly translated by our terms 'right', 'wrong', 'moral' and 'immoral', etc.) and it is meant as well to identify a group of people who can properly be seen as all thinking and talking about (as we would put it) what is right and wrong, moral or immoral." (p. 934)

Sayre-McCord's conception of *we* is roughly in accord with my conception of *ordinary people*. Although his conception of *we* excludes by stipulation people that are not engaged in genuine moral thought and discourse, his primary concern seems to be to exclude people whose thought and language is deviant, perhaps due to incompetence, unfamiliarity, or idiosyncrasy (or even insincerity). Sayre-McCord furnishes the example of a person who insists that "God is love and mystery." And, since love and mystery clearly exist, obviously, *God exists*. Anyone interested in whether God exists is not likely to be impressed by this brand of casuistry, but we all recognize it. When pressed, people will bend and contort terms in the service of some disingenuous or at best misleading argument, and this will sometimes result in employing a word that plausibly has some central or primary cluster of referents in a way that is so strained we are justified in simply denying that their use of the term is appropriate. *God* plausibly refers to some supernatural being or force.

revisionary accounts of how we should use moral terms concepts will depend in part on how much of a departure their account is from how we are already disposed to think about morality, and how negotiable we find our commitments to our current terms and concepts, since any account that is *too* revisionary risks changing the subject (Loeb, 2008, p. 826).

In sum, even if our conclusions of metaethical inquiry don't end with whatever people are referring to when they engage in moral discourse, any satisfactory account of metaethics should at least begin there. And since such questions are empirical, the scientific study of folk metaethics is a natural place to start. Philosophers may be content to ponder the nature of morality from the armchair, but at least part of the grist for their mill must begin with ordinary people: what they *say*, what they *think*, and what they *do*. And the final arbiter of the content of folk metaethics is empirical data, not armchair speculation.

## S2.2 Uniformism, pluralism, & indeterminism

Since a great deal of metaethics is concerned with descriptive questions about ordinary moral thought and language, it is reasonable to wonder why moral philosophers have not traditionally engaged with or conducted empirical research (Fraser, 2014). I cannot provide a complete explanation for why they have failed to do so. But I can at least gesture towards some reasons why they have not.

A host of sociological and institutional explanations may account for much of the failure of philosophers to engage with or conduct empirical research. There is substantial pressure to specialize, and a person trained in analytic philosophy may lack the requisite competence to conduct or evaluate research in linguistics, psychology, or anthropology. Even if they were so inclined, there may be little incentive to do so. Philosophers can be territorial. Directly engaging with other fields, especially the empirical sciences, may be seen as methodological heresy, or at best misguided or of little value in advancing the field. Status and career advancement might instead be predicated on producing

publications of a certain kind (e.g., *non-empirical* philosophical work), discouraging philosophers from interdisciplinary pursuits.

Given the lukewarm reception, and even contempt philosophers have shown for experimental philosophy[33], this could be a substantial impediment.[34] Coupled with norms that favor single authorship over collaboration and coauthorship, lack of institutional support for cross-disciplinary collaborations between philosophers and scientists, and induction into a discipline that has developed a snobbish disdain for the plebeian study of *concreta* over *abstracta*, academic philosophy shows all the hallmarks of a discipline that has walled itself off from the sciences.

Even in the absence of these barriers, we may still have seen little engagement with the empirical study of folk metaethics, at least with respect to whether people are realists. The presumption that ordinary people are moral realists may be so common among philosophers that it didn't seem necessary to gather empirical data. This may *still* be the case. While philosophers would probably acknowledge that some forms of antirealism are popular among some subpopulations, such as relativism among college students (Beebe & Sackris, 2016; Nichols, 2004; Kohlberg & Kramer, 1969; Paden, 1994; Pfister, 2019; Satris, 1986), such populations may be dismissed as outliers who deviate from an otherwise nearly universal commitment to realism. Even antirealists typically either concede that it seems like ordinary people are moral realists, but claim that people are simply mistaken (e.g., error theorists), or concede much of the way people speak and think at least *seems* realist, even if it isn't. In short, insofar as philosophers have traditionally considered the answer *obvious*, there was

---

[33] It's difficult to provide a reference for this claim. I have little more to appeal to in this case other than my own experiences and to call on the experiences of people associated with experimental philosophy.

[34] When applying for PhD programs, I was explicitly discouraged from emphasizing my interest in addressing traditionally philosophical questions via empirical research, and on more than one occasion colleagues objected that my work "wasn't philosophy." I remain puzzled as to the nature of this objection. Even if I grant that my work wasn't philosophy, *so what*? If the best tool for addressing a philosophical question isn't philosophy, this seems more like an objection to philosophy than to my work. It's a sad irony that philosophers raising this objection don't recognize their unreflective conformity to a self-limiting methodological toolkit. If fundamental questions of metaphysics called for studying paleontology, then I expect philosophers to break out the picks and chisels.

simply no need to gather data. And it doesn't help that most philosophers are moral realists (Bourget and Chalmers; 2014; 2021), and may be subject to the *typical mind fallacy*[35], the tendency to erroneously presume others think like oneself (Alexander, 2009; Bervoets, Milton, & Van de Cruys, 2021).[36]

Dovetailing with this explanation, philosophers seem to consider armchair methods (e.g., conceptual analysis) adequate for resolving questions about folk metaethics. This can be readily inferred by looking at how they defend their preferred analyses of folk metaethics. Gill (2009) describes a two-stage process that exemplifies the primary method philosophers have historically employed: First, gather examples of ordinary moral thought and discourse along with putative instances of commonsense moral judgments and intuitions. Then attempt to demonstrate that a particular semantic analysis does a better job of accommodating the commitments, intuitions, and platitudes present in everyday use of moral terms and concepts than rival analyses. For instance, both realists and antirealists point to features of ordinary moral discourse that purportedly fit with their preferred analysis. And while each side acknowledges that there are features of how people speak or think that fit more naturally with rival analyses, these outliers can be explained away more readily than other accounts can explain away features of folk metaethics that don't conform to their own analyses.

There are two striking features of this procedure. First, it doesn't involve any appeal to systematic, representative sampling of ordinary people. Instead, philosophers assume that they can extrapolate from their own judgments and intuitions about paradigmatic moral sentences to people in general. Second, such analyses presume that there is a single uniform and determinate analysis of folk metaethics. This means that *all* ordinary people think and speak about morality in the same way, and

---

[35] The typical mind fallacy seems to be attributed to William James who referred to it as the psychologist's fallacy: "The great snare of the psychologist is the confusion of his own standpoint with that of the mental fact about which he is making his report. I shall hereafter call this the 'psychologist's fallacy' par excellence." (James, 1890, p. 196)

[36] There may also be a selection effect among people engaged in metaethics to be moral realists. Discussing the nature of morality may be less appealing to people who believe much of the discussion is a load of confused nonsense for reasons similar to the unsurprising lack of atheists among theologians.

that this shared metaethical standard exclusively conforms one or the other traditional metaethical positions, e.g., either *cognitivism* or *noncognitivism*, *realism* or *antirealism*, etc., which are depicted as "mutually exclusive and jointly exhaustive" accounts of folk metaethics (p. 218). That is, *folk metaethics* is *either* cognitivist or noncognitivist, realist or antirealist, and so on. Gill refers to this as the Uniformity-Determinacy (UD) assumption. The UD assumption is implicit in the vast majority of arguments for particular metaethical positions.[37] I have sometimes encountered skepticism about the prevalence of the UD assumption. I provide evidence of its prevalence in Supplement 1.[38]

*This* is the assumption that I hope to overturn, and it was Gill who first challenged the UD assumption by proposing the *Indeterminacy-Variability (IV)* thesis. Gill suggests that folk metaethics may exhibit some unspecified degree of *indeterminacy* and *variability*, but leaves exact proportion or ways in which folk metaethics is indeterminate or is instead determinate but variable. Since most research on folk metaethics characterizes variability as *pluralism*, I will stick with the latter term.

This leaves us with three competing hypotheses about folk metaethics: *uniformism*, *pluralism*, and *indeterminism*.

> *(i)* **Uniformism** holds that there is a single determinate account of folk metaethics with respect to a particular metaethical distinction.

> *(ii)* **Pluralism** holds that there is ineliminable but determinate variability with respect to one or more traditionally competing accounts of folk metaethics.

> (iii) **Indeterminism** holds that there is no determinate position that characterizes some or all folk metaethics with respect to one or more metaethical distinctions.

---

[37] According to Gill, the UD assumption was prevalent throughout much of 20th century metaethics, though I suspect it has persisted up to the present as well.

[38] When discussing this topic with people, I have sometimes encountered skepticism that philosophers could really be so adamant about the uniformity and determinacy of folk metaethics. The possibility that people hold fundamentally different views from another, or are riddled with inconsistent beliefs, or don't have any particular stance or commitments on a matter often strike people as plausible, even obvious. Surely philosophers realize that their totalistic language ignores very real and committed folk diversity (WEIRD)? In truth, some philosophers *have* commented on the implausibility of something like the UD assumption. [Blackford quote, Kane B remark]. But such comments are rare, and almost never appear in print.

Pluralism and indeterminism are not mutually incompatible, but may instead characterize subsets of folk thought and language. According to Gill (2009), it is:

> ...quite likely that meta-ethical indeterminacy characterizes much of ordinary discourse. But it may not characterize all of it. The best descriptive analysis of some other uses of moral terms might involve robust meta-ethical commitments. Those commitments, however, might not all be uniformly consistent with one side of the traditional meta-ethical debate over the other.[39] (p. 218)

I agree. But where Gill and I differ is that I emphasize indeterminacy to a much greater extent. Each of the three hypotheses may also apply to one dispute but not another, and are thus not incompatible in that respect, either. It may be that folk metaethics is determinately cognitivist, and that prototypical moral claims are thus best understood as propositional claims (i.e. *uniformism* towards cognitivism), while there could also be no determinate answer as to whether these assertions are relative or not (i.e. *indeterminism* towards relativism vs. nonrelativism), or there could be instances of relativism and nonrelativism that cannot be explained away as conceptually confused or nonstandard (i.e. *pluralism* about relativism vs. nonrelativism, Gill, 2008, p. 218).

Finally, pluralism can come in multiple forms. There could be stable individual differences in metaethical stances and commitments, or *interpersonal variability*, e.g., some individuals consistently think or speak in realist terms while others think or speak in antirealist terms. But there could also be *intrapersonal variability*, e.g., individuals could have different metaethical stances or commitments towards different moral issues.[40] For instance, someone could be realists about whether murder wrong, but adopt a relativist stance towards abortion. There may even be *context variability*, where a person

---

[39] Note that the compatibility of pluralism and indeterminacy could take multiple forms. It could be that some people's moral thought and language is consistently determinate but other people's moral thought and language is consistently indeterminate, or it could be that moral thought and language tend to be determinate in particular contexts but not others, or towards particular moral issues but not others, or among the members of some communities but not others, or among adults but not children, etc. There are too many possibilities for me to entertain all of them, but it is worth noting that the possible ways indeterminacy and determinate variation could manifest suggests folk metaethics could be incredibly messy.
[40] There are a variety of possible ways that the meaning of moral claims could vary with respect to various metaethical positions. The meaning of moral claims could vary depending on the tense, the social context one is in, the targets of one's moral judgments, the specific terminology one uses, and so on.

exhibits a stable tendency to speak or think in accordance with a particular metaethical stance in one social setting (e.g., at work), but to speak or think in accordance with a different stance in another social setting (e.g., church).

All of this potential nuance is enough to scare away even the most ardent moral psychologist. Independently evaluating the plausibility of each of these hypotheses for each of the distinctions that interests metaethicists could fill volumes. This is one reason why I focus exclusively on the dispute between realism and antirealism. In doing so, I am potentially oversimplifying the space of possibilities and conflating or ignoring considerations that may be relevant to a more fine-grained characterization of folk metaethics, but this seems like an acceptable sacrifice both because available empirical data is grossly inadequate to tease these questions apart in a satisfactory way and because doing so would be beyond the scope of even a very long discussion.

Setting these complications aside, my only concern will be whether uniformism, pluralism, or indeterminism (or some combination of the latter two) provides the best account of folk metaethics *with respect to moral realism and antirealism*. At present, the vast majority of published interpretations of folk metaethical data support either uniformism or pluralism. To my knowledge, no one has argued that indeterminism (with a qualified splash of pluralism) may offer a better explanation for existing findings, and is the position most likely to be vindicated by future research. In the remainder of this chapter, I review the state of the empirical research on folk metaethics and argue that existing data is too methodologically flawed to support uniformism or pluralism.

### S2.3 Inadequate response options

There are two general reasons why response options are inadequate: (1) they tend to present only a limited selection of available metaethical positions, and (2) response options typically lack the specificity to distinguish particular metaethical positions from one another. While these problems could in principle be corrected by including more response options, doing so comes at a significant

cost in length, complexity, and risk of introducing ambiguity. The poor specificity of questions is typically bad enough that, even if participants were realists or antirealists, response options interpreted as indications of an opposing view would be consistent with their positions, threatening the validity of such measures.

### S2.3.1 Response options are not exhaustive

Response options are often not exhaustive. I discuss the lack of distinct options below, but a more general issue is the tendency to present participants with the opportunity to endorse or reject a metaethical position, even though the rejection of that position is too non-specific to reflect a meaningful metaethical position. For instance, discovering that someone rejects realism does not entail that they endorse relativism, yet some scales appear to interpret responses in this way (e.g., Collier-Spruel et al., 2019).

### S2.3.1.1 Missing distinct option for noncognitivism

One problem with standard versions of the disagreement paradigm is that it does not offer response options that reflect the full range of possible metaethical stances and commitments. Standard versions of the disagreement paradigm only allows participants to respond that either *both can be correct* or *at least one must be incorrect*. These response options seem to reflect *relativism* and *realism*, respectively. However, relativism is just one form of antirealism. By requiring participants to choose whether moral truth is either relative or nonrelative, participants that deny moral claims can be true or false have no way to properly express this view. In other words, there is no response option for *noncognitivism*, the view that moral claims are not truth-apt, because both response options presuppose cognitivism (Beebe, 2015).

If there were strong *a priori* grounds to suspect folk noncognitivism were extremely unlikely, excluding a noncognitivist response option might be a safe bet. However, this bet would not pay off. When Beebe (2015) included a response option for noncognitivism ("neither belief is true or false"),

it proved incredibly popular. Out of seven moral claims, it was the most common response for three, and was chosen more frequently than the *relativist* option for the remaining four items. In a more recent study, Davis (2021) also found that noncognitivist responses were *the most common response option* among the vast majority of participants in his sample, comprising 34.2% of responses when aggregating across all conditions (p. 17). These findings demonstrate that many participants may favor versions of antirealism other than relativism, but have no way to express their views. As a result, we cannot estimate the proportion of people who are noncognitivists rather than relativists. And because we cannot know what proportion of noncognitivists would favor the *relativist* response option over the *realist* response (at least not without further, independent evidence), versions of the disagreement paradigm that exclude a response option for noncognitivists cannot provide an accurate a valid estimate of the true proportion of realists and antirealists.

### S2.3.1.2 Missing distinct option for error theory

There are also no appealing response options for error theory (the view that all first-order moral claims are false). A response option such as "both people are incorrect" might reflect error theory, but even this may not be adequate, since an ideal measure of error theory would allow the participant to express the view that *all* moral claims are false. If a participant selects *"both people are incorrect"* for some moral issues, but not for others, it is at least conceptually possible to imagine that this person is an error theorist about some moral issues but not others. Yet this would be a very bizarre position to hold. Error theory is typically understood to represent the view that moral claims are *uniformly* committed to some mistaken presupposition.

While it is possible in principle to hold that *some* moral claims are subject to an error theory but others are not, this would be a very strange view to hold. That a person judges that both people are incorrect for some moral issues but not others would more plausibly represent something other than such a view. But even if we grant the possibility of folk pluralistic error theory, it is possible for

someone to judge that two people could be incorrect about seemingly conflicting moral claim*s without this reflecting error theory*. Suppose you believe that abortion is permissible under some circumstances, but not under others. If you are told about a disagreement between two people, one of whom maintains that abortion is morally wrong, and another who does not, you may interpret the first person to believe abortion is *always* wrong and the latter to believe it is *never* wrong, or some equivalent insensitivity to variation in the circumstances that you regard as relevant to whether or not a given instance of abortion is wrong. If so, you may judge that both are incorrect, *even if* you endorse realism or relativism. This highlights one serious shortcoming with the disagreement paradigm: to interpret the questions as intended, participants *must view the disagreement to be mutually exclusive*.

Among philosophers, it may seem obvious that "X is wrong" and "X is not wrong," are intended to reflect jointly exhaustive and mutually exclusive claims, but it is unlikely that ordinary people would be uniformly inclined to interpret statements this way; indeed, this may be an extremely uncommon way for people to interpret such statements. Rather, people may believe these statements reflect something more closely approximating two extreme ends of a continuum, according to which X is *always* wrong or *never* wrong. The participants themselves may judge that X is wrong *in some circumstances* but *not others*. Even when researchers specify *some* of the contextual details that may be relevant, they cannot (and don't attempt) to provide all of them. For instance, take this ethical statement used by Goodwin and Darley (2008):

> Providing false testimony in court about the whereabouts of a friend who is being charged with murder (i.e., to protect that friend by offering an alibi) is morally permissible.

It would be entirely reasonable to deny that this must be wrong or not wrong, and to instead insist that *it depends*. Why are they being charged with murder? Were they framed? Do you have knowledge that the friend committed the murder? If they did, would the murder have been justified? The authors don't even specify whether you have reason to believe that the friend committed the murder! These

considerations *may* be irrelevant, but it is bizarre to assume that they aren't, and that one must be able to offer a decisive answer to this question. Realists and others who do not endorse error theory might readily judge that any decisive answer to this question is mistaken without this entailing that they believe all moral claims are systematically false. Thus, there may be no easy way to provide an appropriate response option for folk error theorists; simply permitting them to judge that both people are mistaken will not work, since this response is consistent with all conflicting cognitivist positions.

However, there may be little reason to include a response option for error theory. Error theory is a sophisticated metaethical position, and it seems implausible that many lay people would endorse it. As such, it may not present much of a methodological problem to exclude it. Sure enough, studies that include an option to endorse error theory find that very few participants do so. For instance, Beebe (2015) included an option for error theory alongside noncognitivism, by permitting participants to select "both beliefs are false," but only a small proportion of participants chose this response. Davis (2021) likewise included a response option for error theory, but also found that, averaging across conditions, only 2.9% of participants were inclined towards this response option, while Pölzler and Wright (2020a) found that, across conditions, 3-15% of participants selected error theory. Even so, we should not ignore small but relevant subpopulations who favor alternative metaethical positions. Even if such people are a minority, this does not in itself entail that their responses are defective. We cannot simply assume a particular response must be the result of incompetence or performance error merely *because* it is an uncommon response. We would need independent data outside the paradigm in question to corroborate such a conclusion. If data collected by means that fall outside the paradigm could confirm similar comprehension rates and general competence with questions about metaethics among these participants, it would be difficult to provide a principled justification for dismissing these responses. On the other hand, if we do wish to regard these participants as incompetent, then failing to provide a response option for error theory will result in dispersing these participants across the

forced choice between other response options even though if a response for error theory were provided we would have excluded these participants, adding further noise to studies that use more restrictive response options.[41]. Thus, *even if* the noncognitivist response option is indicative of incompetence or some other shortcoming that warrants discarding these responses, failing to include would *still* be a mistake.

In addition, the potential for ordinary people to endorse error theory presents another problem for standard versions of the disagreement paradigm. Error theory is a version of *antirealism*. But consider the response options available to participants: that *both are correct*, or *at least one must be incorrect*. Since the error theorist believes that *both* moral claims are incorrect, the judgment that *at least one is incorrect* is consistent with their views, and may be the response option they would favor in the absence of an option to express that both views are false. But notice that *this response option is interpreted as realism*. In other words, any error theorist responding to standard versions of the disagreement paradigm ought, if they understand what they are being asked, choose a response option that is interpreted as *the opposite position*. As such, standard versions of the disagreement paradigm could be systematically miscategorizing an entire subset of antirealists as realists.

First-person versions of the disagreement paradigm also face another problem. Imagine you are asked whether you believe that "murder is wrong," and you judge that it is wrong. Then, you are told about a person who disagrees with you, and are asked whether:

> *(i) Both of you are correct*
> *(ii) Both of you are incorrect*
> *(ii) At least one of you is incorrect*

Even if (ii) were a reasonable option for error theory in principle, there is something odd about asking a person to judge that *their own judgment is incorrect*. First, error theorists have the resources to treat their

---

[41] This assumes that the inclusion of this response option doesn't confuse participants, who are committed to a position other than error theory, but mistakenly select this response option when it is given.

own moral claims differently from what they take others to mean (Joyce, 2011a; 2019). If so, they might not regard their own moral judgments as false in the sense that other moral judgments are false. For instance, they might use moral language metaphorically because it facilitates their practical goals, even if they do not regard their own moral judgments as literally true. If so, it would not in fact be the case that their own moral claims were false. It is worth emphasizing, once again, that this would be an incredibly sophisticated stance for a layperson to take, and it is plausible that few, if any, participants would adhere to such a view. Yet we can still extract a more general problem with presenting people with response options like these: these studies require participants to consistently interpret terms like *correct*, *incorrect*, *mistaken* etc. in ways consistent with researcher intent, and consistent with the interpretation necessary for their response option to reflect their metaethical stance, rather than something else. If a person did not believe there were literal moral facts in the realist sense, they might nevertheless employ moral language and say things like "that is morally wrong." After all, that is conventionally how people tend to speak (at least in English). Speaking in a radically divergent manner might prove difficult, and might signal undesirable social characteristics. If we are already willing to suppose that people have implicit commitments or stances about metaethical issues, it is not *that* much of a stretch to suppose that people are unconsciously sensitive to these considerations, and that they might speak or think in ways that really would evince a commitment to, e.g., error theory, but would nevertheless struggle with a question like this. More importantly, people may not interpret "correct" here in a full-fledged truth-correspondence way, but in some other way, e.g., whether it was socially acceptable for them to hold or express the view in question, or whether they were justified in expressing this view. If so, a person may have some inchoate sense that moral claims cannot be *true* in the sense relevant to error theory; that is, they might think moral claims on some truth-correspondence account are uniformly false. Yet they might bristle at the suggestion that their previous judgment is *incorrect*, since this could bundle truth-correspondence with other implications of being incorrect: that

the participant holds an unreasonable view, that they are worthy of criticism, that they are confused or made some mistake in expressing a first-order moral stance (that, it is worth noting, they are *required* to provide), etc. They may simply pivot in their interpretation of *correct/incorrect* when asked, leading their response to reflect an equivocation that avoids judging themselves to be mistaken *even though* if probed under more careful circumstances they would acknowledge they would find error theory highly appealing.

Although my own response to these surveys counts for little, I am an antirealist myself, and the way I'd walk through this problem illustrates the problem error theorists (and antirealists such as myself) face. Suppose I am first asked how strongly I agree that murder is wrong. If I interpret the question as one about whether there is an *objective fact* about the moral status of a claim, I may choose "1 = strongly disagree," to reflect my denial that there are objective moral facts. However, selecting *strongly disagree* may reflect the view that I believe murder is permissible. This is a common misunderstanding antirealists face. We do not believe that murder (or anything, for that matter) is wrong. Sometimes, laypeople (and the uncharitable or confused philosopher) will imagine that, since we do not think murder is *impermissible*, that we therefore think it is *permissible*. After all, one might think, if something isn't impermissible, then it must be permissible! But this is a mistake. We deny that there can be facts about whether an action is permissible *or* impermissible. So while we must acknowledge that murder is *not impermissible* this is *not* the logical equivalent of saying that it *is permissible*. Given this distinction, our nascent lay error theorists may be disinclined to select "1 = strongly disagree," since this could reflect the view that we think murder *is* permissible. As such, the antirealist might favor "4 = neither agree nor disagree," since this might be the best option to reflect our view that there is no fact of the matter about whether murder is wrong.

Yet suppose we do choose this option, for just this reason. Now, we are told someone disagrees with us, and we are asked whether we are both correct, both incorrect, or whether neither

of us is correct or incorrect. *Now* what response option are we supposed to choose? Well, *my* response option reflects my metaethical stance (as it's supposed to!), *not* my normative stance. And since my normative moral stance is downstream of my metaethical stance, *I am not incorrect*. Thus, I ought (as an antirealist, and the same applies to error theorists) to select the response that at least one of us is mistaken: *the other person*. Yet this response option is interpreted as *realism*!

Finally, suppose I instead interpret the question to be asking me about my evaluative stance towards murder; that is, do I approve or disapprove of murder. This is, personally, the most attractive response to me, as an antirealist. I do not typically go around in everyday speech responding to real-world events by pointing out that nothing is morally wrong. When people bring up the latest terrorist attack or a politician caught accepting bribes, I do not respond by saying "Technically, terrorism and taking bribes aren't morally wrong because nothing is morally wrong." This is not likely to win many friends, and I don't know of any antirealists (and I know quite a few, we tend to track one another down, mostly so we can gripe about realists) who speak this way. On the contrary, we tend to use our moral speech to reflect our evaluative attitudes. We say that terrorism and accepting bribes is *bad* and that we think these people *should be punished* because we *do* think these things are bad and that these people should be punished. We just don't think there is some stance-independent fact that they ought to do so. This raises an extremely serious methodological problem for the disagreement paradigm. Questions about the rightness or wrongness of first-order moral claims are intended to be interpreted in a purely truth-correspondence way. In a purely academic setting, my interlocutors and I are careful about how we speak, have training in metaethics, and are familiar with the relevant terms and concepts used in metaethics and philosophy more broadly. There is little risk of suffering serious reputational consequences for denying that atrocities are morally wrong. And our emotions are dampened by the cold, intellectual formality of the circumstances. *Even then* the antirealist's denial that atrocities are morally wrong is sometimes met with incredulity and muted outrage. Why? When we make first-order

moral claims, we rarely intend to merely express our metaethical stances, yet in denying that a given atrocity is *morally wrong* this is all the antirealist wishes to do.

This is likely to occur only in the context of philosophical dialogue. In almost all ordinary circumstances, whether we are inclined to say that something is morally right or wrong is inextricably bound up in our evaluative attitudes towards the act in question. As such, to say that *murder is not wrong* in most circumstances does not merely convey that we think it is technically the case that there is no stance-independent fact about the moral status of murder, but that *we don't disapprove of murder*. That is, to claim that *murder is not wrong* may not be interpreted as an abstract intellectual position, but will instead entail a whole host of implications for how we are likely to behave in everyday social circumstances. If we do not disapprove of murder, this may suggest we lack empathy, and that we are callous, cruel, manipulative, even violent. In short, it conveys a whole host of undesirable *antisocial* personality traits. If this is not obvious, think about how you would react if you overheard someone at a social gathering (*not* a gathering of philosophers) casually remark that "there is nothing morally wrong with torturing children for fun." I'm an antirealist. Technically, I *must* concede that I think this, too. But I suspect my reaction would be the same as yours: I'd want to get as far away from this person as possible. And I am certainly not going to ask them to babysit! And yet, as an antirealist, I *do* disapprove of murder. I have very similar reactive attitudes towards what most of us regard as morally praiseworthy and blameworthy actions. As such, I may wish to choose the response option that most closely reflects my *disapproval* of murder, rather than my judgment that it is *incorrect*. This causes a schism between my normative stance (*murder is not wrong*) and my evaluative attitude (*murder is wrong*), where 'wrong' means something different. Now, after expressing my disapproval of murder, I am expected (as an antirealist, but not a noncognitivist) to interpret the disagreement paradigm as a question about my first-order moral stance, *even though this is not what my response to these questions reflected*. In other words, questions of agreement and disagreement *presume cognitivism*, and the noncognitivist is

expected to be responsive to this, and to judge that that *their own moral stance* is neither correct nor incorrect. This is fine, as far as it goes, but it does mean that some antirealists who are not noncognitivists will be disposed to choose the noncognitivist option. While this would accurately categorize them as an antirealist, it would mischaracterize them as the wrong kind of antirealist.

This is, perhaps, a relatively minor problem. Exploring the various ways I might respond to the disagreement paradigm reveals a far more serious problem: people with *the same metaethical position* could plausibly respond to the disagreement paradigm *in almost any way* due to variations in how they interpret it. At the same time, these considerations reveal how *complicated* these questions are. How we respond to statements about murder and abortion do not *merely* reflect our moral stances and commitments; they may also reflect our *evaluative attitudes* towards these actions, or at least will be interpreted in this way by ordinary people. When asked whether we agree or disagree that murder is wrong, which is the correct interpretation? Is there a correct interpretation? Presumably it must be the former, since telling me that someone disagrees with how I feel is incoherent[42], but if so, then the only responses to the disagreement paradigm that accurately reflect the participant's metaethical stances are those that reflect this interpretation. Is this how people interpret the disagreement paradigm? I have no idea.

Worse still, even philosophers struggle to disentangle the distinction between normative positions and evaluative attitudes, even when the distinction is salient and they are making an active effort to do so. What hope do laypeople have to succeed at drawing the distinction? They have no knowledge of the distinction, no experience drawing it, and are completely oblivious to the expectation that they do so. All throughout their lives, moral claims are understood not merely to express one's normative position, but one's evaluative attitude towards an act. For a person to say that murder is

---

[42] Unless *disagree* is not understood in the truth-correspondence sense that I am literally *mistaken*, but if that is the case, then the disagreement paradigm would be invalid anyway.

not wrong does not merely convey that they don't think murder is morally impermissible in some abstract, philosophical respect, but that they, personally, don't disapprove of murder. It is perfectly natural to respond to this with incredulity, outrage, and disgust, and expressing that this person is *incorrect* may be the only adequate way to express one's moral disapproval, *even if* you would not endorse moral realism on reflection. It also has the added bonus of being the only appropriate way to signal your own disapproval. Even I would be inclined to respond this way. Or I might respond that we are both incorrect. Or I might respond that none of us could be correct or incorrect. Given my views, I could make a reasonable case for *any* of these responses. So too, I suspect, could any participant. A host of convoluted considerations lurk beneath the superficial simplicity of the disagreement paradigm. For any given participant, it is unclear which (if any) will be salient and how they will influence that person's response. It is unclear whether there are any patterns in how these hidden complexities would influence people's reactions. It is even possible that they systematically bias people who hold a particular metaethical stance or commitment to select a response option that is interpreted in a way that doesn't reflect their metaethical views.

### S2.3.1.3 Missing distinct options for varieties of relativism

Another problem with standard versions of the disagreement paradigm is that the relativist option is underspecified and excludes an appropriate response option for some forms of relativism. For relativists, the truth of moral claims depends on the mental states of one or the standards of different groups. Yet as Pölzler (2018b) notes, the relativist response option in standard versions of the disagreement paradigm presents only a generic conception of relativism, which leaves underspecified *whose* mental states the truth of moral claims depends on and which mental states matter. Individual subjectivism holds that a moral claim can only be judged true or false relative to the moral standards of *individuals*, while cultural relativism holds that moral claims can only be judged right or wrong relative to the standards of different *cultures*. Although rarely discussed, it is also possible for the truth

of moral claims to be relativized to other standards, such as species (see Bush, 2016). Unfortunately, standard versions of the disagreement paradigm cannot distinguish which form of relativism people endorse. Thus, even if it can distinguish realism from antirealism, its ability to distinguish different kinds of antirealism is limited. This does not demonstrate that the disagreement paradigm is *invalid*, but it does show that even at its best, it could provide only very coarse-grained data about folk metaethics.

The distinction between subjectivism and cultural relativism is further complicated by the distinction between *agent* and *appraiser* relativism (Quintelier, De Smet, & Fessler, 2014). To illustrate the difference between these views, consider the following scenario:

> Alex believes that slavery is wrong, and slavery is considered immoral in Alex's culture. One day, Alex hears about a culture where people believe it is morally acceptable to own slaves. One member of that culture, Sam, owns many slaves, and believes "It is not wrong for me to own slaves.". Alex claims that "It is wrong for Sam to own slaves."

According to agent relativism, the truth status of moral claims must be judged relative to the standards of the person engaging in the action (*agent individual subjectivism*) or the standards of their culture (*cultural agent relativism*). Since owning slaves is consistent with Sam's standards (and the standards of Sam's culture) it is not wrong for Sam to own slaves. However, according to appraiser relativism, the truth status of moral claims depends on the moral standards of the person expressing the moral judgment (*appraiser individual subjectivism*) or the standards of their culture (*cultural appraiser relativism*). According to appraiser relativism, Alex would be correct to judge that slavery is wrong, because slavery is not consistent with Alex's moral standards (or the standards of Alex's culture).

Given this distinction, there are at least four major forms of relativism people could endorse. In fact, there are more, since one could endorse both agent and appraiser relativism. In principle, one could endorse any combination of these four possibilities, however implausible some combinations may be. Yet conventional versions of the disagreement paradigm are incapable of distinguishing which

of these people endorse by selecting the response that *both could be correct*. This does not, by itself, invalidate standard versions of the disagreement paradigm. However, it once again illustrates that standard versions of the disagreement paradigm cannot provide detailed information about people's metaethical beliefs. More importantly, standard versions of the disagreement paradigm are not designed to assess appraiser relativism. As Quintelier et al. (2014) point out, "Existing studies about folk moral relativism most often vary only the appraisers" (p. 214). In other words, participants are only asked whether two people who disagree about whether a given action is morally right or wrong can both be correct. The moral standards of the agents that perform these actions are not specified, and participants are not presented with cases that involve the same action but agents with standards that are consistent or inconsistent with the relevant action in order to evaluate whether the moral status of the question varies based on the *agent's* standards rather than those appraising the action. As such, these studies don't merely fail to disentangle agent and appraiser relativism, but frame questions in a way that only prompt a choice between realism and appraiser relativism *in particular*. Even if they did mention the moral standards of an agent performing a given action, the standard response options given to participants would still fail to disambiguate agent and appraise relativism, since there would still only be one relativist response option available to participants. Thus, just as standard versions of the disagreement paradigm do not provide a response option for noncognitivism and error theory, it also does not provide a distinct response option for agent relativism.

Once again, this concern would be moot if we had good reasons to believe that ordinary people would not endorse agent relativism. Yet there is some indication that people consider both the moral standards of the agent and of the appraiser to be relevant to assessing whether an action is morally right or wrong. Quintelier, De Smet, and Fessler (2014) conducted the only study that attempts

to tease apart the distinction between agent and appraiser relativism.[43] They found that participants consider *both to* be relevant to judging whether an action is right or wrong. Although there are significant shortcomings with this study[44], there is little reason to presume without evidence that agent relativism is not a feature of folk metaethics.

Finally, some response-dependent theories hold that the truth of moral claims depend on how people would respond to a given action under particular circumstances. For instance, they might hold that moral facts depend on particular emotional responses, such as disapproval or outrage. Since different people may exhibit different emotional responses to the same action, the truth of a moral claim could vary relative to different patterns of response.

Some versions of response-dependence allow for truth status to vary depending on response, but others do not (ideal observer). Thus, standard versions of the disagreement paradigm cannot distinguish response dependence theory, since people who endorse different forms of it would respond differently. Once again, we might wish to rule these positions out from the armchair. But it is worth asking why researchers studying folk metaethics would wish to do so. One of the primary rationales for conducting this research is the inadequacy of relying on armchair assumptions. Researchers presumed folk realism for the most part, and present findings challenge that assumption. We should not be so confident that we know what people do think, and a half-measure, where we are willing to test for the possibility of some metaethical stances but not others seems like an indefensible and unprincipled half-measure that recapitulates the very overconfidence that empirical research is

---

[43] In accounting for the distinction, they echo my concerns about the validity of standard versions of the disagreement paradigm, explicitly stating the need to avoid study designs that are insensitive to it since "failure to do so may lead to an underestimation of the prevalence of folk moral relativism, as respondents may employ relativist intuitions of a kind other than that being measured" (Quintelier, De Smet, & Fessler, 2014, pp. 209-210)

[44] These studies are replete with a variety of ways participants may have interpreted what they were asked in unintended ways. In addition, how could we distinguish realist reasons for thinking one's standards are relevant from agent relativist reasons? We'd need yet further studies to disentangle these possibilities, while having some means of confirming intended interpretation and not corrupting participants with instructions to such an extent that we're no longer probing the intuitions of ordinary people, but people with minimal (and potentially biased) tutoring in philosophy.

intended to correct in the first place. Why not simply explore all possibilities, if for no other reason than to rule them out empirically?

### S2.3.1.4 Missing distinct options for other positions

All of these omissions gloss over the absence of yet more metaethical positions and distinctions, including the distinction between naturalism and non-naturalism, constructivism, relation-designating accounts, incoherentism, indeterminacy, quietism, hybrid accounts, quasi-realism, and perhaps more I haven't heard about. With the exception of Davis (2021), who provides naturalist and non-naturalist response options, none of these options are represented in the stimuli, despite being legitimate positions. Merely because a particular position is unpopular doesn't necessarily make it less plausible or unworthy of consideration. After all, I personally endorse a view that falls square within this swashbuckling band of less orthodox positions. Indeed, there's an immense irony that studies purportedly providing evidence that ordinary people are metaethical pluralists who sometimes use moral claims in realist ways and sometimes use them in antirealist ways, yet researchers present participants with a forced choice between realist or antirealist options, rather than giving them options to express pluralism directly. If most people are pluralists, why ask them questions that *require* them to treat each moral claim as though one can only be a realist or antirealist towards it?

In short, numerous metaethical positions cannot be adequately captured by standard versions of the disagreement paradigm. By limiting participants to only two (or at best three) response options, researchers are forcing so much potential variation and complexity through a narrow and oversimplified sieve that *at best* the resulting pattern of responses loses much of the detail and richness that may be present. Matters are far worse than this, however. We cannot rule out, *a priori*, the possibility that people hold views that are not accurately captured by the response options given in standard versions of the disagreement paradigm. As such, we have no way of knowing what proportion of people may hold these views. And since these people are forced to choose among

available response options, this can give the misleading impression that people hold the views standard versions of the disagreement paradigm is designed to measure. Since we don't know what proportion are forced to choose unrepresentative responses, and people who hold various metaethical positions may find various features of *any* of the response options attractive, we cannot use the results of standard versions of the disagreement paradigm to draw confident inferences about the proportion of realists and antirealists. Any particular unrepresented metaethical stance or commitment might introduce a tolerable degree of noise on its own, but given how many possibilities there are, their cumulative impact may be enough to tip standard versions of the disagreement paradigm towards being uninformative, or even actively misleading.

**2.3.2 Poor specificity**

The issue of poor specificity is fully addressed in the main text. While the poor specificity of folk metaethics studies doesn't invalidate them, it's worth noting that I have not even addressed a related cluster of more serious issues with the way folk metaethics studies are designed. One of these issues is that metaethical positions in the academic literature incorporate a variety of distinct theses: semantic theses, which concern what people mean when they make claims, metaphysical theses, which concern the nature of reality, epistemic these, and so on (see Sinnott-Armstrong, 2009). One issue with measuring folk metaethical positions is that existing research on folk metaethics ignores these distinctions. Take error theory, for instance. Error theory require at least two theses: (a) a *semantic* thesis about what people mean when they make moral claims (b) a *metaphysical* thesis: that people are in systematic error because their claims commit them to false presuppositions, e.g., mistaken claims about moral properties. The latter claim *cannot* be a feature of *how people speak*. As such, are researchers studying folk metaethics attempting to identify how ordinary people speak or think when making first-order moral claims, or are they trying to figure out what ordinary people's second-order stance *towards* these first-order moral claims is?

In other words, it's one thing to determine that a participant thinks that when people make moral claims, that those people are attempting to refer to stance-independent moral facts, and another thing to determine whether such utterances successfully refer to such facts. The former is a semantic claim, and the latter is a metaphysical claim. *Which of these is any particular metaethical paradigm intending to measure? Both?* Why bundle them together in this way? It's not even clear, much of the time, if researchers are attempting to evaluate (a) what ordinary people mean when they make moral claims (b) what ordinary people think they or others mean when making moral claims (i.e., their *metalinguistic* position), or (c) people's *metaphysical* stance towards the nature of morality. These are not identical, and yet not only is there no effort to disentangle them, their entanglement isn't even acknowledged or discussed. I suspect that researchers studying folk metaethics are so caught up in the way contemporary analytic philosophy engages with metaphysics—by embedding it so thoroughly in assumptions about how language works and the relation between language and metaphysics—that they are either unaware of the degree to which distinct philosophical presuppositions are thoroughly embedded in the very way questions are framed, asked, and interpreted.

This is regrettable, and frustrating to me personally, since I reject orthodox notions about the relation between language and metaphysics, or at least my critical appraisal of what I take to be the unquestioned dogmas hidden in contemporary analytic practice, including those contemporary philosophers would disavow or insist is a misrepresentation of their practice. In short: I am pointing to the fact that the very way questions in folk metaethics are structured buy into a very particular, idiosyncratic, and recent way of thinking about morality, and of philosophy more generally. If I thought this practice were untroubled, or at least approximated an appropriate method, I would take little issue with baking these assumptions into folk philosophical research. But I don't. I think contemporary analytic philosophy has deep and serious flaws that a small but persistent contingent of philosophers have continued to criticize (e.g., Baz, 2012; 2015; 2017; Horwich, 2015). As such, I am

in the unfortunate position of watching folk philosophical research largely recapitulate many of the errors and misconceptions of contemporary analytic philosophy in the very process of rejecting or critiquing its methods. Why do experimental philosophers think the appropriate way to critique traditional philosophical approaches requires borrowing their entire suite of terms and concepts, and assuming that the way ordinary people think is largely isomorphic?

This is a bizarre assumption to make when one of the central motivations behind the experimental philosophy movement was a deep skepticism about whether ordinary people think the way philosophers do. Yet rather than think ordinary people think in ways that fundamentally differ from philosophers, experimental philosophers have simply assumed ordinary people think largely in terms of the same categories and distinctions, but simply differ in their proportion of allegiance to one position over another. For comparison, imagine members of the Bigfoot Appreciation Society were in a bitter dispute over whether Bigfoot tends to be left-handed or right-handed. 80% of the membership is convinced Bigfoot is left-handed, but an insistent minority maintain that Bigfoot is right-handed. Now suppose you became skeptical of the Bigfoot Appreciation Society. You want to resolve this dispute. What should you do? Here's one option: construct a survey for nonmembers, and present them with the following question:

> *Do you think Bigfoot, the most noble and majestic of creatures, is left-handed or right-handed?*
>   o  *Left-handed*
>   o  *Right-handed*

Suppose you found that about 70% of respondents in a small Appalachian town agreed Bigfoot is right-handed. Would this settle the matter?

No. This would be a ridiculous and embarrassing waste of time. Most people wouldn't have a position on this matter. Maybe most would pick one response over another if asked, but the problem with the Bigfoot Appreciation Society isn't that they have mistaken views about Bigfoot's handedness,

but because they mistakenly think there's some fact of the matter in the first place. If you became skeptical of the Bigfoot Appreciate Society's methods, and not merely their conclusions, why in the very act of criticizing those methods would you lend credibility to the legitimacy of the dispute in the first place? Why not ask people whether they think Bigfoot exists in the first place?

This is a frivolous analogy to experimental philosophy, but the point stands. Rather than question the legitimacy of the disputes and distinctions central to analytic philosophy at a more fundamental level, experimental philosophers have largely bought into the way analytic philosophers frame philosophical issues, presumed ordinary people would, too, and constructed their studies accordingly.

## S2.4 Conflations with unintended concepts

*Formal conflations* occur when researchers construct stimuli that conflate realism/antirealism with other distinctions. In such cases, participant responses may reflect the intended interpretation, but researchers mistakenly take such interpretations to reflect a metaethical distinction that they do not in fact reflect. In other words, the mistake results from a failure of operationalization on the part of the researcher, and not a result of unintended interpretations by participants to otherwise well-constructed stimuli. Fortunately, researchers rarely misoperationalize measures in metaethics. However, one error did occur in Goodwin and Darley's (2008) seminal paper on folk metaethics. Participants were asked to rate their level of agreement with a series of moral statements (e.g., "Consciously discriminating against someone on the basis of race is morally wrong"), then asked them whether, for each of these statements was a:

(1) True statement
(2) False statement
(3) An opinion or attitude

G&D interpreted (1) and (2) as *realism* and (3) as *antirealism*.[45] There are several methodological problems with these response options. *Opinion* can be interpreted in multiple ways, some of which are not consistent with the interpretation G&D intended: G&D either require participants to interpret *opinion* in nonpropositional terms, or else (3) is not mutually exclusive with (1) or (2). *Attitude* is in little better shape, since it too can be understood in both propositional and nonpropositional terms. In addition, response (3) collapses two distinct response options into one, leaving participants with no way to specify whether they regard the statement exclusively as either an opinion or an attitude (but not both).

Yet as Pölzler (2017, pp. 461-463; 2018b) points out, the primary problem with these response options is that they reflect the distinction between cognitivism and noncognitivism, *not* the distinction between realism and antirealism. Options (1) and (2) only indicate that moral claims are truth-apt, not that they are true in a stance-independent way. Both are compatible with antirealist positions, such as cultural relativism, individual subjectivism, and, in the case of (2), error theory. In fact, not only are they compatible, they are the most appropriate responses for people who hold these views. As a result, the response options G&D use cannot consistently distinguish realists from antirealists, since many antirealists ought to find regard (1) and (2) as the best reflection of their views.[46],[47] Unfortunately, this

---

[45] They used the terms objectivism and non-objectivism, but the explanation they gave of the distinction they were measuring is equivalent to the realism/antirealism distinction as it is used here (REFER).

[46] They also do not provide an exhaustive list of noncognitivist options. Some noncognitivist philosophers maintain that moral claims are prescriptions, yet such a response option is not available to participants. It may be reasonable to presume such views are uncommon among ordinary people.

[47] One might hope for some consolation in the possibility of using this question to distinguish cognitivists from noncognitivists, but these response options may not be able to achieve that goal, either. This is because cognitivists need not deny that moral statements express attitudes; they merely deny that they *only* express attitudes. If there are cognitivists in these samples, most may favor (1) or (2), since they may see the assertoric role of moral claims as central or primary. Yet it might strike them as odd to select these options since the implication is that such claims do not also express opinions or attitudes; the cognitivist might very well think that (3) is also correct. In other words, participants may think that at least two response options are both correct. Simply put, the response options G&D provide are not mutually exclusive. It is unclear how participants should be expected to respond to a forced choice between compatible views. Imagine presenting participants with a work of art, and asking them to select one of the following responses:
(1) The artwork displays technical skill
(2) The artwork is beautiful

invalid measure of realism/antirealism was used in several studies (e.g., Goodwin & Darley, 2008; Wright, Grandjean, & McWhite, 2013), and participant scores were combined with responses to the disagreement paradigm to form a composite measure of realism/antirealism.

Fortunately, this particular paradigm has been mostly abandoned in favor of others, so concerns about its validity are largely moot. Most other methods used to measure folk metaethical belief are instead subject to informal conflations. *Informal conflations* occur when researchers use stimuli that, if interpreted as intended, would reflect the relevant metaethical distinctions, yet, due to ambiguity or inadequate specification, significant numbers of participants interpret stimuli in ways that differ from researcher intent. When this occurs, participant responses no longer reflect the distinction of interest. Such conflations are not the result of an explicit and demonstrable failure to operationalize the variables of interest in line with the proper metaethical distinction, but are instead an incidental byproduct of ambiguity that may arise even when researchers have the correct understanding of the distinction and have implemented measures that, if interpreted in line with researcher intent, would yield valid responses (assuming other conditions for validity are met).

There is no sharp dividing line between formal and informal conflations. A study may present well-operationalized stimuli, but include instructions or other details that encourage unintended interpretations. Such stimuli would plausibly fall somewhere between having formal and informal conflations. Yet the distinction is helpful in that it highlights the source of the conflation, and draws attention to the difficulties that accompany efforts to prompt participants to respond to subtle, unfamiliar, sophisticated, and often non-obvious considerations. Sometimes the fault lies with

---

This is a strange question to ask, because it implies that if it displays technical skill it isn't beautiful, and vice versa. Likewise, participants who favor (1) or (2) may recognize that in selecting these responses, they are denying that moral claims reflect opinions or attitudes. Yet for cognitivists, this would be an odd thing to deny. This is not to say that cognitivists might not favor (1) and (2) regardless. If so, this item could be an effective measure for cognitivism or noncognitivism. Even so, the response options on offer here are less than ideal, and, at any rate, it is not a valid measure of realism or antirealism. More importantly, as we will see, however, there are many other reasons to doubt participants interpret questions like these as intended, including other ways they might conflate metaethical questions with other, unintended distinctions, e.g., descriptive claims.

researchers who erred in some correctable way. But some conflations are far more difficult to avoid, and result instead from inherent difficulties in specifying what one is asking without including significant context or clarification (Bush & Moss, 2020). I will focus primarily on these latter, informal conflations.

### S2.4.1 Conflating metaethics with normative ethics

Since the disagreement paradigm is intended to distinguish realism from antirealism, it is exclusively concerned with a *metaethical* distinction. Metaethical distinctions concern second-order questions about the nature of morality. Metaethics is distinct from *normative ethics*, which is concerned with first-order questions about what is in fact morally right or wrong, permissible or impermissible, etc. When a person judges that e.g., "murder is wrong," this is a *first-order* (normative) position. When they judge that there is a stance-independent fact about whether murder is wrong, this is a *second-order* (metaethical) position. In other words, metaethical positions are "philosophical views *about* such first-order moral judgment" (p. Pölzler, 2018b, p. 657; emphasis original; see also Huemer, 2005, pp. 1-2).

Most ordinary people are unfamiliar with metaethical considerations. They are far more familiar with expressing a first-order normative judgment about what is right or wrong. In other words, ordinary people are far more experienced with judging an action to be moral or immoral than with determining what it *means* for an act to be morally right or wrong. When confronted with a disagreement, it would be natural for participants to be primarily concerned with evaluating the normative stances of the people who disagree, not a second-order consideration about what it would mean for one or the other of the people who disagree to be "correct." This preoccupation with normative rather than metaethical considerations may render participants vulnerable to the potential influence for normative considerations to influence how they respond to the disagreement paradigm, even when such considerations should play no role in how they respond. And since participants must choose from a limited set of options, all of which are interpreted as evidence of their metaethical

beliefs, researchers would be interpreting their responses in purely metaethical terms even when this does not reflect how participants interpreted the questions.

The risk that participants will conflate normative and metaethical considerations is sometimes exacerbated by the wording used by particular versions of the disagreement paradigm. After presenting participants with a disagreement between two hypothetical members of different cultures about whether, "It's okay to hit people just because you feel like it," Nichols (2004) asked: "*Which of the following do you think best characterizes your views?*" and gave participants one of three options:

(1) It is okay to hit people just because you feel like it, so John is right and Fred is wrong

(2) It is not okay to hit people just because you feel like it, so Fred is right and John is wrong

(3) There is no fact of the matter about unqualified claims like "It's okay to hit people just because you feel like it." Different cultures believe different things, and it is not absolutely true or false that it's okay to hit people just because you feel like it. (pp. 9-10)

Although (3) may represent a metaethical stance, it is not clear whether (1) and (2) are most plausibly interpreted as metaethical positions (Beebe, 2015). Suppose you are a moral subjectivist, and do not believe that there are stance-independent moral facts. If you hold such a view, you are an antirealist. Yet it is consistent with such a view for you to judge that your moral stance is correct, and someone who holds a contrary moral stance is incorrect *relative to your moral standards*. (1) and (2) do not cleanly separate stance-independent from stance-dependent stances, and thus should not be used as a method for distinguishing these views from one another. At best, (1) and (2) could reflect a cognitivist stance while (3) could reflect a noncognitivist stance, but this distinction would not represent a measure of realism versus antirealism. Yet it seems at least as plausible that participants presented with this question are simply choosing the response that most closely reflects their first-order moral judgments. Even if this were treated as a measure of cognitivism versus noncognitivism, (3) would not be an ideal choice. The most straightforward interpretation of (3) is not that there are no stance-independent moral facts, but that there may be no uniform answer to whether ""It's okay to hit people just because

you feel like it," absent further specification about the context in which such an event takes place. Unfortunately, (1)-(3) are all subject to multiple, reasonable interpretations that would result in participant responses not consistently representing distinct metaethical stances or commitments based on their response option. As a result, these response options are unlikely to provide a valid measure of metaethical belief.[48]

Even if researchers are careful to avoid prompting first-order moral judgments, such judgments may be so familiar and easy to process that it may still influence participant response in undesirable ways. For instance, participants may interpret agreement with the possibility of both positions being correct to imply that the participant is willing to tolerate someone who holds or even acts on either belief. Contrary to popular misunderstanding, relativism does *not* require us to tolerate moral beliefs that conflict with our own moral standards (Bush, 2016; Collier-Spruel et al., 2019, Rai & Holyoak, 2013). A cultural relativist may believe that another culture's practice of slavery is morally permissible *according to that culture's standards*, and believe that this is what a person from that culture means when they say, "slavery is morally permissible." But this in no way requires the relativist to tolerate the practice of slavery. Tolerance is itself a substantive normative stance that may or may not correctly reflect the relativist's subjective moral standards, or the standards of their culture. Yet the layperson may not draw this distinction, or may be sensitive even to the possibility of implying tolerance for abhorrent moral beliefs. If so, they may judge that only one person could be correct so as not to imply tolerance (a first-order normative stance) for contrary moral beliefs, even if they reject objectivism.

The conflation between normative and metaethical considerations may be simpler and harder to root out than this. According to Pölzler (2018b) "Avoiding first-order moral intuitions in studies

---

[48] If they do, it would be accidental, and their validity could only be confirmed by establishing that results are consistent with some other, well-validated measure.

on folk moral realism altogether may be methodologically infeasible" (p. 658). Pölzler suggests that even if we could successfully prompt metaethical interpretations, normative moral judgments may be the output of automatic psychological processes that could influence participant response even if the participant is deliberately and consciously attempting to answer the question in line with researcher intent. In my own experience, people who challenge antirealists frequently seem to conflate normative and metaethical considerations. When the antirealist claims that there are no stance-independent moral facts, they are often met with the misguided *"what about Hitler?"* challenge. This challenge involves the antirealist's rival declaring that the antirealist has no grounds on which to disapprove of Hitler's actions, or to state that what Hitler did was bad. Yet this is not true. The antirealist claims only that Hitler's actions are not *stance-independently* bad, not that they aren't "bad." The antirealist's conception of what it would mean for Hitler's actions to be bad may not be satisfactory to the realist, but it is simply not true that the antirealist cannot say that Hitler's actions are "bad" in a way that comports with how are plausibly inclined to think in nonmoral domains. To illustrate why, consider a similar exchange, not in the domain of morality, but in the domain of food preferences.

The *gastronomic antirealist* denies that there are stance-independent facts about which food is good or bad (Loeb, 2008). Instead, they maintain that to express that a particular dish is "good" or "bad" is to express a noncognitivist attitude of approval or disapproval, respectively, or to articulate one's subjective standards, e.g., to say "Chocolate ice cream is better than vanilla ice cream" is best understood as expressing the claim that "*I find* chocolate ice cream to be better than vanilla ice cream." Most of us are probably gastronomic antirealists (or at least not gastronomic realists). Yet we are perfectly happy stating that we find certain foods good and other foods bad. Now imagine the gastronomic equivalent of the *"What about Hitler?"* challenge. Perhaps we might call it the *"What about bacon?"* challenge. Suppose you said that, with respect to its gastronomic properties (rather than e.g., its health status or the ethics of acquiring it), bacon was *bad*. You do not like the taste of bacon, and

prefer not to eat it. Outraged, the gastronomic realist says that you can say no such thing. Since you deny that there are stance-independent facts about whether bacon is good or bad, there is no meaningful respect in which you could say anything about its gastronomic quality. This would be a decidedly weird response. The realist may not be satisfied with your use of "bad" merely reflecting your attitudes or subjective preferences, but such conceptions of what it would mean for food to good or bad seems well within the reach of how we are ordinary disposed to use such terms, and there is no reason why you would be misusing language or conceptually confused if you said that bacon was bad.

Such challenges are thus misguided regardless of the domain in question. Yet I draw attention to these challenges not simply to explain why they do not work, but to draw attention to *why* people present them, despite their having little philosophical credibility. Namely, these objections could plausibly result from those who present them failing to disentangle normative from metaethical considerations. When the realist challenges the antirealist by insisting the antirealist cannot say Hitler is bad, they equivocate between badness as a stance-independent property and badness as a stance-dependent property, suggesting that the antirealist has access to neither, rather than merely denying the former. That is, the challenger seems to be suggesting that the antirealist can have no first-order, normative stance about whether Hitler is bad, rather than merely failing to endorse (what the realist considers) the correct second-order stance *about that* normative stance. In short: this challenge may very well be the product of conflating first-order and second-order moral considerations. It is difficult to assess how often or in what contexts such conflations occur, but I suspect many readers will be acquainted with such exchanges.

Setting aside anecdotes, are there any more concrete reasons to believe that normative considerations influence how participants respond to the disagreement paradigm? Pölzler draws attention to one suggestive line of evidence: the possibility that normative judgments influencing how

participants respond to metaethical prompts could explain the reliable association researchers find between the strength of agreement participants express towards moral claims. Recall that participants presented with the disagreement paradigm are first asked to rate how strongly they agree or disagree with each claim. Many studies that have assessed the relationship between how strongly participants agree or disagree with each moral claim find that strength of agreement is correlated with the likelihood of a "realist" responses (e.g., Beebe et al., 2015; Beebe & Sackris, 2016; Goodwin & Darley, 2008; Wright et al., 2013).

However, the explanations participants give when asked to explain their answers to the disagreement paradigm provide stronger and more direct evidence. In **Chapter 4**, I analyze open response questions designed to assess how participants interpret the disagreement paradigm. These studies consistently reveal many participants interpreted the disagreement paradigm in normative rather than metaethical terms. I am not the first to collect open response data of this kind. Wainryb et al. asked children to explain their answers to the disagreement paradigm. Wainryb et al. (2004) presented children aged 5, 7, and 9 with disagreements between two people in each of four categories: morality, taste, facts, and ambiguous facts.[49] Children were first asked what they believe about each issue before judging a disagreement between two other people.[50] Wording for each disagreement was standardized:

> [Person 1] believes that [Belief A], and [Person 2] believes that [Belief B]. Do you think that only one belief is right, or do you think that both beliefs are right?

For example, a moral disagreement would be phrased as follows: "Sarah believes that it's okay to hit and kick other children, and Sophie believes that it's wrong to hit and kick other children," and a taste

---

[49] "Facts" consisted of non-evaluative claims with presumably uncontroversial answers, such as whether a pencil will go up or fall down when let go. "Ambiguous facts" consisted of non-evaluative claims about situations open to multiple plausible interpretations, such as why a dog did not eat its food.

[50] Wainryb and colleagues included disagreements between two children (e.g., "Sarah and Sophie are first graders, just like you," p. 691) or a child and an adult (e.g., "Sarah is a first grader, just like you; Mrs. Davidson is a grown-up," p. 691) to assess the role of authority.

disagreement would be phrased as "Daniel believes that chocolate ice cream is yucky, and David believes that chocolate ice cream tastes yummy" (p. 691). Wainryb and colleagues found that, with respect to moral disagreements, 100% of 5-year olds and 7-year olds and 94% of 9-year olds judged that one person was right, a pattern that closely paralleled the near unanimity for factual disagreements (100% for 5-year olds, 97% for 7-year olds, and 94% for 9-year olds). This seems to provide compelling evidence that the vast majority of children are moral realists.

However, when asked to explain why only one moral belief was right, W&C report that "[t]he majority referred to concerns with others' welfare and with fairness as the grounds for judging that moral beliefs are not relative," offering the example that "Kicking other kids is mean because it hurts them, so what that kid said is just wrong, very wrong" (p. 697). This response is not a justification for moral realism. Rather, it demonstrates that the participant did not understand the question to be about the truth conditions of moral disagreements. A proper metaethical justification would require stating e.g., "they can't both be right because that's not possible" or "there is only one right answer to moral actions." Granted, children may not be able to articulate sophisticated views like these, and there are legitimate reasons to worry that even if children did interpret the disagreement paradigm in metaethical terms that, when asked to explain their answers, their responses may fail to reflect this. However, insofar as their answers seem to suggest a normative rather than metaethical interpretation, the onus is on those defending these findings as evidence that children are realists to demonstrate that their explanations are not accurate reflects of how they interpreted the disagreement paradigm, but their initial response to the multiple choice portion do accurately reflect their views. One is not entitled to merely dismiss open response data that conflicts with one's' preferred interpretation.

The answers children did give don't merely fail to demonstrate proper understanding of the questions, but are far more consistent with alternative interpretation: instead of responding to the questions as intended, children who appealed to welfare or fairness instead simply expressed their own

first-order judgments about which of the two beliefs they agreed with. In effect, what were intended as metaethical questions appear to have been interpreted no differently than direct questions about what is morally right or wrong, e.g. "If one child believes it is okay to kick other kids, and another believes it is not okay to kick other kids, which of them is correct?" A response to this question would not reflect a second-order stance about the nature of moral truth, but a first-order moral stance, yet antirealists can and do have first-order moral stances, so responses to this question could serve as valid measures of metaethical stances or commitments.

The justifications children gave across all age groups appear to fit this pattern of first- order interpretation. Wainryb et al. coded justifications for moral disagreements that appealed to normative considerations like welfare and fairness into a single category, "fairness." 100% of 5- year olds, 100% of 7-year olds, and 94% of 9-year olds appealed were included in this category. Since these responses are best explained as unintended interpretations of the question, they serve, in effect, as a comprehension check that *nearly every participant failed*. As a result, Wainryb et al.'s study provides no evidence that children are moral realists.

This problem alone may be sufficient to invalidate Wainryb et al.'s findings. However, there are two caveats to this objection. First, as already noted, it is possible children's explanations misleadingly suggest normative interpretations when in fact they interpreted the disagreement paradigm in metaethical terms. Second, it is possible that many of the participants W&C coded as offering a "fairness" response in fact offered metaethical explanations. Unfortunately, I was unable to obtain their data, so I could not analyze these responses myself to determine whether this was the case. Without the ability to directly assess participant explanations, we cannot confirm or disconfirm whether some alternative explanation for what appears to be widespread tendency to interpret the disagreement paradigm in normative terms.

## S2.4.2 Conflating metaethics with epistemic concerns

Metaethical realism and antirealism are typically construed *metaphysical* positions on the truth status of moral claims.[51] Such considerations are distinct, but related, to *epistemological* questions about how we can acquire moral *knowledge*, whether (and how) our moral beliefs can be *justified*, and whether we can be *certain* of our moral views (Pölzler, 2018b). There is no easy way to disentangle metaphysical and epistemological considerations, since metaphysical stances often have epistemological implications, and vice versa. For example, if error theory or noncognitivism are true, then it is impossible to have moral knowledge because there are no moral facts. On occasion, philosophers also incorporate epistemic stances in their characterizations of moral realism, e.g., some claim that moral realism requires that we have (or can have) knowledge of at least some moral facts (Miller, 2009; Sinnott-Armstrong, 2009).[52]

Despite their close relationship, questions about folk metaethical realism and antirealism are not intended to directly assess epistemic considerations about the means or possibility of moral knowledge or justification (and to my knowledge, no researchers have suggested otherwise). Rather, they are concerned exclusively with whether there are moral facts (a metaphysical or conceptual question), and if so, whether those facts are stance-independent. Skepticism about moral knowledge

---

[51] I say "typically" because Parfit and Scanlon apparently maintain non-metaphysical notions of realism (see Veluwenkamp, 2017). This is likely a very uncommon view, and it's unclear how plausibly it can be maintained. Nevertheless, Parfit is quite explicit on this point. As Veluwenkamp notes:

Parfit maintains that in the normative domain these truths have "no positive ontological implications" and are not "about metaphysical reality" [...] And for Scanlon, normative truths "need no natural or special metaphysical reality in order to have the significance that we commonly grant them" [...]. (p. 751, see Parfit, 2011, vol. 2, p. 479, p. 747, and Scanlon, 2014, p. 52)

I have no idea what they are talking about. Of course moral realism has metaphysical implications. At the risk of sounding impertinent towards eminent scholars, I think these remarks are implausible and desperate attempts to insulate realism from objections.

[52] This is a reasonable criterion to include, since it allows the realist to exclude undesirable forms of skeptical moral realism: that there are moral facts, but we can't know any of them. People who believe that there are stance-independent moral facts are typically animated not just by the belief that they exist, but by the confidence (or at least hope) that we already know, or could eventually come to know, at least some of those moral facts.

is entirely consistent with the belief that there are stance-independent moral facts. Conversely, the belief that we can have moral knowledge may not directly entail whether that knowledge is of stance-independent facts or knowledge or more relativized or response-dependent moral standards. In other words, skepticism about moral knowledge is consistent with realism, while a belief that moral knowledge is possible is consistent with antirealism. As a result, questions that do not neatly distinguish epistemological and metaphysical considerations from one another risk being unable to identify whether a response reflects an epistemic stance, a metaphysical stance, or both, and thus cannot serve as valid measures of realism/antirealism.

Unfortunately, G&D's early version of the disagreement was worded in such a way that participants could be readily misled into believing they were being asked an epistemological question rather than the metaphysical question. Participants were presented with the standard set of tasks. They were first asked to rate how much they agreed with a series of moral and nonmoral statements, and were then told that a previous study participant disagreed with them. However, they presented participants with the following response options:

*(1)* The other person is *surely* mistaken

*(2)* It is *possible* that neither you nor the other person is mistaken

*(3)* It *could be* that you are mistaken, and the other person is correct.

*(4)* Other (Goodwin & Darley, 2008, p. 1344, as quoted in Pölzler, 2018, p. 659)

Terms such as *surely*, *possible*, and *could be* all carry epistemological connotations. Participants presented with these options may interpret these questions to be asking them about how confident they are in their moral first-order (normative) moral beliefs, which is distinct from whether those moral beliefs are stance-independently true. Indeed, the focal point of responses (1)-(3) all seem to center on epistemic considerations. As Pölzler (2018b) observers, asking whether the other person is *surely* mistaken or whether it *could be* that you are mistaken seems to be asking whether we can be certain of

our normative stance, not whether the moral issue in question is stance-independently true (p. 659).[53] This is because the phrase "could be" is often used to express recognition of fallibility. Since the overall framing may already invite an epistemic reading, interpreting it in this way may even be the most natural response. To illustrate, suppose Alex believes morally wrong. Alex encounters Sam, who believes abortion is not morally wrong. Alex may believe both that (a) there is a stance-independent fact about whether abortion is morally right or wrong (b) but she could be incorrect that it is morally wrong. If so, Alex (a realist) may favor (3), even though this was interpreted as antirealism. In other words, (3) is just as compatible with realism as the judgment that the other person is "surely mistaken." This means that participants that are less confident in their normative beliefs will be rated as less committed to realism, even when this may not be the case. Worse still, the inclusion of both options (1) and (3) may have the collective effect of encouraging an epistemological reading of both, since they appear to offer a contrast between certainty and uncertainty that a conflicting moral view is correct. In short, certainty in our normative moral beliefs is orthogonal to whether there are stance-independent moral facts. A moral realist may believe there is a stance-independent moral fact of the matter, but simply not know what it is in a particular case. An antirealist may be certain of their moral beliefs, or be certain that another person is mistaken, but not believe that these moral beliefs are stance-independently true.

What about response option (2)? At first glance, this seems to be asking a proper metaethical question, rather than an epistemic one. This is because the focus is on whether it is possible that *nobody* is mistaken. This could be interpreted as the suggestion that both you and the other person are correct (relativism) or that neither of you could be correct or incorrect (noncognitivism). Given the wording,

---

[53] Note also that (2) and (3) are not mutually exclusive. There is no reason why a person could not believe both that it is possible neither they or another person is mistaken *and* that it is possible they are incorrect and the other person is mistaken. Presenting participants with options that are not only consistent, but may be attractive choices for similar reasons, is especially problematic, since the reasons why participants may ultimately opt for one or the other of these response options may be especially uninformative with respect to the measure of interest.

it is not possible to distinguish between responses due to subjectivism or to noncognitivism, which poses one limitation on this response option, but both are at least antirealist stances, so as long as participants were interpreting it in one of those two ways, (2) would not be in too much trouble. Unfortunately, "it is possible" is ambiguous between two potential readings, only one of which would reflect a metaethical stance:

(1) Is it possible, *conditional on your metaethical position*, that neither of you is mistaken?

(2) Is it possible, *not conditional on any particular metaethical position*, that neither of you is mistaken?

The first interpretation would comport with researcher intent. If participants judged that, given their view, it was possible two seemingly-conflicting moral claims could both be made without error, then that person would plausibly endorse some form of antirealism. However, on the second interpretation, this response option would effectively reflect that the participant thinks it is possible their *metaethical* stance is mistaken. And metaethical uncertainty is consistent with both realism and antirealism. (1) also seems like an unusually sophisticated consideration that is altogether unclear from the wording of the response option, and may not be the most likely interpretation. Of course, how people interpret these response options is an empirical question. Absent confirmation that it is interpreted in line with researcher intent, it is unclear how to interpret it when participants select this response.

Fortunately, later versions of the disagreement paradigm dropped most instances of epistemic language in their instructions and stimuli, reducing the risk that participants will mistakenly interpret questions in epistemic rather than metaphysical terms. However, even improved versions are still subject to potential conflations between epistemic and metaphysical interpretations. Beebe et al. (2015) and Beebe and Sackris (2016) present participants with the options

(1) It is possible for both of you to be correct.

(2) At least one of you must be mistaken.

(1) was interpreted as an antirealist (and in particular, a relativist) response, while (2) was interpreted as a realist response. Although these response options are an improvement over Goodwin and Darley's wording, these choices still make use of the terms "possible" and "must," both of which may encourage epistemic readings. Unfortunately, the use of epistemic terms in response options undermines the validity of any folk metaethical measures; whenever such language is included, we can no longer be sure that a participant who responds in a particular way does so due to their metaethical stance, or due to an epistemic stance.

Yet even if active efforts were taken to minimize epistemic conflations, they might still persist. This is because the response options commonly used in response options are vulnerable to modal operator scope ambiguity (Millhouse & Bush, 2016). I describe an example of scope ambiguity in the main text (chapter 2, section 2.3.5.2). However, Wainryb et al.'s (2004) findings also suggest that modal operator scope ambiguity may play a role in how participants respond to the disagreement paradigm. Children in all age groups (5, 7, and 9) were more likely to judge conflicting beliefs about ambiguous facts and taste to both be right, a tendency that increased dramatically across age groups from about one third for each among 5-year olds to nearly 70% for disagreements about ambiguous facts and 95% for disagreements about taste among 9-year olds. It is reasonable to expect children to be realists about facts and antirealists about matters of taste, and their responses are consistent with this expectation. This shows that children are not responding in an indiscriminately uniform way across domains. And given that they respond in the same way to moral claims as they do to factual claims, this provides some evidence that they are moral realists.[54],[55]

However, Wainryb and colleagues include two kinds of facts: unambiguous and ambiguous

---

[54] Conventional norms were not included, though they would have been a better domain of comparison.

[55] These findings also provide some support for an early-emerging capacity to distinguish moral from nonmoral norms, and are consistent with the possibility we have an evolved, innate predisposition to distinguish moral norms from nonmoral norms.

facts.[56] Unambiguous have uncontroversially correct answers, e.g., "Paula believes that when you let go of pencils the pencils go up, and Leah believes that they fall down" (p. 691). Ambiguous facts described situations where no information is given that would make it clear which side of the disagreement was correct, e.g., "Ben believes that the dog is not eating because it doesn't like the food, and Lucas believes that the dog is not eating because it's not hungry."[57] If participants reliably understood disagreements across all domains in the intended way, i.e., as questions about the metaphysical grounding of truth claims that are made true by stance-independent facts, facts about the standards of individuals or groups, etc., then epistemic differences should make no difference. To illustrate why, consider two scientific claims: "Hydrogen atoms have one proton" and "Abiogenesis took place on Mars." Both of these claims are either stance-independently correct or incorrect (at least, for those of us who are realists about facts of this kind). It is irrelevant that we happen to know hydrogen atoms have one proton with extremely high confidence, but have yet to discover whether life arose on Mars. Whether Mars ever had life isn't true or false on the basis of subjective attitudes or cultural consensus merely because we don't currently have definitive evidence one way or the other.

Yet this is not how children treated ambiguous facts. While unambiguous facts were judged almost identically to moral disagreements, ambiguous facts exhibited a pattern that more closely resembled taste preferences, ostensibly suggesting high levels of antirealism for ambiguous facts. The percentage of judgments that both people were right when they disagreed about taste were 35%, 66%, and 94% for 5, 7, and 9-year olds, respectively, while the percentage of judgments that both people were right when they disagreed about ambiguous facts were 37%, 48%, and 69% for 5, 7, and 9 year olds, respectively (compared to morality and unambiguous facts, where the antirealist response rate

---

[56] Wainryb et al. refer to these as "facts" and "ambiguous facts," rather than explicitly describing the former as "unambiguous" (p. 692). I have chosen to characterize them as ambiguous and unambiguous to make the distinction more explicit.

[57] Note that the children themselves were not given sufficient information to know why the dog was hungry, so not only did the two people who disagree not know, neither did participants.

approached 0% across all age groups).

Why would so many children treat unambiguous facts as having stance-independent truth conditions, but ambiguous facts as having no stance-independent truth conditions? Is it plausible that when an answer is well-known or easy to confirm that there is some stance-independent fact, but when it is not well-known or easy to confirm that there just is no stance-independent fact of the matter at all? People *might* think this, but this would be an extraordinarily bizarre position to endorse, since it would make stance-independence subordinate to what we have epistemic access to. This would be a very strange kind of realism. Alternatives are no less appealing. Are children arbitrarily endorsing realism towards some factual issues but not others? If not, what pattern is driving this difference in responses, if we take it to reflect a valid measure of whether they are realists or antirealists? And is it plausible that straightforward pluralism is true about people's factual beliefs? That is, that some there are stance-independent truths about some factual issues, but not others?

It is more likely that children didn't interpret what they were asked as intended, and simply conflated epistemic and metaphysical considerations. The alternative is to endorse the implausible conclusion that many children are antirealists about facts, but only when they aren't sure which of two competing claims is correct (or are antirealists about some factual disputes for some other reason). Unless children have extraordinarily baroque metaphysical beliefs about the external world, we should favor the far simpler conclusion that they are simply not interpreting what they are being asked as intended.

Children may even interpret questions about disagreements across domains differently based on the domains themselves or with respect to specific disagreements within domains. When asked to explain why two people with different tastes both could be right, children were readily capable of appealing to the subjectivity of taste, with older children exhibiting far greater likelihood of doing so. For instance, one child justified their judgment that two people with conflicting taste preferences

could both be correct by stating that "People have their own tastes, so both beliefs are right actually" (p. 697). Yet almost no children appealed to subjectivity to justify conflicting beliefs about ambiguous facts both being correct. Instead, the majority appealed to notions of Wainryb categorized as either "truth", i.e., "the beliefs' correspondence with reality" and "uncertainty." Uncertainty reflects precisely the sort of unintended epistemic interpretation I propose here, and accounted for 28%, 41%, and 66% of reasons given for judgments about ambiguous facts for 5, 7, and 9-year old, respectively. For instance, one child explained that "They can both be right because there's no way to know for sure, maybe the dog is hungry and maybe it doesn't like the food" (p. 692). This remark seems far more consistent with an exclusive reading of the truth value of conflicting beliefs, i.e., that one *or the other* could be correct, *but not both.* "Truth" justifications made up most of the remainder, but this category seems to entail first-order judgments about what is true or false, *not* second-order judgments, which would be necessary for participants to have understood the questions as intended. This is because judging that one view is correct (but the other is not) *because* that belief corresponds to reality (i.e., that it's "true") is not necessarily a judgment about realism and antirealism. It is more likely that these responses indicate the judgment that one of the people shares the participant's own (correct) first-order belief about what is true. Since both realists and many antirealists (namely, relativists) can respond to first-order questions about what is true in the same way, such explanations do not necessarily disambiguate realists from antirealists, so categorizing participants who gave these responses as realists is inappropriate.

To illustrate why, suppose a relativist believes abortion is immoral, and is given the following question:

*Alex thinks abortion is immoral. Sam thinks abortion is not immoral. Are both correct, or is only one correct?*
The relativist would not have to state that both beliefs are correct. Instead, they could state that only one belief is correct: Sam's belief that abortion is not okay. This is because the question could be

understood to be asking "Which of these two beliefs do you hold?" Children's explanations for why one belief is correct often seem to reflect this interpretation. For instance, one child said that "What that girl says is wrong and what this one says is right because pencils fall down, for sure, they never fall up" (p. 692). To this child, the judgment that only one of the two views (that pencils either go up or fall down when let go) is correct is simply a recapitulation of the child's own belief about what happens when you let go of pencils. Such explanations were common for ambiguous facts, and since almost all children appealed to the truth of unambiguous facts, these, too, may have been interpreted in an unintended way as first-order questions. Taken together, the explanations children gave indicate that almost all of them interpreted questions about ambiguous facts as epistemic questions or questions about their first-order beliefs about what is true, not questions about the stance-independence of factual claims. This also reveals how epistemic and normative conflations may interact with each other and may mutually reinforce one another. If an issue is unknown, and thus remains epistemically open in such a way that either view could (in an exclusive way) be correct, you may be more inclined to judge which of the views you agree with, rather than judge whether it is possible for two seemingly-conflicting claims to be correct "at the same time." In other words, epistemic considerations may naturally motivate normative interpretations.

Of course, this does not show that participants interpreted moral questions in unintended ways. But even if they did interpret moral questions as intended, if they reliably interpreted questions in other domains in different ways, we would be unable to compare judgments about moral disagreements to what appeared to be the same kinds of judgments in other domains of disagreement, because the content or topic of a disagreement will have resulted in different interpretations of questions that were otherwise structurally identical. This would pose significant concerns about the methodology used in this study, along with any other study that presumes structurally identical questions are interpreted in the same way across domains, items, and conditions. If questions are

interpreted differently based on content while structure is held constant, then many studies will have inappropriately treated differences in responses across domains, between individual items, and between conditions, as measures of the same variable, when in fact context, pragmatic considerations, and other factors resulted in systematic differences in interpretation. This would make sense of the differences in the explanations participants gave for ambiguous and unambiguous facts. Participants frequently appealed to uncertainty in the former, implying an epistemic interpretation, but rarely did so in the latter. This does not indicate antirealism about ambiguous facts and realism about unambiguous facts. Instead, it suggests these questions were interpreted differently, such that many participants interpreted the former as epistemic questions and the latter as questions about their first-order judgments. Wainryb and colleagues do not comment on or appear to acknowledge this possibility when describing explanations participants gave for their judgments, but instead appear to assume all questions were interpreted in the same way, and that the way they interpreted these questions does not threaten the validity of the measures used.

It's unlikely that these differences in responses are an anomaly unique to this particular study. The same cross-domain, cross-item, and cross-condition comparisons are widely used in folk metaethical research among adult populations (e.g. Beebe, 2015; Beebe et al., 2015 Beebe & Sackris 2016; Goodwin & Darley, 2008; 2012; Heiphetz & Young, 2017; Wright et al., 2013), and because these comparisons often play a central role in inferences about the prevalence of folk objectivism, this cross-participant interpretive inconsistency may threaten more than just Wainryb et al.'s conclusions, but findings in metaethics and other areas of research more broadly. If, on the other hand, people did interpret moral items in the same way as facts or ambiguous facts, then participants will have not interpreted what they were asked in the moral domain as intended either, which would invalidate Wainryb et al.'s findings.

And if this is the best explanation for how participants interpreted this and other disagreements in the factual domain it raises doubts that participants interpret moral disagreements any differently. Even if they do interpret questions about morality differently than scientific questions, this would raise a separate methodological concern, since it would demonstrate that questions that are worded in an otherwise identical way are interpreted differently when the domain (moral, factual, etc.) changes. This would suggest that pragmatics play an important role in how participants interpret the questions used in the disagreement paradigm, and that researchers cannot confidently assume any particular interpretation is consistent across domains.

It would also suggest that researchers could not readily make cross-domain comparisons using the disagreement paradigm, since, if participants interpret disagreements differently depending on what the disagreement is about, then they are effectively responding to different questions, even when the stimuli used are superficially the same. Even if modal operator scope ambiguity is not a confound many participants are subject to, it is not plausible that approximately half of people in four different cultures are antirealists about historical events. Whatever the explanation for these responses may be, it is not unlikely that ordinary people are "realism pluralists" about science, history, or mundane facts. This should be a conclusion of last resort.

### S2.4.3 Conflating realism with universalism

Moral realism is often confused with *moral universalism*. Moral universalism is the view that a given moral principle or standard applies to all moral agents, regardless of their location in time or space.[58] For instance, if it is a universal moral fact that it is wrong to own slaves, then it is not only wrong to own slaves in the United States, it is also wrong to own them in any nation on earth, or anywhere else

---

[58] A *moral agent* is any entity that is appropriately subject to moral appraisal. A typical adult human is a moral agent, while babies, nonhuman animals, and inanimate objects are not. This restriction is intended to limit the scope of universality to appropriate targets. A position may still count as *universal* even if it fails to hold lightning morally accountable for striking people.

in the universe, so it would also be wrong for aliens to enslave one another.[59] Moral universalism is sometimes contrasted with moral relativism, in that the former holds that moral standards apply to all people, while the relativist may hold that moral standards can vary depending on an individual's subjective values, or the standards of their culture or group. For instance, a relativist may claim that it is morally wrong for Catholics to have abortions, but it is not morally wrong for non-Catholics to have abortions.[60,61]

Universalism represents one end of a continuum of the *scope* of moral facts. At one extreme, moral facts apply on an individual basis: the individual subjectivist may believe that each of us ought to do that which is consistent with our personal moral standards. On such a view, the scope of a given person's moral standard, e.g., "do not murder," applies only to themselves. Someone else may endorse the same moral rule, but to the extent that moral rule applies to them, it does so in virtue of it being *their* standard, not someone else's. At the other extreme, the moral standard "do not murder" applies

---

[59] Assuming they possess the relevant characteristics to be appropriate subjects of moral consideration, e.g., relevant forms of agency such that they can be held to the same moral standards as humans. For the record, moral philosophers do not reference aliens as often as they should. For instance, most forms of group relativism seem to implicitly refer only to differences between human cultures. Yet in principle one could advocate species-relativism, and defend the view that moral standards can be correct or incorrect relative to the standards of an entire species, rather than to particular cultures within that species. This position is rarely explored, presumably because there are no known alien civilizations to compare ourselves to, but if there were, this might very well be a popular position. It is interesting to note, then, that the conceptual space of metaethical positions that people happen to defend seems to some degree circumscribed by contingent features of our circumstances. If multiple advanced species had evolved on earth (e.g., advanced elephantine or cephalopod civilizations) and existed today, species-relativism might be a common position.

[60] I provide this example rather than an example based on different cultures as a revolt against the common tendency to speak of relativism only in terms of cultural standards, rather than other potential group-based standards. Relativists need not hang their hat on one, and only one way in which one's standardized can be relativized.

[61] Note that universalism is not the same as absolutism. An absolute rule is a moral rule that admits of no exceptions, e.g., an absolute moral rule against abortion would hold that abortion is always wrong. But *always* in the absolutist sense differs from the universalist sense. Universalism concerns who a moral rule applies to; it is not a feature of the content of the moral rule itself. Such rules can be absolute or non-absolute. For instance, it could be a universal moral fact that abortion is prohibited in some circumstances but not others. For instance, abortion could be permitted up until the third trimester. This would mean the non-absolute rule "abortion is morally permissible until the third trimester" would apply to all people. Moral absolutism, on the other hand, is about the content of the moral rule. The rule is absolute when a certain action admits no exceptions, e.g., an extreme pacifist who believes violence is *always* wrong adheres to an absolute moral principle: violence is wrong, *even in self-defense*, *even to prevent someone from committing a greater amount of violence*, etc. Absolute rules need not be universal. It could be that some people or groups are subject to a particular absolute rule, but others are not. For instance, it could be that Catholics are prohibited from using any form of birth control for any reason (an absolute moral rule), while non-Catholics are not.

to everyone, everywhere, in all times and places. Other positions may fall somewhere between these two extremes e.g., cultural relativists may believe a moral rule applies to all members of a particular culture. It is also possible to believe some moral rules apply universally but others do not.

Regardless of where one falls on the continuum between moral facts being more or less universal in scope, universalism/localism is orthogonal to the distinction between realism and antirealism. Some researchers have drawn explicit attention to this and have sought to carefully avoid conflating universalism with realism (e.g., Goodwin and Darley, 2008).[62]

Recently, however, Ayars and Nichols (2018) and Nichols and Rose (2019) have opted to use the disagreement paradigm to distinguish folk universalism from relativism, rather than to distinguish folk realism from antirealism. That is, rather than using the disagreement paradigm to assess whether ordinary people believe there are stance-independent moral facts, they use it to determine whether ordinary people think there is only one correct moral standard that applies to everyone (universalism), or whether there is more than one correct moral standard (relativism).

The rationale for this move is simple: when participants select the response option "at least one of you must be mistaken," this implies that the participant believes there is a single correct answer. Yet this does not tell us whether they think that what makes that moral claim true is *stance-independent*. This would be moot if universalism entailed stance-independence, and relativism entailed antirealism, but this is not the case (Rose & Nichols, 2019). It is possible to believe that there is a single correct moral standard, but that it is not stance-independent, and it is also possible to believe there can be

---

[62] Goodwin and Darley are very clear not to mix the two up:

"[T]he question of whether ethical standards should apply to all cultures is a question about the *scope* of ethical standards, and is independent of the question of whether such standards and beliefs are objectively or subjectively true. Our interest centers on this second question, which concerns the *source* of such beliefs or standards - whether they derive their truth (or warrant) independent of human minds (i.e., objectively) or whether instead, their truth is entirely mind-dependent or subjective." (p. 1341)

more than one correct set of moral standards, but that those standards are stance-independent.[63] For instance, some universalists believe that moral facts

In short: universalism and relativism are both compatible with realism and antirealism, so a measure of the former cannot be used as a measure of the latter. And since the response options in conventional versions of the disagreement paradigm are more suited towards capturing the universalism/relativism distinction than the realism/antirealism, distinction, since they at best can only tell us whether people think there is one or multiple moral standards (but cannot tell us whether they are mind-independent), they are not a valid measure of realism/antirealism. Joyce (2015) is especially careful to tease out the distinction between stance-independence and the universalism/relativism debate. Joyce characterizes relativism as the view that moral claims "contain an essential indexical element, such that the truth of any such claims requires relativization to some individual or group." The difference between individual subjectivists and cultural relativists, for instance, is that the former relativize moral claims to the subjective standards of each individual, while the latter relativize moral claims to the standards of different groups. Yet, Joyce adds,

> [...] it *may be* that what determines the difference in the relevant contexts is something "mind-dependent" —in which case it would be anti-realist relativism—but it need not be; perhaps what determines the relevant difference is an entirely mind-independent affair, making for an objectivist (and potentially realist) relativism.

Joyce offers the example of *tallness*. Tallness is a relative notion. Whether a person is "tall" only makes sense relative to some standard. However, whatever that stance may be, whether a person is tall or

---

[63] To put it in their own words, Rose and Nichols (2019) observe that:

"[T]he term 'moral objectivism' often implies something stronger than the rejection of relativism; on one such usage, 'objective' moral claims purport to describe facts or properties that are independent of anyone's feelings or attitudes about the claims. One can, however, reject relativism without committing to mind-independent moral facts. The core claim that relativism rejects is that there is a single true morality." (p. 61)

Note that because I have opted to use *realism* to refer to stance-independence, these references to objectivism could be substituted for realism with no change in meaning. Rose and Nichols are thus quite clear that studies designed to evaluate realism/antirealism are in fact only suited to distinguishing universalism from relativism.

not is not stance-dependent since it is not *made true* by our thinking it is the case. For example, at the age of ten, Alex may be tall relative to other children of the same age, but not tall relative to adults. But whether Alex is tall relative to a particular standard, e.g., "children of the same age" is made true by stance-independent facts, *not* subjective standards or the consensus of a culture. It is possible for someone to endorse a relativistic moral standard that is also stance-independent. For instance, someone could believe that everyone is obligated to abide by the dictums of a council of elders or a supreme leader. Such views aren't a genuine form of relativism, since moral facts all depend on the same stance-dependent source, and thus cannot vary according to different standards.

The distinction between such *relation-designating*[64] accounts and relativism is especially relevant to the disagreement paradigm, since the disagreement paradigm relies on antirealists not selecting the response option that at least one person must be incorrect. But this is exactly what a person who endorses a relation-designating account ought to choose, since they believe there is only one universal (but stance-dependent) moral standard.

It is possible that, among ordinary people, there is a close (albeit contingent) link between stance-independence and universalism on the one hand, and stance-dependence and relativism on the other, but this is not a position we are entitled to presume from the armchair. After all, the whole rationale for conducting empirical research is predicated on the methodological inadequacy of armchair speculation about folk metaethics. If measures that capture the universalism/relativism divide also capture the realism/antirealism divide, this would itself have to be established empirically. At present, there is no data to support such a connection. At present, the disagreement paradigm may be a face valid measure of the distinction between universalism and relativism, but it is not a valid measure for determining whether ordinary people are realists or antirealists. It *may* be a valid measure

---

[64] Joyce (2015) draws on an obscure label, describing positions like these as "relation-designating accounts" (see Stevenson, 1963, p. 74).

of the folk realism/antirealism distinction, but if so, it may be using universalism and relativism as proxies for realism and antirealism when there is no evidence that they can adequately serve in this role.

In short, the goal of the disagreement paradigm is to determine whether ordinary people endorse realism or antirealism, but the response options participants are typically given only allow us to determine whether they believe there is a single moral standard (*universalism*) or whether there can be more than one (*relativism*). However, universalism is compatible with antirealism, and while relativism is often categorized as a form of antirealism, this is not always the case. As Joyce (2015) observes:

> Moral relativism is sometimes thought of as a version of anti-realism, but (short of stipulating usage) there is no basis for this classification; it is better to say that some versions of relativism may be anti-realist and others may be realist.

Given this, the "relativist" response option does not conceptually entail antirealism. Since it is theoretically possible for participants to interpret the question in line with researcher intent-that is, to accurately indicate that they endorse relativism-but for that participant to still be a moral realist (since they moral facts are stance-independent), the disagreement paradigm formally conflates an unintended metaethical distinction with the distinction it was constructed to measure.

This regrettable flaw in design reveals how even highly competent researchers can conflate subtly distinct dichotomies. The distinction between universalism/relativism and realism/antirealism is subtle, requiring us to carefully tease apart considerations that are non-obvious, difficult to grasp, and even more difficult to articulate. Adequately characterizing these positions often calls for the introduction of technological jargon and clumsy neologisms that trip up people with decades of philosophical training. If specialists struggle with these distinctions, it is easy to imagine that ordinary people would, too.

Thus, even if the disagreement paradigm were carefully reworded to more directly probe realism and antirealism, ordinary people may still conflate the distinction between universalism/relativism and realism/antirealism, and respond accordingly. Ensuring participants do not conflate the realism/antirealism distinction with similar, but distinct metaethical distinctions would require dedicated efforts to validate any proposed paradigm by providing evidence that participants reliably interpret what they are asked in the intended way, rather than in some other way. Given that the very researchers designing these studies, and who are explicitly aware of the distinction, nevertheless struggle to devise questions that adequately disentangle the two, it may be quite difficult to frame questions in a way that doesn't risk participants mistaking a question about stance independence for a question about scope.

### S2.4.4 Conflating realism with absolutism

Nothing about realism and absolutism conceptually bundles them together, such that one must be closely linked to the other. In practice, however, there may very well be a relation between the two. Relativism (and perhaps antirealists positions in general) are often associated with *tolerance* (Bush, 2016, Collier-Spruel et al., 2019). Realism, in contrast, may seem comparatively more dogmatic, rigid, and moralistic, and for some may have an unappealing religious vibe. And perhaps, in practice, precisely those people who are most inclined to endorse moral realism really do also tend to endorse more rigid and exceptionless moral rules. People may be picking up on a genuine, if contingent, connection.

This could mislead participants into believing that to endorse a response option that ought merely to reflect a belief in stance-independent moral facts would also (or instead) commitment them to an absolutist standard towards the moral issue in question, even when this is not the intent of the question nor an entailment of a realist response. Once again, whether or not participants are prone to such an error is an empirical question. While there is some evidence that a handful of participants interpret the realist response to the disagreement paradigm to reflect absolutism, only a handful of

participants respond in a way that suggests they might be conflating realism with absolutism. In one study, I told participants that I had asked another participant the following question:

*When two people disagree about a moral issue, do you think they can both be correct, or must at least of them be incorrect?*

And that the participant responded:

*"When people disagree about a particular moral issue there can be at most only one correct answer"*

This response is intended to be similar to the realist response option in the disagreement paradigm. While few participants interpreted this response in a way that clearly indicated they interpreted as indicating a commitment to realism, their descriptions rarely indicated a conflation between realism and absolutism. Even so, a few responses hinted at the possibility of this conflation:

**Response #1**: *I think he means that there is only one answer to a moral dilemna. That there is no exceptions. Only one right way.*

**Response #2**: *He means that there are not two ways to look at an issue, it is either right or wrong despite any other circumstances.*

Neither of these responses unambiguously demonstrates a conflation between realism and absolutism. Although the first respondent references "That there is no exceptions [sic]" while the second refers to something being either right or wrong "despite any other circumstances," both of which could suggest a realist/absolutist conflation, the former may be endorsing something closer to universalism, while the latter might mean that there is a definitive answer to moral issues, even when they are tricky and involve some potentially exculpatory circumstances, which could reflect an epistemic or metaethical interpretation. Open response questions that assess how participants interpret other questions about metaethics aside from the response options used in the disagreement paradigm likewise reveal little evidence that ordinary people tend to conflate realism with absolutism.

This is hardly definitive, but it does cast doubt on the possibility that participants tend to conflate realism with absolutism. So why bring it up? First, it is important to consider all plausible

conflations, even if we ultimately rule them out. Second, while participants rarely seem to conflate realism with absolutism, understood in narrow philosophical terms, they do appear to associate the realist response with stances and attitudes that, at least to ordinary people, may seem conceptually adjacent, such as *rigidity*, *narrow-mindedness*, "*black and white*" thinking, and a "*closed*" attitude towards moral disagreement (Goodwin & Darley, 2012). In other words, the realist response may be seen not to reflect the view that there is a stance-independent moral fact about the issue in question (or at least not *only* reflect such a view), or that there are exceptionless moral rules, but may indicate that the person favoring the "realist' response is unwilling to consider contrary perspectives or change their mind, is intolerant towards people with other moral perspectives, pushy, self-righteous, and dogmatic about their moral stance, unwilling to consider nuance or exculpatory considerations, convinced that answers to moral questions are definitive and perhaps obvious, and so on. There is no well-established philosophical term that adequately captures this cluster of concepts, nor is it clear there ought to be. These concepts don't perfectly overlap with one another, and run the gamut from personality traits to epistemic standards to genuine normative and metaethical beliefs.

At the risk of oversimplification, this cluster of concepts seems to center on a perspective towards morality that is close-minded and lacking in nuance. Insofar as participants associate the realist response option with these generally undesirable qualities, they may judge that this does not accurately reflect their view towards morality or towards a particular moral issue, and select some other response instead. Although the realist response option is not strictly intended to reflect such attitudes, it is not hard to see why people might interpret the response option in this way. Ordinary people often invoke the idea of "grey areas," intermediate space between two extremes. There may be a simple and straightforward answer to many moral questions, e.g., there are no conceivable circumstances in which *genocide* would be morally permissible. Yet other actions, just as *lying, stealing*, or *killing* represent a much broader range of actions, which include actions that are clearly permissible and clearly impermissible,

but that, crucially, also include issues where people are uncertain or ambivalent. Killing *just for fun* may seem obviously immoral, while killing *a group of terrorists intent on torturing you and your family* may be obviously morally permissible, but would about *killing a terminally ill patient in severe pain who is begging to die*? Ordinary people often consider euthanasia a "grey area." This is a loose and colloquial notion, but roughly speaking, people recognize that such situations are difficult to judge in a confident and conclusive manner. There are at least a couple of reasons this is the case, though there is likely more to it than this.

One reason is that the issues that people think of as falling into a "grey area" often invoke conflicting intuitions or moral standards. Consider the classic case of a person violating a strong norm in order to achieve an altruistic goal, e.g., "stealing from the rich to give to the poor." Such actions involve both a moral violation (stealing) and a supererogatory moral act (taking on personal risk to aid the impoverished solely for their benefit despite having no obligation to do so). This action is not strictly speaking *good* or *bad*, it's a bit of both. Ordinary people recognize this, and may find that there are circumstances where they are unsure whether the bad outweighs the good.

It could also be that whether a given instance of euthanasia is permissible or not will depend on details that are highly specific to each instance, and that are difficult to assess. For instance, is the patient in full possession of their cognitive faculties? If the person considering euthanasia has a history of severe mental illness, we may oppose euthanasia in their specific case, while if they have a terminal illness and psychiatrists judge them to be of sound mind, we might judge that it is permissible in that case. But what if a person has mild dementia? We may be uncertain. Regardless, if asked whether euthanasia is permissible, many of us might be inclined to say that *it depends*.

When a given issue is difficult to judge, involves conflicting intuitions and moral standards, and is highly contingent on highly variable circumstances, people may deem the issue in a "grey area." Under such circumstances, *even if* people felt that, with perfect access to all of the relevant nonmoral

considerations, there may be some stance-independent fact of the matter, in practice, we do not have access to these facts. And in the spirit of epistemic humility, tolerance, and ecumenicalism, we should cordon off moral issues that fall into this "grey area" as issues people are permitted to reach different conclusions about. It is not at all implausible for someone to judge that, e.g., they would not personally get an abortion, but that it is permissible for others to do so. Such a person may feel that abortion is probably wrong, but that they are not in a position to impose this moral standard on others.

Given this folk notion of moral "grey areas," it should now be apparent why participants may perceive the "realist" response to go beyond merely expressing that there is a stance-independent answer to any given, well-specified moral issue, even if we don't know what that answer is. When the realist reacts to a *specific* moral issue by judging that *at least one person must be mistaken*, they may seem to be saying something like, "this moral issue does not fall into the grey area we all recognize and mutually respect, but instead has a clear and decisive answer." Such an attitude towards moral issues may imply a whole host of unpalatable characteristics. A person who held this attitude may be violating an implicit compact of tolerance for people whose moral disagreements fall within an acceptable range of views. Most participants are unlikely to tolerate members of their community who openly endorse genocide, but would not ostracize people that disagree about euthanasia. Someone who insists that both issues have definitive answers may seem like they fail to appreciate that many issues are complex and difficult to judge, they may seem insensitive to context, intolerance of opposing positions, and overconfident in their particular stance. This last possibility is especially troubling.[65]

When a person judges that at least one side of a disagreement must be mistaken, it is unlikely that they have in mind their own side. Rather, they think anyone who disagrees with them must be

---

[65] Even if participants do recognize that they have an inflexible and dogmatic moral stance, there may be social incentives to not admit this in the context of an experiment. And participants may anticipate (and wish to avoid) reputational costs for selecting the realist response option, even if they recognize that choosing it does not necessarily entail these undesirable traits. I discuss this possibility in **section S2.10**.

mistaken. Such a person thus doesn't merely think that there is some fact of the matter, even if they don't know what it is. Rather, they think there is a fact of the matter, and it is exactly what *they* think it is. It is one thing to express confidence that people who support murder and genocide must be mistaken. It is quite another to say that anyone who disagrees with you about a complicated topic most of us regard as controversial, e.g., abortion or murder, *must be* mistaken. This carries both epistemic connotations and implications about the stance someone takes towards their first-order moral judgments: that their moral standard is the only acceptable one regarding this particular issue, which connotes rejection of the ecumenicalism and tolerance people often reserve for controversial moral issues.

The first glimmers of skepticism that motivated my doubts about the validity of the disagreement paradigm were driven more by the suspicion that found the questions underspecified and confusing, and that people were motivated to hedge against their uncertainty about controversial moral issues. It was not obvious to me that identifying realism with a dogmatic, unsophisticated, "black and white" stance towards moral issues would be a dominant factor. But when participants were asked to evaluate the disagreement paradigm, their responses frequently reflected just these kinds of concerns with the realist response. In fact, they were so frequent it may represent one of the most common conflations. In one study, I told participants that they would be reviewing a question and response from a previous participant. The previous participant was asked:

> *When two people disagree about a moral issue, do you think they can both be correct, or must at least of them be incorrect?*

John: *"When people disagree about a particular moral issue there can be at most only one correct answer."*

We then asked participants:

> *In your own words, what do you think the respondent means in the statement above?*

Here are some of the responses participants gave:

**Response #1**: *It means that he believes there is only one correct answer. He sees thing is black and white, either you're write or you're wrong, either you're moral or immoral.*

**Response #2**: *He means that an answer to a question is black and white, there is a right answer and a wrong answer and no in between.*

**Response #3**: *That there can be only one answer when two people disagree about a moral issue. That there is no gray area, only black and white.*

**Response #4**: *That there is no grey area in situations. Either you are right or wrong.*

**Response #5**: *"Everything is black or white and there is always a right or wrong answer. Thinking about both sides is a waste of time."*

**Response #6**: *There is one true right/wrong and there's not shade of grey.*

About 20% of the participants explicitly reference John as having a "black and white" view of morality, or denying that there are "grey areas." This may not seem like much, but this is just *one* of the ways participants can interpret the realist response in a way that diverges from researcher intent. Given all the other ambiguities and unintended interpretations discussed here, 20% is a lot. It also represents a lower bound on the total number of people who regard the realist position as dogmatic, narrow-minded, or unsophisticated. Many participants provided shallow or underdeveloped responses, e.g., merely repeating that John thinks that one view is correct:

**Response #7:** *There is a right and there is a wrong. Only one thing can be right.*

**Response #8**: *Most of the time there is only one correct answer.*

These participants did not explicitly reference any other interpretation (e.g., correctly interpreting the statement to reflect a stance-independent view about moral claims), so it is hard to judge whether they may have also regarded the realist response as expressing a "black and white" or dogmatic view of morality. Since it is plausible at least some would attribute these qualities to John, 20% is an underestimate, and perhaps a substantial one. More importantly, participants often expressed in various direct or indirect ways that John's realist response indicated that he thought of moral issues as

having clear and discrete answers without explicitly using the terms "black and white" or mentioning "grey areas":

> **Response #9**: *I think he means that there is only one answer to a moral dilemna. That there is no exceptions. Only one right way.*

> **Response #10**: *I think that he means that their can only be a right or wrong, or good or bad when it comes to actions, no middle ground.*

Finally, some participants emphasized that John's response indicates that he's not receptive to changing his mind, or that he is convinced that his moral stance is the correct one:

> **Response #13**: *That what you believe is the only true and not open to hear more*

> **Response #14**: *I think the respondent means that there is only one way to see morality and that he is not open to other's opinions.*

> **Response #15**: *He feels certain that his moral view is the only answer*

> **Response #16**: *That his moral standpoint is the correct one.*

When asked what a person would mean if they expressed a view towards moral disagreement that mirrors the wording typically employed in the disagreement paradigm, a substantial number of people judge the realist response option to reflect a closed, dogmatic, narrow-minded, black-and-white perspective on morality that is insensitive to context. Thus, it would appear that many people interpret the realist response option to reflect stances and attitudes towards moral disagreement that do not reflect realism. Since these qualities are likely to be unappealing to participants, this probably discourages many participants from selecting the realist response option for reasons unrelated to their endorsement of some antirealist position.

The conflation between realism and dogmatism is further confounded by the possibility that the degree to which people regard a realist response option to reflect dogmatism (and the other qualities mentioned here) will systematically vary in accordance with the content of the moral issue in question. In other words, it is not simply that people would regard the realist response option to

uniformly suggest dogmatic or black and white thinking for *every* moral disagreement, but that the degree to which people associate the realist response with these negative qualities will depend on the particular moral or nonmoral disagreement. For instance, people may think that insisting that there is only one correct answer to whether rape or genocide are immoral is not close-minded, dogmatic, and so on. Yet someone who thinks there is a single correct answer to whether controversial (e.g., abortion or euthanasia) or underspecified moral issues (e.g., killing or stealing), *would* be perceived as dogmatic and close-minded.

In other words, the conflation between realism and dogmatism is not uniform. This interpretative inconsistency threatens the validity of the disagreement paradigm in a way distinct from the simple fact that many participants interpret the realist response option in unintended ways. In order for researchers to make cross-item comparisons, participants must interpret the response options to the disagreement paradigm in a uniform and consistent way across all items. In other words, they must understand the response option "at least one person must be mistaken" to mean the same thing when it is presented as a response option to a disagreement about abortion as they do when it is presented as a response option for a disagreement about murder, or for disagreements about nonmoral issues such as disputes about aesthetic or scientific claims. If participants do not interpret the response options the same way, then no matter how participants interpret each question, we cannot aggregate responses with a domain, or make comparisons between a domain, and treat these as measures of the same variable.

For instance, if participants interpret the realist response option to moral disagreements about murder to reflect realism, but they don't interpret the realist response option to reflect realism for abortion, then we cannot take a rate of, say, 80% realist response for disagreements about murder and 30% for disagreements about abortions and say that more people are realists about murder than about abortion, because these percentages won't reflect the true proportion of people who endorse realism

about murder and abortion. The item would be valid for murder, but not for abortion. If some subset of participants *uniformly* interpreted all questions used in the disagreement paradigm in unintended ways, we could perhaps account for this by using e.g., comprehension checks or open response questions, and excluding participants who described the realist response as indicating a "black and white" view of morality. Aside from the methodological problems excluding large numbers of participants would introduce, this would not be viable if participants interpret response options differently for each moral issue. Instead, we'd have to check how they interpret the response options for *each and every moral disagreement.* This would not normally be a problem; studies can withstand minor interpretative variation. But in this case, we have compelling reasons to suspect that interpretative variation is significant and substantial, and that many interpretations diverge from the required interpretation to such an extent that response options predicated on these interpretations are not valid.

This points to a more general problem with social scientific research that involves taking a standardized wording, then swapping out portions of the content. That is, the disagreement paradigm uses a standardized wording and set of response options. Participants are told that someone disagrees with them about [moral issue] or that two other people disagree about [moral issue]. Then they are asked whether both people can be correct, or whether at least one must be incorrect.

Yet researchers mistakenly presume that if they hold the semantic content of a set of questions constant across items, that this ensures participants will interpret questions in a uniform and consistent manner. With respect to the disagreement paradigm, this means that if they swap out the [moral issue] above for murder or abortion or even nonmoral issues, such as disputes about science or aesthetics, that participants will interpret all of these questions the same way. Yet there is good reason to suspect they would not. Researchers are far too insensitive to the influence pragmatics can have on the meaning of their items and response options, and that pragmatic variation can threaten the validity of a paradigm by introducing cross-item interpretive variation.

Since researchers using the disagreement paradigm have done no work at all to account for variation in response option meaning on the basis of the variation in the meaning of response options attributable to pragmatics, we cannot know whether differences in response options are best explained by differences in interpretation of the question due to pragmatics, or differences in the participant's realist/antirealist stances or commitments. In other words, all cross-item comparisons in rates of realist/antirealist responses are confounded by the potential for these differences to be due instead to cross-item variation in interpretation. Such variation is further compounded by potential variation in demand characteristics and social and reputational concerns.

For instance, people may feel comfortable declaring someone who disagrees with them about racism to be incorrect, but less comfortable declaring that someone who disagrees about euthanasia must be incorrect. The former carries no significant reputational consequences; on the contrary, it may be costly *not* to select the "realist" response, while the opposite may be true of the latter. In other words, whenever participants are presented with concrete moral disagreements, the degree to which a participant would regard a realist response to reflect unappealing dogmatism and narrow-mindedness is likely to vary based on the content of the item. In some cases, a "closed" attitude towards disagreement may be unappealing, but for other moral disagreements it may play little role or even be positively expected of people. In a society that widely condemns racism, slavery, or genocide, dogmatic opposition that refuses to permit exceptions or consideration of context might be expected of us. Much intrapersonal variation in participant response across different moral disagreements could reflect variation in sensitivity to considerations like these, and without knowing precisely how each participant perceives each individual concrete moral issue, it would be difficult to know whether responses reflect sensitivity to these considerations rather than genuine intrapersonal variation in metaethical standards towards different moral issues.

Thus, participants may not simply interpret the disagreement paradigm in unintended ways, or have incentives to answer in ways unaligned with their stance towards realism, but rather that interpretation and social incentives to answer in particular ways will vary on an item-by-item basis that no researchers have accounted for and that would make cross-item comparison at best a methodological nightmare that it may be impractical, or even impossible, to adequately address using conventional social scientific methods. At the very least if these problems were surmountable, nobody has even begun to address them.

### S2.4.5 Conflating relativism with contextualism

Compounding the possibility that many participants may be disinclined to select the "realist" response option when it seems to an unappealing attitude towards moral disagreement (either in general or with respect to specific issues, e.g., euthanasia) disagreement, participants may also interpret relativist and noncognitivist responses in a positive light when presented with moral disagreement in the abstract, or with respect to certain issues where open-mindedness, or sensitivity to context would be seen as desirable.

Note that, once again, such response options would reflect a *conflation* between what is intended to reflect a metaethical stance, but is instead interpreted to reflect in part or in whole some other stance or attitude. A separate, but related issue would occur whenever participants correctly interpret the disagreement paradigm to be asking metaethical questions, but anticipate that the response options on offer would signal desirable or undesirable beliefs or attitudes. Yet in many cases participants may instead interpret the responses to properly constitute an expression of tolerance, flexibility, open-mindedness, or their converse. Just as the judgment that at least one person must be mistaken could be understood to reflect insensitivity to context, the view that both people could be (or are) correct could reflect sensitivity to context, open-mindedness, or other stances towards moral disagreements

that a participant would endorse. Take, for instance, a hypothetical disagreement about whether it is morally wrong to lie:

Alex: *"It is morally wrong to lie."*

Sam: *"It is not morally wrong to lie."*

A natural interpretation of this disagreement is to append an implicit "always" to Alex's remark, i.e., it is *always wrong to lie*. Interpreted in this way, most people would likely agree with Sam, and view Alex as unreasonably dogmatic and absolutist about the moral status of lying. Surely *some* lies are permissible. It is a well-worn trope that we may (or must) lie about someone's location if a deranged psychopath with an ax shows up demanding to know their whereabouts. Many of the moral disagreements participants are given may be less straightforward than this, but the response option that both people could be correct may nevertheless be interpreted to reflect a sensitivity to the fact that a given action type (e.g., "lying") may be permissible or impermissible depending on the circumstances. In short, many participants may simply interpret the "relativist" response to reflect, roughly, that whether a given type of action is morally right or wrong *depends on the circumstances.*

Participants presented with the disagreement paradigm and a supposed relativist response sometimes invoked this explanation. I presented one set of participants with a question similar to the one above, but with John espousing a relativist remark rather than an objectivist one:

*When two people disagree about a moral issue, do you think they can both be correct, or must at least of them be incorrect?*

John: *"When people disagree about a particular moral issue each can be correct according to their own moral standards".*

Once again, I asked participants:

*In your own words, what do you think the respondent means in the statement above?*

A handful of respondents clearly associated the relativist response with open-mindedness:

**Response #1**: *The respondant is being open minded in seeing all views and perspectives*

**Response #2**: *that people shouldn't be close minded, they should be open to other people's views.*

**Response #3**: *You must understand a person's point of view before making a decision about them, their way of thinking may be completely than yours and it is good to hear why a person believes what they believe.*

At least one believed the relativist meant that we should not enforce our views on others:

**Response #4**: *It means when it comes to morals, every one has their own set of morals, so people shouldn't try to enforce theirs on others.*

Finally, a few referenced sensitivity to context:

**Response #5**: *People have different standards regarding their moral judgement. It is unwise to judge one's action solely based on the judge's opinion without considering the person's situation.*

**Response #6**: *Morality depends on the situation and the society.*

Even so, such responses were not that common. Most participants who did not interpret the statement as intended interpreted in ways unrelated to sensitivity to context or open-mindedness. This is not surprising. This particular way of understanding relativism (which does not accurately reflect its academic counterpart) is just *one* way that participants could interpret it. It is also possible, given the design of this study, that participants would recognize such associations if prompted, but did not consider them central or primary to the meaning of the statement.

Yet additional evidence that a significant number of participants interpret the relativist response in this way comes from asking participants to directly respond to an abstract version of the disagreement paradigm, then asking them to explain their response. I asked participants:

*When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?*
*Please briefly explain why you chose this response.*

Many explanations alluded to sensitivity to context or a rejection that morality is "black and white":

**Response #1**: *The nuance can be correct as morality can be situational.*

**Response #2**: *Because morality is very complicated. There is no one true "truth." People can differ and both share aspects of the truth.*

**Response #3**: *I don't think the world is "black and white", it's mostly "grey".*

**Response #4**: *Morals aren't black and white, there is almost always a gray area.*

Even if such views are only explicitly articulated by a minority of participants, such reactions are common enough that they pose a threat to the validity of the disagreement paradigm. It seems that many people conflate the relativist response option with various considerations other than the belief that moral claims are best understood to reflect indexicalized truth claims about the moral standards of the speaker or the speaker's culture. While some participants do explicitly interpret the relativist response option in this way, when asked directly what such a response means, most do not, with a handful referencing the sensitivity to context or the view that morality has many "grey areas." And when asked to explain their own preference for the view that two people can be correct, a substantial subset of participants explain their reasoning by appeal to sensitivity to the circumstances or other notions that would allow both people to be correct about some circumscribed aspect of the moral issue in question, or to have a valid or justified perspective, or to have part of the moral truth, and so on. That this *particular* conflation does not comprise a majority of interpretations does not minimize the threat it poses. The interpretation variation that undermines the disagreement paradigm is more a death by a thousand cuts than a single fatal confound. The conflation between realism and a dogmatic and unsophisticated moral stance on the one hand, and the relativist (antirealist) response with sensitivity to context is just one among several ways participants do not interpret the disagreement paradigm as intended.

### S2.4.6 Conflating relativism with descriptive claims

As it is used here, *relativism* refers to the metaethical position that there are moral facts, and that those facts are true or false only relative to the standards of different individuals or groups. Yet moral

relativism is sometimes used to refer to *descriptive moral relativism*, the empirical hypothesis that there are pervasive and fundamental differences in the moral standards of different individuals or cultures (Bush, 2016; Gowans, 2021; Levy, 2003).

Among philosophers, metaethical relativism draws much of its justification from the alleged truth of descriptive relativism. If it seemed that there was little cross-cultural or interpersonal variation in moral standards, this would plausibly undermine much of the motivation for supposing that metaethical relativism was correct, while if there are widespread and seemingly irreconcilable differences, their persistence may provide some indication that there is no single, correct moral system. Given their association, it is possible participants sometimes conflate metaethical relativism with descriptive relativism. In other words, participants may interpret the disagreement paradigm to be asking whether, as a matter of psychological fact, people can or do hold conflicting moral standards. This may seem unlikely, given the wording of the disagreement paradigm. After all, participants are asked to judge whether one or both positions is *correct*, not whether two people have different beliefs *about* what is correct. Nevertheless, as I demonstrate in **Chapter 4**, when participants are asked to explain why they selected their response to the disagreement paradigm, they sometimes offer an explanation that suggests that they interpreted the question in descriptive rather than metaethical terms. For instance, I asked participants the following question:

> *When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?*

After answering, they were asked to briefly explain why they chose their response. A handful of responses do appear to reflect a descriptive interpretation:

> **Response #1:** *Each person has their own set of moral beliefs. The way moral beliefs work is that they can vary.*

> **Response #2**: *They can be right in their own way of thinking. It changes a lot, when the perspective is different.*

**Response #3**: *Everyone believes different things*

**Response #4**: *People have different opinions, values, and beliefs. What is moral in person's eyes may be immoral in another.*

**Response #5:** *Everyone is different when it comes to their beliefs and how they view what is right or wrong.*

These responses are all consistent with the intended metaethical interpretation. After all, participants could be explaining why they endorse metaethical relativism by appealing to descriptive relativism. Yet it is also possible that they interpreted the question to be asking about whether two people with different moral beliefs could be correct *according to their own standards*; that is, they could have interpreted the question to be one about the plausibility or acceptability of moral disagreement. This might seem implausible, but this could be because researchers (myself included) know what the question is *supposed* to be asking, and are more familiar with using the term "correct" in a strict, truth-correspondence sense. Yet in colloquial speech, people often use "correct" not to refer to which views they themselves think are correct or not, but to describe what people *believe* is correct. For instance, people often say things such as "It's correct *according to her*." Even so, it might seem unlikely that people would interpret the question to be asking something as mundane as whether people have different moral beliefs. Who would deny that?

Nevertheless, these explanations hint at the possibility that participants did interpret the question in this way. None of these examples illustrate any attempt to explicitly connect the fact that two people disagree to the view that conflicting moral standards can both be correct in a truth-correspondence respect, so while this *may* be what they have in mind, it is not clear that it is. And in a few instances, their remarks seem more in line with the descriptive reading than an implicit justification for their metaethical stance. Take respondent #2. They state that "They can be right *in their own way of thinking*." This participant seems to be more concerned with what participants believe is correct than

what is in fact correct. Other responses seem to even more clearly reflect a descriptive reading of "correct":

> **Response #6**: *Being correct about a moral decision is in the eye of the beholder, what is correct to one person isn't always correct to the other.*

This participant does not appear to be using the term "correct" to refer to what is in fact true, but is instead using it to describe what people *believe* is true. Note that they say that what is correct to one person isn't always correct *to the other*. This indicates that the participant is focused on people's beliefs about what is correct, not what is in fact correct. Still others do draw connections between moral differences. Finally, consider this response:

> **Response #7**: *Because we all have different perspectives and our perspectives determine our beliefs about right or wrong.*

This participant appeals to the fact that we each have different perspectives on what is morally right or wrong, but rather than concluding that each of these perspectives is *correct*, they state that each of these perspectives determine *what we believe* is morally right or wrong. It is possible this participant interpreted the disagreement paradigm in the intended way, but their explanation puts some strain on possibility. Responses like this suggest that participants are often inclined to think in terms of what people believe is correct rather than what they (the participant) thinks is correct.

There may be some social incentive to interpret the question in this way. In ordinary social settings, it may seem rude to declare that another person is incorrect. One way to avoid stating that others are incorrect is to focus not on which beliefs are correct, but on the fact that each of us has a different perspective on what is true, or the fact that it is acceptable for us to do so. Doing so may signal prosocial personality traits, such as tolerance for divergent moral beliefs. Of course, it is also possible that participants interpreted the disagreement paradigm itself in metaethical terms, but offer explanations that signal these traits. If so, then these participants would have interpreted the disagreement paradigm as intended, but their explanations would give the erroneous impression that

they didn't interpret what was asked as intended. This possibility strikes me as fairly plausible, and highlights one of the shortcomings in asking participants to explain their answers: namely, that participants may have interpreted the disagreement paradigm as intended even when open-ended follow-up questions suggest that they didn't. My impression is that a substantial number of participants really do conflate descriptive relativism with metaethical relativism, but at present, there is insufficient evidence to decisively support this conclusion.

## S2.5 Evaluative standard ambiguity

The response option "at least one person must be incorrect," is intended to reflect realism. However, this may also be the appropriate response option for cultural relativists, even though cultural relativism is an antirealist position. This is because some versions of the disagreement paradigm suffer from *evaluative standard ambiguity.* This occurs whenever the participant isn't given enough information about the speakers to know whether their claims could be indexing the same normative standard. For instance, consider the claim:

*Abortion is morally wrong.*

Without knowing who is making this moral claim, and what moral standards their claim indexed to, *there is no way in principle for a moral realist to know whether this statement is true or false.* Recall that cultural relativism holds that moral claims are true or false relative to the standards of different cultures. If two members of the same culture disagree about a moral issue, the cultural relativist would still judge that at least one of the people who disagrees must be mistaken.[66] This is because, if people who disagree are members of the same culture, they could both be making claims that refer to the same moral framework. But if they are members of different cultures, they may not be. Such information is not merely relevant, but *necessary* for cultural relativists to judge whether both people could be correct or if at least one must be incorrect. More importantly, *if* two people who disagree are referencing the

---

[66] More generally, so long as the people who disagree are referring to the same set of moral standards, an antirealist would judge that at least one of those people must be incorrect. This is technically true for all forms of relativism, but indexing the truth of moral claims to cultural standards (or the standards of groups more generally) is the most plausible form of actual ambiguity. It would be much stranger for subjectivism. Subjectivists hold that people's moral claims index the moral standards of individuals, typically *themselves.* However, in principle if Alex says, "stealing is wrong" and Sam says, "stealing is not wrong," they could both be referring to the same moral standards, e.g., Alex's or Sam's. If so, then even the subjectivist would judge that at least one of them would have to be incorrect. For instance, if both statements refer to Sam's moral standards, and Sam thinks stealing is wrong, then Alex would be correct and Sam would be incorrect. However, in practice it would be strange for someone to say, "x is wrong," and intend for this to be indexed to someone *else's* moral standards, without surrounding context or additional remarks suggesting that they were doing so. For instance, Sam could say "stealing is wrong," and Alex could respond, "you're lying, Sam. You actually think stealing isn't wrong. So according to you it's true that 'stealing is not wrong.'"

same moral standard, then relativists *would judge that at least one must be incorrect even though this is intended to be the "realist" response.*

Many versions of the disagreement paradigm fail to explicitly specify the cultural backgrounds of the people who disagree. Even when they do, the moral statements themselves don't *explicitly* index one or another of possible moral standards. For instance, when Alex says, "*abortion is morally wrong,*" this statement does not include any explicit content that would allow us to know whether Alex's remark appeals to an unindexed (that is, stance-independent) moral standard, or an indexed moral standard. It *couldn't* do this, or asking the question would be pointless: such a statement would either explicitly convey a realist or antirealist standard, and there'd be no point in asking participants about *their* metaethical standards. Thus, the evaluative standards at play in the moral statements used by the disagreement paradigm must *always* be implicit, since the disagreement paradigm's purpose is to assess what evaluative standard the *participant* will infer that people who disagree are appealing to.

This creates a potentially serious problem for the disagreement paradigm: if a participant is a *cultural relativist* and they are asked to judge a moral disagreement, but they do not know which cultures these people are in, they do not have enough information to judge whether both people can be correct or whether at least one must be incorrect. Such participants are nevertheless presented with a forced choice that *requires* them to resolve the ambiguity, despite the study itself lacking the requisite information they would need to respond in line with their own metaethical position. Thus, unless enough background information about the cultural context in which the relevant moral statements occur, the disagreement paradigm will have an irresolvable ambiguity that could cause a significant proportion of antirealists to choose the "realist" response, which would threaten the validity of the disagreement paradigm altogether.

Some versions of the disagreement paradigm imply that the people who disagree are from the same culture (e.g., Goodwin & Darley, 2008). When this occurs, the ambiguity is resolved in a way

that undermines the validity of the disagreement paradigm: the proper response for a cultural relativist would be to judge that "at least one person must be incorrect," i.e., the correct response is the "realist" response. Since the correct response for both realists and some antirealists would be the same, the measure is no longer capable of determining whether the participant is a realist or antirealist. In other cases, researchers may reference the cultural backgrounds of the people who disagree. This *may* resolve the ambiguity. But it will only do so if this information is salient to the participant when judging the disagreement, and it may only work for *some* forms of relativism. For instance, if the disagreement is third-personal, this may be irrelevant if the participant is an *appraiser relativist*, since their metaethical position depends on the moral standards of whoever is judging the moral disagreement, not the moral standards of those who disagree. It may be possible to resolve these ambiguities with enough clarifications and instructions, but doing so will once again increase the length of the study and increase the cognitive load on participants.

Some findings are consistent with the possibility of evaluative standard ambiguity. Sarkissian et al. (2011) asked participants to judge moral disagreements between a member of their own culture and either (a) a member of the same culture, (b) a member of a very different human culture or (c) a member of an extraterrestrial civilization with very different norms and goals. The greater the cultural distance between the two, the stronger participants disagreed with the statement that "at least one of them must be wrong." Since people were far more likely to judge that they could both be correct when the cultural differences between two people who disagree was explicit, this suggests that many people who judge that one person must be incorrect could be cultural relativists who were either explicitly informed that the two people share the same culture (and thus the same moral standards) or assume that this is the case when their cultural backgrounds are unspecified. If so, then some responses may reliably fail to reflect their metaethical stance using conventional versions of the disagreement paradigm that don't specify the cultural backgrounds of the people who disagree.

Some studies attempt to minimize evaluative standard ambiguity by explicitly informing participants that the people who disagree are from different cultures (e.g. Nichols, 2004), but others make no reference to the cultural backgrounds of the people who disagree prior to participants judging the disagreement (Beebe, 2014; Beebe et al., 2016; Beebe & Sackris, 2016; Heiphetz & Young, 2017; Wright et al., 2013). Some may even amplify evaluative standard ambiguity, since they *do* provide information about the presumptive cultural backgrounds of the participants, but imply that they are members of the *same* culture. For instance, Goodwin and Darley (2008) told participants that "If an event is described, assume that it occurs within the U.S.A.," and asked participants to consider disagreements with real people who participated in previous research (p. 1343). In a second study, Goodwin and Darley (2012) first ask participants about the proportion of US citizens that agree and disagree with the moral claim, then ask them to evaluate a disagreement between themselves and a previous participant. Fisher et al. (2017) likewise ask participants to judge disagreements between themselves and previous participants. None of these cases explicitly state that the person who disagrees with the participant is from the same culture, but they do imply that the other person is from the same country and speaks the same language, which suggests at least some overlap in cultural background.[67] Even when cultures differ, information suggesting that people who disagree are similar to each other or to the participant could plausibly suggest a greater likelihood of a shared moral standard. For instance, Wainryb et al. (2004) asks participants to judge a moral disagreement between other children that "are first graders, *just like you*." (p. 691, emphasis mine).

Even if researchers do provide some background, this may not be adequate, since the fact that two people are from different nations or grew up in different cultures does not ensure that one or the other of the people who disagree have adopted the moral standards of a different culture. Adequately

---

[67] Even when cultures differ, information suggesting that people who disagree are similar to each other or to the participant could plausibly suggest a greater likelihood of a shared moral standard. For instance, Wainryb et al. (2004) asks participants to judge a moral disagreement between other children that "are first graders, *just like you*." (p. 691, emphasis mine).

ensuring cultural relativists are provided with an appropriate response option might require more robust efforts to indicate that the participants adhere to different cultural standards. Even when these cultural details are specified, they would need to be salient to participants and understood in the intended way, which may be a lot to expect of participants. And without adequate context or information that would allow participants to know which standards each person is referring to, participants may interpret the disagreement differently from one another, further undermining its validity. Excessive emphasis on the cultural backgrounds or differences in standards might also bias participants towards relativist response options, since the inclusion of such details could imply their relevance. Participants motivated to give researchers the answers they think they want, or who are trying to get the "correct" answer may be sensitive to specification of cultural background, and opt for the relativist response option even if it does not reflect a genuine stance or commitment towards relativism.

The possibility of specifying the cultural backgrounds of the participants points to yet another difficulty with the disagreement paradigm. The purpose of the disagreement paradigm is to determine whether the *participant* is a realist or antirealist. To do so, participants are expected to apply their own understanding to the meaning of moral statements to the disagreement. This works especially well when participants are asked to adjudicate a disagreement between themselves and someone else. Yet whether the disagreement is between themselves and someone else, or two third parties, another problem still emerges. Suppose I am asked whether two people can both be correct. I am a moral realist myself. However, I am told that each person comes from a very different culture. I infer that each person is attempting to make a claim about what is morally right or wrong according to their culture's moral standards. In other words, I interpret the disagreement to be one between two cultural relativists. Even if I am a moral realist, I should still judge that both people are correct. After all, each of these people would be making a claim about what is true relative to their own standards. The

disagreement paradigm only works if I impose my own understanding of the meaning of moral utterances on others, and render my own judgement about whether there is a stance-independent fact of the matter about the moral issue in question. Yet a belief in stance-independent moral facts is compatible with a recognition that other people are referencing their own stances, i.e., that *others* are speaking as relativists. Imagine, for instance, a realist is told that two people made the following claims:

> Alex: *"Euthanasia is inconsistent with my society's moral standards."*

> Sam: *"Euthanasia is consistent with my society's moral standards."*

If other people are interpreted as making relativist claims, then judging that both people are correct is an appropriate response for the realist. This works even if the disagreement is between the realist and someone else. In such cases, the realist would be asserting that there is a stance-independent fact about the issue in question, while the other person is stating that the moral issue is or isn't consistent with their own moral standards. Again, both statements can be correct, even if the participant is a realist. Excessive efforts to specify the cultural background of the person who disagrees with the participant or the different cultural backgrounds of two third parties could, along with any other efforts that would induce participants to believe the people who disagree are referencing their own standards could result in an extremely confusing question: is the participant supposed to interpret the moral claims in accordance with their *own* metaethical stance, or the metaethical stance *of the person making the claim*? It is not obvious which of the two interpretations is the intended one, resulting in yet another form of evaluative standard ambiguity.

It may seem implausible that this ambiguity would play a significant role in how participants interpret the disagreement paradigm. Even if it did, it poses an additional challenge to the first form of evaluative standard ambiguity, by creating a dilemma: the less information we give about the frame or frames of reference two people who disagree are referring to, the more ambiguity there is about

what standards they are appealing to, while the more we specify, the more we risk the second form of ambiguity influencing participant interpretation.

However, there are also positive reasons to worry that specific versions of the disagreement paradigm are especially prone to the latter unintended interpretation. After presenting participants with a moral disagreement between two other people, Wainryb et al. ask participants:

*Do you think that only one belief is right, or do you think that both beliefs are right?*

The problem with this question centers on the use of "are right." Relativists believe that two conflicting moral views could be correct according to different moral standards, but this does not mean that the relativist believes different moral standards can be correct according *to the relativist's own moral standards*. Suppose, for instance, the participant believes that an action is morally wrong. They are told that two people disagree about the action: one also thinks it is morally wrong, but the other doesn't. The participant is then asked whether both people are correct. Correct according to what standard? Their own standards, or the standards of the participant? If the former, even a realist may judge that both are correct *according to their own standards*, even if they *don't think one or either of them is correct according to their own moral standards*. None of the disagreements Wainryb and colleagues describe indicate that the participants are from a different culture, either. On the contrary, they are given familiar names (e.g., Sarah and Sophie) and told that the children who disagree are "first graders, just like you." If anything, this implies they are members of the same culture as one another and the participant. At best, Wainryb et al.'s findings could at best only distinguish realism from subjectivism, *not* cultural relativism (much less other versions of antirealism).

## S2.6 Abstract norm ambiguity

Moral realists may believe that there are stance-independent facts about whether *specific* actions are morally right or wrong. However, they may also believe that certain *abstract* moral principles are stance-independently true. We can distinguish norms about how, locally, to comply with an abstract moral

principle from the obligation to comply with the principle itself. We may think of the former as *norms of compliance*, and the latter as *norms of obligation*. A norm of compliance is a norm about *how* to comply with an abstract moral rule, while a norm of obligation *is* an abstract moral rule itself. Realists can (and typically do) recognize that there are multiple means of complying with some abstract moral rules. For instance, they may believe that all people have a moral duty to "show respect for the dead." Yet, *how* one shows respect for the dead will depend on the local customs and norms of one's community. Some cultures show respect for the dead through burial, others through cremation, and still others through ritual endocannibalism. Each of these actions may be immoral if performed in a different community, insofar as it provoked outrage and was perceived as an act of desecration, while that very same act would be seen as morally obligatory in a different cultural context. In other words, different practices may be equally consistent with the same stance-independent moral facts.

As a result, a realist may believe that if two people hold contrary moral views, they could both be correct since both positions could be consistent with the same abstract moral facts, *not* because each is correct relative to a different moral standard. If so, this could cause many moral realists to judge that if two people disagree about a moral issue, that they can *both* be correct. Yet, rather than this reflecting a form of moral antirealism, it would simply reflect the view that there is more than one way to conform to the same (stance-independently true) moral principle.

Such a recognition may play a role in how participants interpret the disagreement paradigm. If so, participants who endorsed realism could nevertheless judge that if two people disagree about a moral issue, they could both be correct. This is because participants may understand a disagreement between two people to reflect conflicting norms of *compliance*, rather than a disagreement about a norm of *obligation*.

If the judgment that two people can both be correct for this reason cannot be distinguished from people who judge that both are correct relative to different moral standards, this represents yet

another way that people's responses fail to consistently reflect the relevant metaethical position, which we may call *abstract norm ambiguity*. It would take additional research to assess how often this ambiguity influences interpretations of the disagreement paradigm. If it does occur often enough to raise methodological concerns, however, mitigating it may require additional instructions that lengthen and complicate research.

**S2.7 Misattributing source of disagreement**

In order for the disagreement paradigm to be valid, participants must attribute the difference in moral belief to a *fundamental moral disagreement*. A fundamental moral disagreement cannot be attributed to a difference in nonmoral beliefs or attitudes, but must instead be due to a difference in moral values themselves. In other words, even when people are referring to exactly the same situation and are not subject to errors in judgment and reasoning about all relevant nonmoral facts, they still disagree about what is morally right or wrong because they have different beliefs about what the moral norms *themselves* are. [68] Not all moral disagreements are fundamental moral disagreements. There are many reasons why two people might find themselves on competing ends of a moral dispute that cannot be attributed to differences in their moral standards:

For instance, people could disagree about the nonmoral facts. Two people could both agree that we should favor whichever policy would minimize human suffering, but disagree about which policy would in fact do so. Alex may think we should raise taxes, because doing so would provide us with more tax revenue, which could be used to fund welfare programs that would minimize suffering.

---

[68] This can include both abstract moral norms and the application of those rules to specific situations. With respect to the former, one person might believe that we have private property rights, while another person may simply deny that outright we have private property rights. With respect to the latter, two people could both agree that lying is sometimes permissible, but disagree about when it is permissible. What is necessary for the latter to be a fundamental moral disagreement is that the difference in application must result from a difference in what they believe the moral norms themselves dictate, *not* some nonmoral consideration. For instance, one person may believe it is permissible to lie whenever it would minimize harm, while the other may believe it is permissible to lie only when some overriding moral duty takes precedence (whether or not, and independent of, whether acting in accordance with this overriding moral duty minimize harm).

But Sam might believe that raising taxes would drive business away, which would reduce overall tax revenue, which would reduce the amount of money available to fund welfare programs. Both want to minimize suffering, and both even want to fund the same welfare program; they just disagree about the best way to fund it. This is not a fundamental moral disagreement, because they share the same moral goals. They just disagree on how to realize those goals. There are many other ways people could reach different moral conclusions because they have different nonmoral beliefs. Disputes about the moral status of abortion or using animal products could result from differences in the amount of suffering these actions cause, rather than the moral value of suffering itself. Differences about gun control or the death penalty may result from differences in the impact these policies have on society, and so on.

Other moral disagreements may be due to far more mundane factors. People could simply misunderstand one another, or be thinking of different situations, or be imprecise in their language, or use qualifiers that are not meant literally, or speak in ways that imply universal claims even where no such claim is intended. For instance, when someone says "lying is wrong," they may be thinking only of prototypical cases where it would be fairly uncontroversial that lying is wrong, e.g., when it is done to further the interests of the liar at other people's expense, but are not thinking of atypical cases where it may be justified (or required) to lie, e.g., to refuse to provide the whereabouts of someone to an enraged psychopath. Ordinary discourse is often highly underspecified, context-sensitive, and prone to minor and often major misunderstandings and miscommunication. A great deal of moral disagreement may result from such misunderstandings, and in many cases does not reflect genuine moral disagreement at all. If Alex says that stealing in situation X is *wrong*, and Sam says that stealing in situation Y is *not wrong*, they are not really disagreeing; they are just talking past one another.

These unintended interpretations could in turn lead participants to judge that two people could both be correct simply because the participant doesn't know the details of the imagined specific

circumstance the disagreement is about, or what each individual has in mind in expressing their moral stance. The explanations many participants offered for their answers suggest just this unintended interpretation. Wright, Grandjean, and McWhite (2013) note that participants who judged that both they and another person could be correct if they disagreed about a moral claim:

> [...] participants frequently pointed out the importance of a *situational influence*—for example, that fact that circumstances could influence whether a given action was right or wrong (e.g., "it depends on the seriousness of the situation," "I don't know the situation," "reasons that make it okay can come up"). (pp. 15-16)

These participants seem to be responding in one of two ways. The response "I don't know the situation," seems to implicitly assume that the disagreement is not about the moral status of a general type of moral act or principle, but is instead about the moral status of a *particular* moral act. Since they don't know the details of the situation, which may be relevant to whether the act is morally permissible, they cannot judge definitively whether the act is permissible or not. This view is fully compatible with realism. The best interpretation of how this person views the conclusion that they "could both be correct" is that it is a concession that either one or the other of them could turn out to be correct (not both), but since the details are unknown, they cannot say which.

When participants are asked to explain why people disagree about a moral issue, they frequently point to nonmoral differences (See **Chapter 4**, **Study 1**). For instance, when asked to explain why someone who disagreed with a participant about the statement:

"Opening gunfire on a crowded city street is a morally bad action"

One participant responded:

*The other person could be thinking about certain circumstances like the protection of others if there was a threat.*

In other words, the reason why someone disagreed was because they were thinking of a situation that was *different* from the situation that the participant was thinking of. If they judged that both they and the other person could be correct, this could merely reflect that each person is correct about a *different*

moral issue. Yet, the disagreement paradigm only serves as a valid measure of a person's metaethical stance or commitment *if* the participant interprets the disagreement to refer to the *same* moral issue. As a result, all instances in which participants attribute the source of the disagreement to nonmoral differences cannot serve as valid measures of realism and antirealism. In the absence of additional instructions, participants do seem to frequently attribute disagreements to nonmoral differences. Studies could include additional instructions to mitigate these interpretations, and researchers could include additional questions to assess whether people attributed the source of disagreement to nonmoral differences. As always, doing so would require making studies longer and more complicated.

The second type of response does not presuppose the disagreement concerns any specific act. Instead, it involves the recognition that the same general act, e.g., "killing" or "stealing," may be morally right or wrong depending on the circumstances (Wright, Grandjean, & McWhite, 2013). For instance, the participant may believe that abortion is permissible when the mother's life is threatened, but not otherwise. If the participant is required to judge a disagreement between someone who claims abortion is permissible, and another who claims it isn't, the participant may conclude that there is no single correct answer to this question without knowing the additional detail of whether the mother's life is threatened. The judgment that both people could be correct may be the best way to capture the underspecificity of the disagreement, even though the participant is a realist.

The potential for underspecificity and a sensitivity to context to influence how participants interpret what they are asked is exacerbated by the realist response option, which emphasizes that only one side of the disagreement can be correct. Even if participants would agree that there is a stance-independent fact of the matter about the truth of a well-specified norm of obligation, they are not given enough information to know whether the people who disagree do so as a result of a fundamental difference in moral values, or because they disagree about how to comply with the same abstract moral rule.

There is some indication that participants who favor a relativist response do so because they interpret the source of disagreement in this way. Goodwin and Darley (2008) told participants that another person disagreed with them, and collected data on what participants thought was the source of the disagreement between them and another person. Many participants thought the disagreement could be due to each person imagining a different context, one in which the action in question would be acceptable and one in which it wouldn't be. For instance, some participants were told that someone else disagreed with the following claim that "Opening gunfire on a crowded city street is a morally bad action."

> **Response #1**: *Depends on the context (?) Possibly, if the streets are full of rapists trying to kill you. I know it's a stretch.*

> **Response #2**: *A difference in perception of a situation in which gunfire was opened on a crowded city street. I was thinking gunfire from terrorists/ criminals; other person may have thought gunfire from police officers to catch a criminal.*

> **Response #3** *The other person could be thinking about certain circumstances like the protection of others if there was a threat.*

Similar explanations were offered for other moral disagreements as well, but I will not belabor the point with additional examples. Across virtually every moral issue tested, participants regularly appeal to the possibility that people who disagree about a moral issue do so because they are imagining different situations.[69] When participants judge that both people can be correct in these cases, it cannot

---

[69] Some participants attribute the source of disagreement to a difference in how the people who disagree conceive of morality itself:

**Response #1**: *They have different thoughts about what constitutes a "morally bad" action.*
**Response #2**: *They define morality differently from me.*

It is hard to tell whether these participants are expressing a relativist stance or something else. While these participants could think that both people are each correctly referencing their own moral standards, it is unclear whether the participant *themselves* judges both positions to be correct; in other words, it is not enough to believe that two people who disagree do so because they disagree about what morality is; one must also believe that both people are making genuine moral claims and that both are correct because one's own conception of moral truth is relativistic. Participants may instead believe something like "Each person is using the word 'morality' to mean something in particular." But this is consistent with both realism and antirealism. After all, both realists and antirealists could recognize that people have different stances on how to define morality.

tell us whether that participant is thinking in realist or antirealist terms. The disagreement paradigm's validity *requires* that participants interpret the disagreement to concern *the exact same moral issue*. Otherwise, each person could simply be correct about a different moral situation. Judging that both people are or could be correct in such cases would not indicate relativism or any other form of antirealism, it would simply involve a recognition that some abstract act like "stealing" or "killing" may be morally acceptable in some circumstances and not others, which would simply be a normative judgment fully consistent with all forms of realism and antirealism.

Goodwin and Darley (2008) took the prescient step of asking their participants what they thought the source of the disagreement between themselves and the other person could be. However, I reanalyzed this data, and found that *most* participants pointed to reasons other than fundamental disagreements (Bush & Moss, 2020). Many of the participants in Goodwin and Darley's study attributed the disagreement to far more prosaic causes than fundamental differences in moral values. Many attributed the moral disagreement to the other person thinking of a different circumstance than the participant:

> **Response #1:** *Depends on the context (?) Possibly, if the streets are full of rapists trying to kill you. I know it's a stretch.*

> **Response #2:** *A difference in perception of a situation in which gunfire was opened on a crowded city street. I was thinking gunfire from terrorists/ criminals; other person may have thought gunfire from police officers to catch a criminal.*

> **Response #3:** *The other person could be thinking about certain circumstances like the protection of others if there was a threat.*

My goal here is not to quantify how often such interpretations occur, but it is worth noting that they are frequent enough all on their own to chip away at the validity of the disagreement paradigm. But taken in isolation, a handful of open-response questions don't mean much. Studies can easily absorb the loss of a handful of participants potentially interpreting what researchers are asking, provided most interpret the question as intended. Yet quantifying comprehension rates reveals that very few

participants unambiguously interpret the disagreement paradigm in the way researchers intend. Here, I want to make a more general point. The disagreement paradigm may be well-suited for use among philosophers. Whether by training or disposition, philosophers can suspend extraneous considerations when entertaining questions like the ones posed by the disagreement paradigm.

Consider the trolley problem. Philosophers recognize that their options are constrained to those provided; they cannot jump on the tracks themselves or call for help. Yet ordinary people often propose such measures, and have to be told that these options aren't available. Philosophers accept that unless otherwise specified, the people on the tracks are generic strangers whose moral worth is ostensibly equal to any other person. Yet ordinary people often ask *who* the people on the tracks are. They have to be told that it doesn't matter, that they are strangers, and so on; even then, people push back: But what if one of them is a doctor? What if one of them is a criminal? Ordinary people present a litany of considerations that philosophers have to patiently (or not so patiently) bat away: will the police find out? Why isn't anyone else there? Why can't I shout at the people on the tracks? Can I live with the guilt? What about the families of the victims? And one by one, the philosopher has to dispense with each one of these concerns, often by adding various details, e.g., "It's a one-off event. There are no witnesses and no legal repercussions," "No, you can't shout. Because they can't hear you. Why? Well...uh, because they have ear protection on. No, you can't wave, they have their backs turned. No, it doesn't matter that you are on the track team in real life. Imagine in this scenario you can't run that fast."

Philosophers often find themselves responding to a fusillade of questions like these, questions that, to the philosopher, *miss the point*. People just don't seem to understand that this is a thought experiment that is conceptually constrained by design in order to focus on one specific consideration. This is not a natural way to think. The kinds of questions ordinary people pose may seem foolish, and in a certain respect, perhaps they are. On the other hand, these are perfectly reasonable questions for

one simple reason: for ordinary people, *context is everything.* Our everyday judgments concern *actual events*, not magical thought experiments that isolate all the variables that would ordinarily be present. People's judgments are tailored for operating in actual contexts where the people are, why they are there, the social ramifications of intervening in various ways, and so on all matter.

Just the same, when a philosopher is presented with an alleged moral disagreement between themselves and someone else, they are attuned to recognizing that this is a thought experiment, and their facility for construing the situation in a way that would allow them to respond appropriately kicks in, leading them to infer that:

> *(i) This is an idealized, imaginary situation, where someone really does disagree with them, and it is not due misunderstanding or differences in nonmoral belief.*

> *(ii) They are both referring to the same scenario, not different scenarios.*

> *(iii) The issue is sufficiently well-specified that whether it is correct or incorrect uniformly applies to whatever circumstance or circumstances both sides of the disagreement are referring to. It is not the case that one person could be correct about some contexts while the other is correct about others, because their judgments quantify over the exact same circumstances.*

> *(iv) The other person holds a sincere moral belief that conflicts with their own. They aren't just a psychopath who lacks moral beliefs at all.*

Responding to the disagreement paradigm appropriately *requires* this entire array of sophisticated inferences in order to function as intended. This is an incredibly tall order, and yet ordinary people are expected to reliably interpret every question having made (consciously or not) all of these inferences.

Yet there is little evidence to suggest that they are up to the task. A casual glance at the reasons people offer for why another person might disagree with them overwhelmingly attest to the fact that people are considering what an *actual* disagreement with another person would be like, and are offering reasons for why two people might report different moral conclusions. And actual disagreements aren't the idealized kinds of disagreements philosophers typically entertain. Actual disagreements can be messy. One or both sides can be confused or talk past one another. And when we disagree with others,

our primary interest may be what side to take on the issue in question, or evaluating the character of the other person, or engaging in a myriad of social goals that are orthogonal to assessing the metaethical status of the conflict between our claims. These are just the kinds of responses people offered. Rather than treating the other person as an epistemic and moral peer who sincerely endorsed an alternative moral standard, some participants simply suggested that the other person disagreed because they were immoral or insane:

>**Response #1**: *The other person is greedy and inconsiderate.*

>**Response #2**: *The person is immoral.*

>**Response #3**: *Greed, laziness, lack of respect for social institutions.*

These participants may or may not be moral realists, but we cannot tell from these remarks. Since they attribute the cause of the disagreement to the other person simply being a bad person, it is not clear whether they think that person has a genuinely distinct moral stance, or just doesn't care about morality at all. If the latter, then their response to the disagreement paradigm would not be a valid measure of their metaethical beliefs. In addition to suggesting that the other person is immoral, some participants also suggest that the person who disagrees may be unaware of the relevant moral facts, could be insane or psychologically damaged in some way, or may simply not even in the business of expressing a contrary moral stance:

>**Response #4**: *The other person must be both ignorant, immoral and insane.*

>**Response #5**: *I don't even understand how they could have their opinion, unless they suffered psychological abnormalities or are morally depraved.*

>**Response #6**: *The person is amoral, and does not realize (or care about) the possible consequences of his/her actions.*

If the participant attributes the source of disagreement to the other person being "ignorant," or "does not realize [...] the possible consequences of his/her actions," then this could be because that person is unaware of the moral facts, rather than because that person has different moral values. If the other

person is "amoral" or "insane" or suffers from psychological abnormalities, it is unclear whether such a person is expressing a legitimately contrary moral stance, rather than just failing to express a moral stance at all. Still others offer psychological explanations or propose that the other person is in denial:

> **Response #7**: *The other person may have personal experiences with this and does not want to admit to taking morally wrong actions.*

> **Response #8**: *Maybe he/she has discriminated and feels that he/she is still a moral person. So he has to call his acts moral.*

One participant even suggested that the other person simply misunderstood the question:

> **Response #9**: *Other individual misread the question.*

I address how often each of these kinds of responses occurred among participants in **Chapter 4**, but my goal here is not to make quantitative claims about how often such unintended interpretations occur (though it turns out that they are far more common than instances where participants clearly interpreted metaethical questions as intended). Rather, it is to highlight how ordinary people engage with questions about moral disagreements. It is clear that ordinary people do not engage with the disagreement paradigm the way a philosopher is trained to. When considering why people disagree, they are open to all the *actual* reasons people in everyday life may express contrary positions.

Philosophers, on the other hand, are trained to limit their concern to an extremely narrow and peculiar consideration: They are expected to interpret the disagreement paradigm to describe a situation in which two people appear to be offering utterances that have the prima facie appearance of logically contradictory assertions (i.e., "X" vs. "Not-X"). They are then expected to recognize that both claims are token instances of some circumscribed *moral domain*, that is, both are instances of a particular normative domain, such that any judgments about the metanormative characteristics of assertions reflect on the metanormative properties of the domain itself. Finally, they must judge whether the two (seemingly) conflicting moral claims can both be correct because the proper understanding of moral claims (or at least the moral claim in question) is that they contain implicit

indexicalizations the truth of which can vary depending on what standard each speaker is referring to, or whether they cannot both be correct because moral claims (or at least the moral claim in question) are not indexicalized in this way.

This is an incredibly sophisticated interpretation of the question being asked, and it requires the person engaging with the question to suspend every irrelevant consideration that might interfere with this very specific interpretation: they cannot attribute the disagreement to nonmoral beliefs, insanity, confusion, ignorance, misunderstanding, amorality, insincerity, different conceptions of what the concept of morality itself entails, or referencing a different scenario or context than the other person. No, they must understand both sides of the disagreement to perfectly understand one another and the issue in question in exactly the same way, to suffer from no psychological deficiencies or abnormalities, and to be completely sincere. And they are expected to understand the disagreement in this way even when one side of the disagreement allegedly believes that issues as manifestly deplorable as mass shootings, bank robbery, and racism are morally acceptable. This, I submit, is *ridiculous*. For ordinary people to understand the disagreement paradigm as intended, they would have to think like well-trained philosophers, when the whole point of engaging them as participants is *because they don't think like philosophers*.

Researchers studying folk metaethics, or many other topics for that matter, fail to appreciate that ordinary thought is well-calibrated for dealing with real-world circumstances that involve actual people. And actual people can be ignorant, immoral, confused, or insane. There is no reason to think that ordinary people would readily suspend such considerations without researchers even asking them to. And nobody conducting the disagreement paradigm has ever asked people to suspend these considerations. Even if researchers did ask people to suspend such judgments, it is not clear that they would be successful. The kind of suspended disbelief, counterfactual thinking, ability to suppress the distorting influence of emotional and cognitive biases, and the capacity to adequately simulate and

engage with thought experiments in ways that cordon off philosophically irrelevant considerations, may require substantial training that ordinary people simply don't have. And as much as they may aspire to think like Vulcans, even philosophers routinely struggle to do so.

The disagreement paradigm requires that participants understand disagreements in a particular way. Yet there is little theoretical rationale for presuming that they would, and at present, what little evidence we have about how people understand the moral disagreements suggests that they interpret them in a variety of ways, few of which are consistent with the interpretation necessary for the disagreement paradigm to be valid.

## S2.8 Domain classification inconsistency[70]

Most versions of the disagreement paradigm rely on researchers' own *a priori* classification of issues as moral and nonmoral. Wright, Grandjean, and McWhite (2013) found that when participants were asked to classify issues as moral or nonmoral, participants (a) exhibited considerable variation in how they classified moral issues, (b) systematically differed from researchers, e.g., the vast majority of participants did not consider donating to charity to be a moral issue. However, they found that even when participants' own classifications were taken into account, they still exhibited about the same degree of intrapersonal variation in realist and antirealist responses to moral issues. This demonstrates that differences in domain classification cannot explain away evidence of metaethical pluralism.[71] These findings suggest that evidence of pluralism cannot be explained away as a result of differences in what people regard as moral or nonmoral issues.

---

[70] Pölzler (2018c) discusses this issue in section 3.8 (pp. 70-71) and recommends including a domain classification task. My attitude towards this solution is decidedly less optimistic.

[71] It could have turned out that if I only examined disagreements that participants classified as moral, that they would show a uniform pattern towards all such issues, e.g., consistently realist or antirealist. This is not what WGM found for the issues used in their study. See also Wright (2018) and Wright, McWhite, & Grandjean (2014).

However, my concern is not with whether the findings appear to show pluralism, but with the validity of the measures. If our goal is to determine whether the participant takes a realist or antirealist stance towards a particular moral issue, or the moral domain as a whole, systematic variation in what participants consider to be moral issues will invariably introduce noise into any attempt to measure rates of moral realism or antirealism towards particular issues or in aggregate. As such, *a priori* classification will always suffer from potentially significant and systematic imprecision.

One potential solution would be to always ask participants to classify issues as moral or nonmoral themselves, then estimate rates of realism and antirealism towards the moral domain based on participant's own classifications. However, it is unclear whether this would be adequate. First, doing so would further complicate studies that employ the disagreement paradigm, since it would require participants to perform yet another task, further increasing the cognitive demands and length of a study.[72]

Second, even if we include a classification task alongside the disagreement paradigm, participants within any given population may still show significant intrapersonal variation in how they classify moral issues. At the same time, they may (a) interpret the classification task differently from one another and (b) interpret the classification task in unintended ways that differ from researcher intent. Recognizing risks like these, WGM used an open response method much like what I present in **Chapter 4**, and found that participants generally offered fairly sensible responses that tracked what they intended with the classification scheme. Setting aside whether their evaluation of their own data is more or less accurate (and we may reasonably question whether it is), *even if* participants *generally* interpreted the classification scheme as WMD intended, they didn't *uniformly* do so. Thus, such measures still introduce a degree of noise. In addition, we don't know whether the same would hold

---

[72] Taken in isolation, this may be a minor issue, but coupled with the many other measures, instructions, and other additions that might be needed to ensure the validity of the disagreement paradigm, such complications contribute to the overall infeasibility of such efforts.

for other populations, once again making efforts to employ the same measures in different populations challenging at best. Given growing evidence that different populations conceptualize the moral domain differently (Levine et al., 2021; Machery, 2018), as well as differ in what they consider to be moral and nonmoral issues, the chances that we'd encounter such problems are very high.

Third, relying on participants' own classifications complicates comparisons between participants. One participant may consider issues {A, B, C} to be moral issues, while another considers issues {B, C, D, E} to be moral issues. Differences between participants would no longer be based on judgments about the same issues. Likewise, it would be difficult to interpret the total proportion of participants who favor a realist or antirealist view towards a particular issue since judgments of realism or antirealism towards a particular issue may not be orthogonal to domain classification. That is, participants who favor a realist or antirealist response may *systematically* differ in whether they consider a particular issue to be moral or nonmoral. WGM are aware of this, but report that domain classification appeared orthogonal to such judgments: "for at least *some* of the issues" (p. 7). Yet, orthogonality may *itself* vary from item to item, and, in any case, high orthogonality is not the same thing as *perfect* orthogonality: to the extent that classification isn't orthogonal to metaethical judgments, classification tasks reveal yet another source of interpretive variation that contributes to the noisiness of the disagreement paradigm.

In short, the problem of domain classification variation draws attention to another source of interpretive variation that threatens the validity of the disagreement paradigm. Augmenting the disagreement paradigm with a classification task necessarily carries costs by increasing the length and cognitive demand. Furthermore, the inclusion of such tasks may be insufficient to fully mitigate the imprecision caused by classificatory variation and may be incapable of resolving potential problems related to measurement invariance. The inclusion of a classification task is itself saddled with its own risk of interpretive variation and unintended interpretation. Even if both are low, they aren't zero,

contributing further noise to our measures. Finally, domain classification may not be perfectly

orthogonal to metaethical judgments, further undermining the value of the disagreement paradigm.

**S2.9 Presumption of correspondence theory of truth**

The concepts of *true*, *false*, *correct*, *incorrect*, *mistaken*, *right*, and so on are central to the disagreement

paradigm. Yet as Pölzler (2018a; 2018b) observes, the validity of measures of folk moral realism and

antirealism rely on the presumption that participants understand these concepts in line with, roughly,

a *correspondence theory of truth*:

> [I]t is important to note [...] that realists and anti-realists disagree about the existence of
> objective moral truths in a very specific sense of moral truth. They affirm or deny these truths
> in a *correspondence-theoretic* sense, according to which for a moral sentence to be true is for it to
> represent a moral fact [...] This means that in order for truth-based measures of moral realism
> to be valid subjects would have to understand moral truth (correctness, rightness, etc.) in this
> correspondence-theoretic sense as well. (p. 662)

If participants do not share a conception of truth such that to say that there are moral facts is to

describe some feature of the world, then it is unclear whether their judging that *two people can both be*

*correct* or *whether at least one must be incorrect* reflect antirealism and realism, respectively. As Pölzler

(2018b) notes, ordinary people could instead have a deflationary (Blackburn, 2000; Gibbard, 2003) or

coherentist (Dorsey, 2006) view of truth.[73]

For instance, according to deflationary views of truth, to say that a statement is true is simply

to affirm the statement in question, i.e., to say that "*it is true* that murder is wrong" is reducible to

simply asserting "murder is wrong," and nothing further can be said about the truth of such claims

(Stoljar & Damnjanovic, 2007). This may seem like a strange view, but we cannot dismiss from the

armchair the possibility that, in at least some domains (such as morality), people do not regard moral

claims as attempts to describe the world, but to instead simply affirm a particular moral stance. This

---

[73] The references for deflationary and coherentist views of truth are those provided by Pölzler (2018b).

may be less plausible for claims about typical claims about physical objects (e.g., claims about whether trees exist), but it is unclear whether people are similarly inclined to regard moral claims as attempts to represent facts that in some way correspond to the world.

There is little research on whether ordinary people have a truth-correspondence notion of moral truth. The only attempt to empirically assess whether ordinary people endorse a correspondence theory of truth with respect to moral claims is briefly described in a footnote in Pölzler and Wright (2020a). They made two attempts, both of which were inconclusive. The first presented participants with a description of correspondence and deflationary views of truth and asked them to choose between them, but they report that this "sparked a lot of confusion, as evidenced by participants' verbal explanations" (p. 19, footnote 15). In their second attempt, participants were asked whether they endorsed a set of claims of the form "X is wrong" and a corresponding set of claims of the form "The sentence 'X is wrong' is true." They found no significant difference between the conditions, which is consistent with both deflationary and non-deflationary views of truth. This is consistent with the pair of studies described in the main text, which did not support the notion that ordinary people readily and uniformly endorse the correspondence theory of truth (Barnard & Ulatowski, 2013; Reuter & Brun, 2021). Quite the contrary, both found little support for the notion that most people endorse a correspondence theory of truth. In fact, Reuter and Brun occasionally found a substantial majority of participants favored coherentist views of truth:

> Perhaps surprisingly or even shockingly – at least from a philosopher's perspective – a substantial number of participants (in some experiments up to 70%) responded in line with the predictions of the coherence account. These results suggest that, even within the empirical domain, 'true' is not used in a uniform way in everyday discourse. (p. 2)

These are not encouraging results for studies that require the correspondence theory. Of course, with only a handful of studies, such results are far from conclusive. Yet with a few points *against* the assumption that everyone endorses the correspondence theory, researchers are not in a position to

just assume most people endorse it anyway. Pending additional research, it is not possible to assess whether folk conceptions of moral truth differ in ways that threaten the validity of the disagreement paradigm or of folk metaethical research in general. However, just as people may hold no determinate stances or commitments about moral realism and antirealism, they may also have no determinate stances or commitments about the nature of moral truth. Indeterminacy about folk conceptions of moral truth would threaten the disagreement paradigm and the validity of folk metaethical research in general. If moral realism and antirealism presuppose a determinate conception of moral truth (i.e., the truth-correspondence theory), then a person could not have a determinate metaethical stance unless they *also* had a determinate stance about truth.

Even if folk conceptions of moral truth are determinate, there could still be significant intrapersonal and interpersonal variability in how people conceive of truth, just as there allegedly is with metaethical claims themselves. Pölzler (2018b) acknowledges this possibility himself, noting that research on folk conceptions of moral truth could reveal that people adopt a correspondence view towards some moral issues but a deflationary view towards others (p. 664, footnote 34).[74] If ordinary people turned out to be moral truth pluralists, this could render attempts to measure folk metaethical beliefs even more difficult than they already are. This is because, if people are pluralists, we could not simply determine whether they hold a correspondence or deflationary view of moral truth in general. Instead, we would need to evaluate how participants think about moral truth *for every moral issue they are presented with*. One participant might have a deflationary view of the truth of claims about abortion but a correspondence view of claims about shooting pedestrians, while another participant favors the opposite, while still a third may favor a deflationary view of both. As Pölzler (2018b) observes, "such a survey design would certainly be extensive and complex. Researchers would have to weigh the

---

[74] There could also be stable interpersonal variability, i.e., individual differences between people, with some people tending more towards correspondence theory and others tending more towards a deflationary view of truth.

resulting benefits in terms of construct validity against potential pragmatic costs" (p. 664). Yet Pölzler is not optimistic about the prospects of testing folk conceptions of truth. After speculating about various ways we might attempt to test whether people endorse a correspondence-theoretic conception of moral truth, Pölzler notes that attempts to articulate the distinction in ways that ordinary people could understand and would allow us to reliably distinguish which position they endorse would be exceptionally difficult to devise, leading Pölzler to the pessimistic conclusion that there "seems to be no way around testing the understanding of *some* of subjects' philosophical concepts" (p. 663). I am sympathetic to this concern. It may be exceptionally difficult to determine whether people endorse a correspondence, deflationary, or some other conception of moral truth. While this question may prove empirically tractable, it will not be easy to frame questions in a way where we could be confident participants are understanding the relevant concepts in the way philosophers do. Even if we can, it has not been done yet. As a result, the validity of the disagreement paradigm rests on the questionable hope that enough ordinary people endorse a correspondence theory of truth that the disagreement paradigm (and other measures) remain valid.

Pölzler and Wright's insights that the meaningfulness of the disagreement paradigm hinges on assumptions about the *other* (non-metaethical) philosophical stances and commitments of participants illustrates a deeper problem with research on folk philosophy: many philosophical positions only make sense against the background of *other* philosophical positions. Philosophical positions do not typically exist in isolation, but are nested within a web of other philosophical stances and commitments. Any particular set of stances and commitments play an integral role in how a philosopher understands *other* stances and commitments. In other words, philosophical positions are meaningful only in a holistic, interdependent sense. The disagreement paradigm is only a valid measure of *everyone's* metaethical stances/commitments if we assume *everyone* shares the same, particular theory of truth. Yet we have

no good evidence that they do. Indeed, people may have no determinate stances or commitments with respect to theories of truth, *either*.

Given the interconnectedness of different philosophical stances and commitments, this possibility points to a much broader philosophical worry with research in folk metaethics: any attempt to study a particular phenomenon using a particular set of materials that ignores or fails to control for variation in other possible philosophical stances or commitments the content of which would be relevant to how a participant interprets the questions in the study may suffer from a large degree of nearly-undetectable levels of interpretative variation. And since participants may have no determinate stances or commitments with respect to some of these other philosophical issues, studies that presume that they do may yield results of questionable value. After all, if ordinary people don't understand truth in the correspondence-theoretic sense, then judging that only person may not indicate a stance or commitment towards realism at all.

In short, if ordinary people's philosophical stances and commitments can only be understood *holistically*, then it may be extremely difficult to study philosophical positions in isolation. If so, research on folk philosophy that treats philosophical stances or commitments towards a specific philosophical issue as autonomous and disconnected from the rest of a person's philosophical stances or commitments may be systematically flawed. Incidentally, virtually all research does treat folk philosophy this way.

## S2.10 Signaling & reputational concerns

In practice, questions about realism and antirealism are entangled with normative moral considerations to such an extent that it may be difficult or impossible to fully isolate questions about the nature of morality from considerations about the substantive normative content of people's moral beliefs. This provides fertile ground for the potential role of signaling and reputational concerns to play a considerable role in how participants respond to the disagreement paradigm and other questions about

metaethics. For instance, to express a realist stance may be perceived to signal a rigid, close-minded, or intolerant attitude towards people who hold contrary moral positions. Conversely, an antirealist attitude may express a more tolerant and open-minded attitude, yet it could also convey ambivalence or indifference towards a particular moral issue.

The degree to which people associate realism and antirealism with non-metaethical attitudes, traits, and behaviors is an open empirical question. There is already some empirical evidence that ordinary people associate relativism with tolerance and openness (Collier-Spruel et al., 2019, Feltz & Cokely, 2008; Wainryb et al., 2004), and realism with intolerance, rigidity, and close-mindedness (Goodwin & Darley, 2012). However, *if* people are sensitive to what realist and antirealist response options to the disagreement paradigm may signal about their character, such that responding in particular ways may suggest socially desirable or undesirable qualities, they may be motivated to select a response in order to signal that quality, rather than that response accurately reflecting their metaethical position. These findings are amply corroborated by the findings I present in **Chapter 4**: participants routinely associate expressions of realism with absolutism, dogmatism, close-mindedness, and the kinds of things a religious person might say, the latter often expressed with implied disapproval.

While signaling and reputational concerns may be relevant to many psychological studies, they may be of greater potential relevance with respect to metaethical considerations, While the results of my open response data hint at this possibility, such findings are hardly conclusive. However, there are at least some reasons to expect future findings to support this connection. For instance, may associate expressions of realism and antirealism with various non-metaethical beliefs, values, or ideological positions. For instance, realism may be more associated with religiosity, conservatism, and a rigid and inflexible attitude, while relativism may be more associated with secularism, progressive values, and

an open-minded and tolerant attitude. Insofar as people are sensitive to these associations, this could

influence how they respond to questions that are *ostensibly* about metaethics.

I also introduce *normative entanglement*, the conflation between metaethics and normative ethics

(see **Supplement 3**). While this conflation is often used as a rhetorical ploy (even if unintentionally)

in academic contexts, its use may not be limited to debates between academic philosophers. Rather,

the conflation may reflect a deeper entanglement between metaethics and normative ethics. It is

possible that when ordinary people are introduced to antirealist positions, such as relativism, they

imagine that these positions have substantive *normative* implications, for good or ill. For instance, there

may be instances in which people infer that antirealist positions entail a disregard for the welfare of

others, or a psychopathic disregard for the welfare of others. As an antirealist, it's not uncommon for

me to encounter people exclaim with outrage, incredulity, or scorn: *"So you think it's totally okay to torture*

*babies just for fun?!"* This is an absurd reaction. The answer is that, no, of course I don't. But even

professional philosophers often posture in this way towards antirealism, either because they want to

win cheap rhetorical points, or because they *also* misunderstand antirealism. This has led me to suspect

there is a deep, but confused association between realism and normative ethics, where realists imagine

that antirealist conceptions of morality are somehow inadequate in normative or practical terms, and

that we don't "really" think anything is good or bad, don't "really" think torturing babies is wrong,

and so on.

I've often countered this by pointing out that we don't speak this way about food preferences.

Think about your favorite things to eat, favorite songs, and so on. Do you think they are *objectively* (i.e.,

stance-independently) good? That is, are you a *normative realist* about food and music? Some readers

may endorse such a view, though I suspect many won't, and will instead endorse an antirealist position,

e.g., subjectivism, believing instead that there is no stance-independent fact of the matter about what

food or music is good. Yet would you say that you don't *really* think your favorite food and music is

good? Are your preferences somehow fake or illusory or not "real"? I don't think so; that seems like a strange way to think about our preferences.

Realism doesn't have a claim on things being "real," or meaningful or important to us. This is not to say the antirealist's opposition to torturing babies is as arbitrary or trivial as our favorite pizza toppings, but rather that, if we don't think of our values in the way realists do, this doesn't render then *meaningless* or *trivial* or *unimportant* or *not real* in some practically meaningful way, at least not unless realists have a very good argument for why this must be the case. But they're not entitled to help themselves to this being true without argument, since that would require them to question-beggingly presume that their own notion of meaning, value, or things being "real," were correct, which *is the very thing we antirealists are disputing*. From an antirealist perspective, realist conceptions of value aren't real, after all! *That's the whole point!*

Yet for some reason many realists seem to struggle to adopt an evaluatively neutral stance when engaging with antirealists. This is, unfortunately, merely a failure of imagination. It is unfortunate not only because it results in realists struggling to understand the antirealist point of view, but because it may be one of the primary reasons why realists turn up their noses in disgust at antirealist positions. From their point of view, at best, we endorse nothing but paltry and insubstantial notions of value that aren't *real*. We're like Cypher from *The Matrix*: We've taken the blue pill, content with simulated steaks, rather than the real thing. Of course, if the whole dispute was whether we were living in a simulation or not, and we sincerely held that we were in the real world, such an objection would make no sense. Just the same, realists who object to antirealists conceptions of value that reject these notions as "real" merely on the basis that the conception in question isn't a realist conception are expressing a view that is warranted only if realism were true, which is of course not something a realist is entitled to presume as a given in discussions with realist, any more than theists are entitled to presume the existence of God in any genuinely open discussion with atheists.

Such speculation is predicated almost exclusively on my interaction with moral realists, who've mostly consisted of professional philosophers, Christian apologists, and autodidacts or people with at least some interest or training in philosophy. It's less clear that ordinary people would fall victim to the same conflations, since such conflations could be caused in part by philosophical education. However, consider how difficult it is for ordinary people (or philosophers, for that matter) to disentangle metaphysical and epistemic considerations. It seems plausible to me that, even if people required adequate instructions to understand what they're being asked, that they'd be similarly inclined to entangle metaethics and normative ethics, and I suspect they'd be even *more* inclined to do so in the absence of adequate instructions. A predisposition to entangle different concepts and fail to draw clear distinctions may be the natural state of the human mind when initially engaging with philosophy. This is, I believe, precisely what we should expect. After all, one of the central tasks of analytic philosophy has been to devise, discover, and develop such concepts and distinctions, and to consider their implications and relationships with other concepts. If we readily drew such distinctions without issue, this work would be unnecessary. The fact that it's apparently not only necessarily but almost exclusively the domain of elite academics is telling.

## S2.11 Lack of realism

In the main text, I refer to three types of realism that, when insufficient, threaten the validity of the disagreement paradigm:

(i) Experimental realism

(ii) Mundane realism

(iii) Psychology realism

I provide additional commentary on each here. None of these problems are so serious and so pervasive to threaten the validity of all research on folk metaethics. Nevertheless, all three pose at least a minor

threat to the validity of the disagreement paradigm itself, while some threaten specific studies that employ the disagreement paradigm.

**S2.11.1 Lack of experimental realism**

With respect to social psychological research, *experimental realism* refers to the degree to which participants take stimuli sufficiently seriously that it "has an impact on them" (p. 131, Gilbert, Fiske, & Lindzey, 1998). If people do not take experimental stimuli seriously, it may fail to prompt the psychological response that would be present in the real-world social circumstances that study is intended to represent. There is no question that many studies conducted in the lab or via online surveys do not represent "realistic" social situations, but insofar as participants are engaged with and take these studies seriously in such a way that induces similar responses as those real-world circumstances, then such studies still have experimental realism. Experimental realism is thus not about whether the experimental context superficially resembles real-world circumstances, but whether it succeeds at engaging participants with experimental stimuli as researchers intend. For instance, a study on social exclusion may involve going into a lab then playing a game on a computer. This situation may be extremely artificial, but so long as participants are engaged with the task, and really believe e.g., that someone else decided to exclude the participant because they didn't like them, the study will have successfully achieved its goal of inducing the emotional response associated with social exclusion.

If participants regard the experimental circumstances as silly or unserious, this can undermine the external validity of the study, since how people respond in situations that they don't take seriously may not generalize to how they'd respond to situations they do take seriously. For example, suppose participants are told that failure on a task will result in another person receiving electric shocks. If participants do not believe another person would really receive real shocks, they would probably respond differently than if they did believe someone else might genuinely suffer. Or suppose

researchers are interested in how people assess moral transgressions, but use stimuli with elements participants find absurd or humorous. For instance, the infamous chicken stimuli used by Haidt, Koller, and Dias (1993):

> A man goes to the supermarket once a week and buys a dead chicken. But before cooking the chicken, he has sexual intercourse with it. Then he cooks it and eats it. (p. 617)

Haidt and colleagues intend for this item to represent a moral transgression that comports with Moral Foundation Theory's *sanctity/degradation* domain (Graham et al. 2013). Other studies exploring the relation between disgust and morality have adopted this item or developed similar items (Horberg et al., 2009; Parkinson et al., 2011). Some of these items defy description. To illustrate just a few examples, consider the items employed by Parkinson et al. (2011):

> Jane's father asks her to stimulate his penis right after he dies to see whether he gets an erection. She never promises, but after he dies Jane stimulates his penis for several minutes with her hand. Jane suffers no ill effects, and she feels sexually aroused.

> Fred goes to a large chain supermarket once a week and buys a fresh whole chicken. At home, he thoroughly cleans the chicken and rubs butter all over it. Then he has sexual intercourse with it, using a condom. He does this only once.

> Ursula occasionally buys leftover pig sex organs at a butcher shop in a large grocery store. After she takes them home, she plays with these sex organs by inserting the male organs repeatedly into the female organs, slowly at first and then faster to simulate sexual intercourse.

My initial reaction to these scenarios is that they are *ridiculous,* so it is important to reiterate that these scenarios are *not intended to be funny.* I find it difficult to imagine that whoever developed scenarios experienced no *mirth* (Martin, 2007) when they did so, and I suspect lab discussions about these items involved a fair number of chuckles. Perhaps even a chortle. Given what strikes me as the transparently humorous nature of these stimuli, is it possible that participants given these scenarios did not take them seriously, and were less engaged with them? I suspect so, but to my knowledge no efforts were made to assess whether participants found the stimuli humorous, and if so, whether this influenced their response in ways detrimental to the study.

In short, when participants disengage or fail to take experimental situations seriously, their reactions cannot inform the hypotheses of interest. Most research on folk metaethics does not suffer from the most serious threats to experimental realism. However, experimental realism may be threatened if participants are presented with scenarios that they find humorous or implausible, since this could cause them to disengage or think in ways that don't reflect how they'd think about moral issues under ordinary circumstances. Unfortunately, one study does use humorous stimuli.

In a series of studies conducted by Sarkissian et al. (2011), participants were given the following scenarios:

> *Horace finds his youngest child extremely unattractive and therefore kills him.*

> *Dylan buys an expensive new knife and tests its sharpness by randomly stabbing a passerby on the street.*

When I first read these sentences it resulted in a literal spit-take. These scenarios are *ridiculous*. There is something comical about the bizarre juxtaposition of completely mundane and understandable attitudes (recognizing your children are ugly) and motives (wanting to be sure a knife is sharp) and the bizarre, over-the-top, psychopathic actions that follow from them. The absurdity of these scenarios is further compounded by the conditions many participants were assigned to. Some participants were asked to consider a moral disagreement between a member of their own culture and a Pentar, a member of an alien species described as follows (with the accompanying picture):

**Figure S2.1**

*Extraterrestrial condition from Sarkissian et al. (2011)*



Imagine a society of extraterrestrial beings called the Pentars. The Pentars have a very different psychology than us. They do not experience love, friendship, pleasure or pain. They do not pursue the sorts of goals that we do. Instead, their entire lives are organized around a single project—the effort to reshape every object they can find into perfect pentagons. They are extraordinarily rational and efficient in the way they work together in achieving this goal, and they can always count on each other's collaboration. However, if it turns out that they can best achieve the goal by killing other Pentars, they immediately go ahead and proceed with the killing (after which they reshape the dead Pentars into pentagons themselves). None of them see anything wrong with this sort of behavior.

As creative and amusing as this design may be, scenarios like this could pose significant methodological shortcomings.[75] As Pölzler (2018b) argues, studies that employ *unrealistic* and *humorous* stimuli may have limited generalizability, a finding supported by research on the impact that such scenarios have on moral judgment in other domains of research (Bauman et al., 2014). Humorous stimuli in particular threatens external validity because it can reduce engagement and can prompt unintended psychological processes.

As Bauman et al. (2014) note, the thought experiments philosophers use often deliberately include humorous elements to lighten the mood and make it easier to digest what might otherwise be otherwise excessively grim topics. A little comedic sugar may help the philosophical medicine go down, but it may be inappropriate to export these elements into the stimuli used in social scientific

---

[75] This is not idle praise. I have both a personal love of aliens and science fiction and a proclivity for over-the-top thought experiments. In fact, I routinely aggravate my family, my wife, and complete strangers with scenarios at least as outlandish as this one. I genuinely love this scenario, and believe it is completely suitable for a classroom of students with sufficient background in philosophy and familiarity with counterfactual thinking to engage with it. But, in my experience, most ordinary people respond to such scenarios with an enormous amount of resistance. Philosophers have good reason to come up with strange and implausible scenarios. Doing so allows them to exclude extraneous considerations that might otherwise be irrelevant. Much as scientists must carefully control laboratory conditions in ways that render lab conditions very different from everyday life, but do so for the purpose of excluding the noisy interaction of real-world variables, philosophers construct thought experiments to exclude irrelevant considerations. Yet in practice, laypeople often respond with an unhelpful (but understandable) focus on the irrelevant features philosophers introduce, and take into consideration the real-world plausibility of the hypothetical scenario. As such, I think it's a mistake for philosophers to present strange scenarios to laypeople and expect them to respond appropriately.

research. People regard situations with amusing elements as less serious and more distant from everyday experiences (Apter, 1982; 2014; Martin, 2007; McGraw, Williams, & Warren, 2014; Morreal, 2009)[76], which can lead them to be less engaged with the moral content of the stimuli or regard it as unimportant (Bauman et al., 2014; p. 541; McGraw & Warren, 2010; McGraw, Williams, & Warren; McGraw, Schiro, & Fernbach, 2015; Yang et al., 2019). For instance, McGraw, Schiro, and Fernbach (2015) found that participants judged the issues raised in public service announcements (PSAs) to be less important when viewed a humorous PSA than when they viewed a more serious PSA. In a second study, they also found that humorous PSAs about sexual health were less effective than serious ones at motivating viewers to sexual health information. Taken together, these findings suggest that humorous stimuli are taken less seriously and that this can influence real-world behavior. However, these findings do not *guarantee* that humor will prompt undesired psychological processes or fail to promote desirable outcomes. Ort and Fahr (2020) found that humorous PSAs were more effective than threatening ones. These differences could be a result of differences in the stimuli each study used, but I am not claiming that humorous content will *necessarily* produce unintended or negative consequences, only that it *can*. As such, its use should be carefully considered, and its impact evaluated. Unfortunately, Sarkissian et al. (2011) did not adequately address the potential interference humor may have had on their results. Their stimuli are the perfect candidate for the inclusion of unnecessary humorous elements that risk being taken less seriously and prompting a different reaction than more serious stimuli. As a result, reactions to these scenarios may not generalize to how people react to other moral issues, including more serious and realistic ones and to more prototypical moral issues in general.

Humorous or absurd stimuli can also prompt participants to reject the information stipulated in the scenarios described in a study. Rejecting experimental stimuli is similar in some ways to

---

[76] These references were primarily provided by Bauman et al. (2014).

*imaginative resistance*, which occurs whenever people have difficulty imagining a scenario as intended, either because they are unwilling or unable to do so (Gendler & Liao, 2016). Likewise, participants may struggle to imagine the scenarios researchers present them because they are too abstract or strange, or may simply refuse to accept the stipulations presented in a set of stimuli. Philosophers may be fond of outlandish and bizarre hypotheticals, but ordinary people often react to scenarios involving intergalactic space pirates or super-advanced AI with incredulity or obstinance. When this occurs, people may reject these scenarios outright or incorporate extraneous information or assumptions that are not part of the stimuli, in order to make better sense of them. Either way, when participants can't or won't consider the scenarios described by a study as intended, their responses *cannot* be a valid measure of whatever it is researchers are trying to evaluate, since such responses are effectively responses to a different question than what the study asked.[77]

There is some evidence participants will reject information explicitly stipulated in a hypothetical scenario. Ryazanov et al. (2018) gave people trolley problems and other sacrificial dilemmas, and explicitly stipulated that the outcomes of different decisions one could make in these

---

[77] It cannot be valid because researchers must opt in advance to interpret responses in accordance with an *a priori* interpretive framework. For instance, if I want to know whether people believe in God, and ask:

*Do you believe in God?*
  - ❏ *Yes*
  - ❏ *No*

My findings can only provide an accurate estimate of the proportion of participants who believe in God if they interpret the question as I intend, i.e., as a question about whether they believe in God. For simple questions like this, wildly divergent interpretations are unlikely. But if we asked a complicated or ambiguous question, interpretations may systematically vary in our population. We may wish to know the proportion of people who believe or disbelieve X, but we may have some unknown quantity who instead interpret the question to be about Y or Z. Of course, unintended interpretations are inevitable. It is always possible a few people will interpret questions in unintended ways, yet this does not undermine a measure's validity any more than a small false positive rate would invalidate methods for diagnosing illnesses. The only relevant question for the validity of a psychological tool is *how much* unintended interpretation occurs. It may be that only a handful of participants are subject to imaginative resistance or engage in substitution. On its own, such concerns might only undermine the precision of the disagreement paradigm. Yet considered in light of the many other criticisms I raise in the sections that follow, the complete picture starts to look more like death by a thousand cuts. I emphasize this now because it is important to keep in mind as I present this and other criticisms. Each, on their own, might seem like a manageable concern, but together, it is far less clear that we should continue to use the disagreement paradigm or other survey methods to explore folk metaethics.

imagined scenarios were guaranteed to occur (i.e., that if the person in the situation pulls a lever, it will

kill one person but save five, but if they do not, the five will die). Ryazanov et al. found that participants

did not accept that the stipulated outcomes were guaranteed to occur, and instead substituted their

own intuitions about the probabilities of various outcomes. These substituted probability estimates in

turn influenced participant's responses to these questions. In other words, participants opted to

interpret the study in a way that differed from how the researchers intended it to be interpreted, and

this difference in interpretation influenced how they responded to the study prompts, leading

Ryazanov et al. to conclude that "It seems clear that people do not understand the scenarios in

precisely the way they are intended" (p. 65). These results are corroborated by Greene et al. (2009).

They found that participants who judged the details of events in sacrificial dilemmas to be unlikely

responded differently than participants who did not judge them to be as improbable.[78]

Other studies have also found that people often struggle to imagine fictionalized scenarios

(Liao, Strohminger, Sripada, 2014). Of course, these findings merely demonstrate that there are some

circumstances where people refuse to or fail to entertain features of experimental stimuli; they don't

demonstrate that people did so Sarkissian et al.'s study, or any other research on folk metaethics.

However, they do suggest that many people struggle to accept the information given in scenarios

commonly used in moral psychology and folk philosophical research. Perhaps they also struggle to

---

[78] Greene et al. (2009) asked participants if they did not accept the sacrificial dilemmas as described. They found that 5% and 12% agreed that *"I did not find the description from the preceding pages to be realistic, and my answers reflect my inability to take seriously the description that was given"* (as quoted in Bauman et al., 2014, p. 543). These percentages are quite low, yet Bauman et al. (2014) note that they may underestimate the true proportion of participants who cannot readily imagine these scenarios. First, because participants may be reluctant to acknowledge their inability to imagine the scenario as described due to demand characteristics that motivate socially desirable behavior (Orne, 1959; Weber & Cook, 1972). Second, participants would require access to the factors that drove their judgments, but such factors may not always be transparent or available to us (Nisbett & Wilson, 1977). Since the exclusion of these results was insufficient to eliminate the impact of differential acceptance of scenarios Ryazanov concluded that excluding participants who explicitly reject stimuli would not be sufficient to mitigate concerns about external validity. Excluding many participants could pose its own set of methodological problems, as well. Studies that exclude too many participants have reduced power and run the risk that remaining participants systematically differ from those excluded in ways that can undermine random assignment (if conducting an experiment with random assignment to condition) and reduce the generalizability of findings, since it is unclear whether findings that exclude skeptics can generalize to skeptics.

imagine what it would be like to be a Pentar, to imagine the mindset of people from unfamiliar cultures, or to consider the perspective of someone who would casually murder others for trivial reasons.

Fortunately, poor experimental realism may be limited to Sarkissian et al.'s findings, since only these employed especially humorous and unrealistic stimuli. To the extent that experimental realism threatens the disagreement paradigm in general, this threat seems relatively weak. There is nothing especially unrealistic about moral disagreements with ordinary people about familiar moral issues, and there is no compelling reason to believe people would be any less engaged by the sober stimuli standardly employed in the disagreement paradigm than they would be by the typical content of social scientific surveys.[79] The other two threats to external validity, however, pose a far more serious challenge to the disagreement paradigm, since they tend to threaten the validity of *all* existing studies.

**S2.11.2 Lack of mundane realism**

A second threat to the external validity of the disagreement paradigm is the extent to which it lacks *mundane realism*, that is, the extent to which the experimental context reflects the conditions participants would encounter in everyday experience. Does the disagreement paradigm capture the kind of situation people would encounter in everyday life? At first glance, the disagreement paradigm seems to present participants with stimuli that have a very high level of mundane realism. First, the moral issues themselves are ones most of us are familiar with. Consider those used by Goodwin and Darley (2008):

(1) Donating to charity
(2) Bank robbery
(3) Racial discrimination

---

[79] On the other hand, it is possible that, precisely *because* people are engaged (emotionally or otherwise) with the substantive content of the moral issues themselves, that metaethical considerations tend to be less salient, such that participants interpret what they are asked in unintended ways or are subject to performance errors that undermine the validity of the disagreement paradigm in ways unrelated to experimental realism.

(4) Cheating on an exam

(5) Abortion

(6) Euthanasia

(7) Embryonic stem cell research

(8) Providing false testimony in court

We are all acquainted with these issues, if not by having direct experience with them, then indirectly by hearing about them on the news, having discussions about them, or by interacting with people who have experiences with them. Moral disagreements are also a part of everyday life. Ordinary people may rarely have sustained philosophical disputes about the moral status of abortion on a regular basis, but most people nevertheless have significant direct and indirect experience with moral disagreement, since we sometimes argue with friends, family, or colleagues about the moral status of the actions or policies of people or nations.[80] It seems then that the disagreement paradigm does not lack for mundane realism. Moral disagreements are a part of everyday life, and the moral issues researchers typically employ are not typically far-fetched and unusual.

Bauman et al. (2014) convincingly argue that research that relies on trolley dilemmas lacks mundane realism because aspects of the situation strike participants as implausible and unrealistic. People often express doubt that a single person's body could stop a train, or question why the workers could not get off the tracks, or wonder whether a person in such a situation could appraise all the relevant considerations in time to deliberate and act. These are all reasonable concerns that highlight the importance that practical, contextualized considerations play in everyday judgment. Yet many of these concerns are less applicable to the situations described in the disagreement paradigm. The disagreement paradigm should be lauded for presenting what are, in many respects, utterly mundane moral circumstances: the simple occurrence of someone holding a contrary moral belief about familiar

---

[80] Though it is probably rare to encounter people who explicitly insist that e.g., robbing banks is morally acceptable. Some moral disagreements are more likely to energy in everyday life than others.

moral issues. It may lack the contextual richness of a real-world disagreement, but it is still the *kind* of experience we could readily imagine ourselves in.

Yet Bauman and colleagues point to another issue with trolley problems. Participants often insist that they do not have enough details (see also Bloom, 2011). Do they know any of the people on the footbridge or tracks? Are any people in a position of authority aware of what is happening? Are there witnesses? Will their decisions have potential legal ramifications? Why are there no safety mechanisms in place? These are reasonable questions, and highlight one of the most important ways a study can lack mundane realism. Everyday moral situations tend to be *situated* within *specific social contexts*. Yet the stimuli used in the disagreement paradigm are abstracted from real-world contexts. We don't typically make judgments about highly artificial and abstract situations in everyday life. In fact, we almost never do. The kinds of moral judgments that researchers are attempting to generalize towards are not about hypothetical situations, but actual situations, and actual situations plausibly include content absent from hypothetical situations that are an essential feature that constitutes everyday moral judgment. For instance, everyday moral judgments often involve people we know or know about. Whether someone is a family member or a friend or a member of our community is not incidental; such details are integral to the situated circumstances that characterize everyday moral judgment. And we may evaluate the moral status of a person's actions differently if that person is a member of our ingroup, or an ally, or shares our ideological values, compared to someone who is an outgroup member, or an enemy, or holds ideological values we despise. Such differences in judgment could be a result of bias or error, but they could be an essential component of everyday moral judgment.[81]

---

[81] Any research that ignores this possibility and imposes *a priori* presumption that moral judgment must necessarily be free of partisanship or partiality may be imposing external, ideologically-motivated standards on what moral judgment must be that are not necessarily justified. Philosophers may insist that a moral judgment must, by necessity, be impartial and apply consistently to all moral agents, and that any deviation from this is some external bias or error not constitutive of moral judgment itself, but such standards may reflect the idealistic aspirations of people inducted into specific, hyper-

In short, everyday moral experiences are not about hypothetical people in hypothetical circumstances, but real people in real circumstances. And such circumstances involve a host of rich and relevant details: what are the motives of the people involved in these situations? Do we know any of these people? What social roles do they have? What is their social standing relative to one another? What is one's relation to the people in these situations? Are they aware of their moral obligations? Consider, for instance, how being a parent or a doctor or a bystander to a crime may endow us with obligations that we possess only in virtue of our social role within these contexts. And consider how a person's motives, emotional state, and epistemic access to relevant details of a situation all play a critical role in our moral evaluations of that person's conduct. Everyday moral judgments are contextually rich. Studies that provide brief descriptions of hypothetical scenarios are a pale shadow of the real thing. It is unclear whether they are capable of activating the psychological processes involved in everyday moral judgment. Since contextual details may play a role in the psychological processes that characterize moral judgment as it occurs in the real world, inadequate context may undermine the external validity of the disagreement paradigm not because it is difficult to imagine the situations that are described, but because abstract scenarios that lack relevant contextual details fail to trigger the psychological processes present in everyday moral judgment.

Another problem with impoverished stimuli is that participants may prompt participants to engage in unintended substitution or extrapolation. This occurs whenever participants (consciously or unconsciously) "fill in the gaps," adding details to the scenario that researchers did not intend, or altering features of the scenario to render it more believable.

This can pose serious problems to external validity, because different participants may add different details. Researchers may lack access to these details, since participants are rarely asked to

---

intellectualized modes of thought that do not reflect how actual moral judgment works. Our job is to describe moral judgment as it really is, not how we'd like it to be.

describe them (and they may be opaque to participants themselves, or they may lack motivation or ability to report these details). This means that researchers may not know exactly how participants are interpreting these questions. A more serious problem is that participants may add *different* details to these scenarios from one another. When this occurs, we cannot be sure whether differences between participants are due to variation in their beliefs, attitudes, or other psychological processes of interest to researchers, or are instead due to differences in how they interpret the question. Finally, suppose participants do not add additional details to these situations. Do the patterns of judgment found in studies that assess judgments about abstract, contextually-impoverished stimuli serve as an adequate proxy for the kinds of judgments that take place in concrete, contextually rich circumstances? It is not obvious that they do.

Most folk metaethical research does not include descriptions of Pentars, but participants are asked to imagine people who hold different moral views than they do. This may be easier than imagining aliens that want to convert all matter into pentagons, but depending on the moral issue in question, it may not be that easy to entertain a genuine commitment to a contrary perspective. Most of the scenarios that participants are given are simple, generic, and nonspecific. There are many good reasons to use such stimuli. If we are interested in variation in people's responses to different moral issues, e.g., abortion, murder, theft, and so on, it would be impractical to provide long, detailed descriptions of specific events. Doing so would carry its own risks, such as introducing extraneous details that influence how people respond that researchers are not explicitly aware of. At best, we could not capture *all* of the features that characterize everyday moral issues in their full context, absent conducting field studies in which participants were led to believe they were entangled in a genuine moral crisis. The situations stipulated in surveys will thus always be impoverished in various ways, e.g., the participant will have no personal connection to the people involved, there are no actual consequences for expressing the wrong moral judgment, and even if a situation is described in exacting

detail, it will never come close to real situations. After all, *descriptions* of sunsets don't even come close to real sunsets. Likewise, descriptions of hypothetical moral scenarios can never truly match actual morally-laden events in the real world; the best we can hope for is that impoverished stimuli are sufficient to prompt the appropriate psychological states or trigger the relevant kinds of judgments. However, it is possible researchers underestimate the gulf between cold and detached "moral judgment" in the lab and bona fide moral judgment in the real world; people's responses to the former may just not provide that much insight into the latter.

For instance, Goodwin and Darley (2008) use the following item: "Opening gunfire on a crowded city street is a morally bad action." Most participants believe this is morally wrong. Yet to interpret this question as intended, they must imagine a person who disagrees with them, not because of some justifying rationale that could conceivably warrant opening fire on a crowd, e.g., to attempt to kill a fleeing terrorist who is plotting to detonate a nuclear bomb, but simply because *there is just nothing wrong with recklessly killing innocent pedestrians*. In other words, they must imagine a person who appears to be a psychopath, and this interpretation is the *required* interpretation for response to this question to reflect the phenomenon researchers are trying to measure. Research participants are not philosophers. They are not familiar with and readily disposed to entertain wildly unconventional mindsets. When given a scenario like this, instead of envisioning a psychopath, they might instead presume that they *must* be dealing with an (at least somewhat) ordinary person, which in turn prompts them to reflect on how such a person could think shooting into a crowd could be morally acceptable. This might prompt them to reason as follows:

> Why would a normal person who isn't all that different from me think it was okay to fire a gun into a crowded city street? Aha! They must be thinking of situations where this *would* be morally justified (in a way consistent with conventional moral standards).

I am not suggesting that *all* participants respond this way. But at least some do. Goodwin and Darley collected data on why participants thought others disagreed with them.[82] Several responses are consistent with the line of thinking proposed above. Here are some examples:

> **Response #1**: *A difference in perception of a situation in which gunfire was opened on a crowded city street. I was thinking gunfire from terrorists/ criminals; other person may have thought gunfire from police officers to catch a criminal.*

> **Response #2**: *The other person could be thinking about certain circumstances like the protection of others if there was a threat.*

In order for their responses to reflect a genuine moral disagreement, participants in this study must regard the person they disagreed with as having a genuine commitment to different moral standards that led them to hold a contrary moral stance towards *the exact same cases*. Yet this is not what these participants imagined. Instead, they imagined that the other person must have some reason to express an only *apparent* disagreement, not a genuine one. To make sense of a situation that they had trouble imagining, they *interpreted the nature of the disagreement in a way that differed from the interpretation necessary for their response to be valid.* This is because such participants effectively responded to a different question than the one that was asked. For such participants, there was no genuine disagreement. Instead, two people simply imagined *different* scenarios. Even a moral realist ought to judge that two people who judge two different scenarios could both be correct. Yet anyone who did so would select a response option that would be misinterpreted by researchers as antirealism.

Note also that such unintended interpretations are not the result of incompetence or inattention. The kinds of interpretations highlighted in these examples are perfectly reasonable ways

---

[82] Specifically, they asked the following question: "Give us your thoughts about why it is that there is disagreement. What could be its source?" I would like to thank Geoffrey Goodwin for providing this data to me. I have been reluctant to offer criticisms of the work of others, especially when they have been generous with their time and their data. So I would like to note that while I raise objections to their findings here and elsewhere, I am deeply impressed with the clarity and forethought that went into their experiments and the philosophical rigor of the surrounding discussion. Unfortunately, I believe research on folk metaethics is like an attempt to cross a minefield with no safe surfaces. No matter how well one navigates the terrain, it's just not possible to make it across unscathed.

to interpret the situation. In fact, such interpretations are, if anything, far more plausible than the interpretations researchers intend. So not only must participants interpret the disagreement paradigm in a very particular way, the particular way they are expected to do may be less plausible than plausible unintended interpretations.

Alternatively, participants may simply insist that there is not enough information to judge the situation. For instance, one participant responded:

> **Response #3**: *There is not enough information about why there would be gunfire - i.e., are there terrorists in the street someone is trying to kill, or is the street filled with innocent people?*

When participants are not given enough information about the scenario, they may find themselves incapable of responding to the question appropriately. Two people who disagree about an ambiguous and underspecified claim may not disagree about the same moral issue. If so, then whether they can both be correct or not has no metaethical relevance and any response the participant gives will be unrelated to the question researchers intended to ask.

Furthermore, even when a moral issue is well-specified, and there are no plausible circumstances where it might be justified (e.g., torturing someone just for fun), it is natural to draw inferences about a person who supposedly disagrees that go beyond what the disagreement paradigm stipulates. In our everyday experience, a person who exhibited a complete disregard for other people's welfare (e.g., by disagreeing that it is morally wrong to torture people for fun) probably *would* suffer from serious psychological deficits. It is hard (without philosophical training, and perhaps even with it) to imagine otherwise. If so, such a person may not be expressing a sincere commitment to a different moral position so much as failing to express a genuine moral judgment at all. Such a person's disagreement would not reflect a moral stance that differs from the participant's, so much as an unwillingness to act in accordance with moral standards they share with participants. For instance, someone who says it's fine to murder others may not deny that it's *immoral*, they might just *not care* that

it's immoral, and in saying it's fine to murder they could simultaneously (a) acknowledge the moral rule against the action (b) express a willingness to violate that rule anyway. Some participants hint at interpreting disagreements about moral issues this way. For instance, in response to the question about why a person might disagree with the participant about firing on a crowd, one participant responded:

**Response #4**: *I don't even understand how they could have their opinion, unless they suffered psychological abnormalities or are morally depraved.*

First, note that this participant explicitly states that they don't understand how someone else could think this. They also suggest that the person who disagrees may suffer "psychological abnormalities." This participant seems to be struggling to envision a person with a normal, functioning brain who has simply reached different moral conclusions. This presents yet another difficulty for the disagreement paradigm. We may have little difficulty imagining people who disagree about abortion or euthanasia. But it could be much harder to imagine a person disagreeing about murder or genocide. Such people are radically unlike the kinds of people we typically encounter. Reactions to such disagreements may differ in unintended ways to reactions to disagreements about controversial moral issues. If so, they may lack external validity, because they may trigger psychological processes that are less relevant to moral judgments in more familiar contexts. This is compounded by the possibility that these processes are only triggered by *some* items, but not others. When this happens, it can lead to systematic variation in how participants interpret sets of items. Yet to serve as a valid measure, participants ought to consistently interpret these questions in the same way, since if they do not, then cross-item comparisons and aggregation of responses may be inappropriate.[83] In other words, suppose participants interpret *some* moral disagreements to reflect genuine moral disagreements, but not others.

---

[83] They do not need to be interpreted consistently in *every* respect. There is nothing incoherent about being a relativist about some moral issues and a non-relativist about others. The problem is, rather, that participants may regard some moral disagreements as genuine and others as not being genuine, or imagine an alternative interpretation of a situation for some moral disagreements but not others. In such cases, participants are effectively responding to *different* questions that only superficially resemble one another. Yet researchers will aggregate such data and treat all interpretations as uniform.

If so, their responses effectively reflect answers to different questions. Yet researchers may combine responses to both sets of questions *as if* they were responses to the same question.

The risk that participants will struggle to imagine a person who disagrees is not limited to questions about firing into crowds. When confronted with a person who disagrees with them about the claim that it is wrong to cheat on a lifeguard exam, one respondent stated:

> **Response #5**: *I honestly don't know how anyone could disagree with this statement - I suppose if they were used to cheating.*

These remarks provide some (albeit minimal) evidence that participants often struggle to understand how people could reach different moral conclusions. While this may be unimportant for some studies, the disagreement paradigm only provides a valid measure of the participant's metaethical stance if they regard the moral disagreement as genuine (Bush & Moss, 2020; Goodwin & Darley, 2008).

Unintended interpretations like these cannot necessarily be dismissed as simple participant error, either. In principle, we could use training exercises to enhance comprehension or use comprehension checks to exclude participants who do not interpret disagreements as intended. Yet doing so risks eliminating people for drawing on precisely those psychological processes that are integral to moral judgment, or at least relevant to how *some* people think about moral issues. First, it is natural, when confronted with a disagreement, to consider all the possible reasons why a person might disagree. It is not obvious that the only reason that a person would reach different conclusions about a moral issue is because they have different fundamental moral values. This is not a natural or even obvious way to interpret moral disagreements. Many (perhaps most) moral disagreements may be attributed to disagreements about the nonmoral facts. Many apparent moral disagreements, in other words, are not really disagreements about what is morally good or bad, but whether a particular action or policy conforms to a moral standard. Disagreements can also be superficial. Sometimes people only appear to disagree when in fact they are talking past one another. Some participants seem to recognize

this possibility when they speculate that another person might disagree because that person is imagining a *different* situation. These are *sensible* reactions to ambiguous stimuli. Researchers do not typically provide detailed instructions that fully disambiguate the possible sources of disagreement (or apparent disagreement) and specify precisely what kind of disagreement they intend. In other words, participants are simply not given enough information. Arguably, it is as much or even more of a mark of competence and engagement that participants would "misinterpret" the disagreement paradigm than interpret it as intended.

Such "misinterpretations" are, if anything, more plausible when a participant is told that another person disagreed with them about shooting into crowds or other egregious moral violations, or asked to imagine such a person. Participants are expected to suspend all knowledge of how actual psychologically normal humans are and somehow merge "psychologically healthy person with no serious cognitive or emotional deficits" with "considers it morally acceptable to shoot innocent pedestrians." This may be difficult to imagine, and may prompt participants to imagine that if another person disagrees there must be a reason why they would do so in a way consistent with ordinary human psychology. In other words, people may be naturally drawn to inferences that differ from or go beyond what the disagreement paradigm stipulates. In our everyday experience, a person who exhibited a complete disregard for other people's lives probably *would* suffer from serious psychological deficits. It is hard (without philosophical training, and perhaps even with it) to imagine otherwise. When this occurs, participants may feel it necessary to draw additional inferences that conflict with the intended interpretation of the study. They may conclude that if a person disagrees with them about fundamental moral issues, that this person is not expressing a genuine moral stance, or is confused, or misread the question, or is imagining some specific situation where the action would be justified. Indeed, when Goodwin and Darley (2008) asked participants to explain what they thought

the source of a moral disagreement between themselves and a previous participant, one participant stated that:

> **Response #6**: *I think the other individual may have misinterpreted the question. I know I had to read it twice to be sure I understood it correctly, so it's very easy to miss a key work. Robbing a bank is morally wrong, especially to pay for a vacation.*

Another participant suggested that the other person might be joking:

> **Response #7**: *They're either joking, or have a very different sense of what's morally acceptable.*

These *specific* ways of interpreting the question are infrequent, and if they were the only difficulty, it would pose little problem for the disagreement paradigm. We could simply exclude the small handful of people who did not interpret questions as intended.

However, there are three reasons why doing so cannot adequately address concerns about unintended interpretations. First, not all participants will be willing, able, or motivated to report that such considerations influenced how they answered, and the influence inferences like these may have on participant response may not even be accessible to participants. Even if such considerations do occur, they may be underreported, because participants may engage in post-hoc reasoning about multiple possible sources of disagreement and report other possibilities. In fact, response #7 presents *two* possibilities, and there are many other instances of participants offering at least two explanations for why a person might disagree. This illustrates that people are often uncertain about which particular interpretation is incorrect. If anything, this makes it even less clear which (if any) of the possibilities they propose influenced their response to the disagreement paradigm. There are also numerous *other* ways participants can interpret questions in ways that don't match the interpretation necessary for their results to reflect their metaethical stances or commitments. Even if each of these unintended interpretations peels off a small proportion of participants from the overall proportion that interpret the disagreement paradigm as intended, their cumulative impact can add up.

There may be no simple solution if large numbers of participants interpret stimuli in unintended ways. Even if researchers include comprehension checks and assess how participants interpret questions, then exclude participants who fail these checks or express unintended interpretations of stimuli, such exclusions may fail to capture residual unintended interpretations. And such comprehension checks would be hard to implement, since we'd simply recapitulate the same concern with unintended interpretations: how can we be sure their response to the comprehension check reflects an intended interpretation of the relevant stimuli? While such concerns may be implausible for a wide variety of conventional questions in psychology, they may be far less effective when asking about subtle and sophisticated philosophical questions. For instance, even if a person explains why they think someone who disagrees with them about a moral issue must be mistaken by saying that it's because they "think morality is objective," this is far from adequate for securing compelling evidence, because my findings suggest most people don't clearly interpret "objective" to mean "stance-independent." Multiple choice and other methods of assessing comprehension may provide at best only shallow insights that reveal marginal, superficial competence with or understanding of the terms and concepts used in the study; they will typically be incapable of demonstrating that participants are interpreting questions as intended.

Even if such measures worked, and we could reliably exclude participants who interpreted stimuli in unintended ways, this will at the very least reduce a study's power, but may also introduce self-selection effects that limit the external validity of the findings, and, when those unintended interpretations differ across conditions, may prohibit confident causal inference due to *posttreatment bias* (Montgomery, Nyhan, & Torres, 2018). As Montgomery et al. show, conditioning on posttreatment variables, which includes excluding participants who fail comprehension checks, can undermine the value of a study by introducing systematic biases in the remaining pool of participants whose results are analyzed. In experimental contexts, this can undermine causal inference because

"[c]onditioning on posttreatment variables eliminates the advantages of randomization because we are now comparing dissimilar groups" (p. 762). However, even if we're not conducting experiments or attempting to draw causal inferences, excluding participants who fail a comprehension check may lead to the remaining pool of participants not accurately reflecting the population they were drawn from.

For instance, it would make little sense to infer that members of a particular population interpreted a question as intended if one of the criteria for inclusion in analysis was passing a comprehension check that demonstrated the intended interpretation! The proportion of people who were excluded using this method would be an essential element of assessing challenges to the validity of the measure, especially when those challenges are similar to the challenges I've raised, e.g., that people struggle to interpret questions as intended. If that number is very large, then the remaining pool of participants may represent little more than an unrepresentative group whose responses would be useless for making inferences about the target population. For comparison, suppose we wanted to know whether ordinary people endorse A-theory or B-theory of time. However, we excluded anyone from the study who was unable to provide a detailed description of special relativity.

Now suppose a handful of participants could do so, and the majority of these participants endorsed B-theory. Should we conclude that people in general favor B-theory? Of course not, for the simple reason that anyone who could pass the comprehension check differs in a variety of ways from anyone who couldn't, and may not represent everyone else. For instance, it could be that people who could explain special relativity are much more likely to endorse B-theory. Given that proponents of B-theory maintain that special relativity entails or is consistent with B-theory but not A-theory, this is actually a very reasonable assumption (Fazekas, 2016; Koons, 2022; Maxwell, 2006). It could be that most people intuitively favor A-theory, but that people who understand special relativity will tend to endorse B-theory. If so, participants who pass the comprehension check would exhibit the *opposite* response pattern as those excluded from the analysis, e.g., it could be that 90% of those who

understand special relativity endorse B-theory, but that only 10% of those who don't understand special relativity endorse B-theory. This would be an extreme case of exclusion criteria dramatically flipping results, but this is a hypothetical example intended for illustrative purposes. In practice, such exclusions may result in more subtle threats to validity.

### S2.11.3 Lack of psychological realism

Low psychological realism represents another threat to the external validity of folk metaethics research. *Psychological realism* represents the degree to which experimental stimuli activate the same psychological processes as those that would be active under the circumstances those stimuli are intended to represent (Bauman et al., 2014). Psychological realism is an essential component of any study, since there is little reason to doubt that "the validity of any study necessarily depends on the extent to which the research setting engages the process of interest" (p. 543).[84] For instance, a study about how anger influences behavior cannot be valid if its method of anger induction fails to make anyone angry.

Does the disagreement paradigm lack psychological realism? This will depend in part on the stimuli used in any particular study. There is no single, canonical set of stimuli, so there may be no *uniform* answer to this question. However, there is one clear way that a study can lack psychological realism: where are compelling reasons to believe that it is activating unintended psychological processes that are known to influence results. Although it seems harsh to pick on Sarkissian et al. (2011)'s study once again, it serves as an excellent example of how psychological realism can threaten the external validity of a study.

---

[84] Low experimental and mundane realism often pose no threat to the external validity of some studies. This is because many studies are not intended to generalize to everyday life in a direct way. Researchers studying visual perception, memory, or other aspects of human cognition may have no need to use stimuli that mirror real-world circumstances. Minimizing or tightly controlling variables present in everyday life may even be necessary for testing some theories, and artificial laboratory conditions may in fact be optimal for testing those theories. Such findings can inform the real world, albeit indirectly: such studies test theories, and we generalize from the theories (rather than the lab settings in which they occur) to the real world (Bauman et al., 543).

When humor is injected into a scenario, it can activate psychological processes that differ from those that would be active in its absence. This threatens the external validity of studies with humorous stimuli, since results may not generalize to circumstances where humor (or at least the psychological processes it triggers) is less salient. And since humor is not a typical feature of ordinary moral judgments, the injection of humor into experimental stimuli may render findings incapable of generalizing towards the primary phenomenon of interest.

There is already abundant evidence that humor can threaten external validity. As Bauman and colleagues point out, there is a considerable that positive states are less motivating than negative ones, (Baumeister et al., 2001; Janoff-Bulman, Sheikh, & Hepp; 2009; Rozin & Royzman, 2001; Vaish, Grossmann, & Woodward; 2008). For instance, Bauman et al. note that people give less to charity when they are given pictures of happy children than when they are given pictures of sad children (Small & Verrochi, 2009). Further, since people seem motivated to overlook unpleasant information that could worsen their mood, participants presented with stimuli that includes both humorous (positive) elements and negative ones may be motivated to place undue emphasis on the positive elements and ignore unpleasant elements of the stimuli (i.e., considerations of murder or abortion, or the fact that someone apparently disagrees with them about these issues) further eroding the degree to which these stimuli reflect and can generalize to the judgments of interest (i.e., prototypical moral judgments).

Notably, Goodwin and Darley (2012) found that people were less disposed to select "objectivist" responses for positive moral acts (e.g., donating to charity) than negative moral acts (e.g., theft), which they attribute to the exact research on *negativity dominance* that Bauman et al. appeal to in their discussion of the impact of humor. *Negative dominance* is the well-established observation that, within a given dimension, people are more attentive to and assign greater weight to negatively valenced events than to positive events. For instance, we are more responsive to criticism or losing than to

praise or winning, respectively (Baumeister et al., 2001, p. 323; Janoff-Bulman et al., 2009). Since negativity dominance already appears to influence metaethical judgment with respect to positive and negative moral acts, this raises the plausibility that humor and other positivity-enhancing factors could influence metaethical judgment by potentially reducing people's inclination to favor objectivism.

Unfortunately, exactly how humor or positive affect in general influences moral judgment remains a subject of contention. Strohminger, Lewis, and Meyer (2011) found that humor's impact on moral judgment may not be driven by a general effect on positive emotion (see also Valdesolo & DeSteno, 2006). In a series of studies, they found that *mirth* (the positive emotion specifically induced by humor) had the *opposite* impact of *elevation*, a positive emotion associated with "witnessing acts of moral beauty," and motivates people to "act in a more noble, saint-like way" (p. 296; Haidt, 2003). Whereas mirth led to a more permissive attitude towards deontological moral violations, elevation increased conformity to deontological moral rules. Strohminger and colleagues conclude that different positive emotions can influence judgment and behavior in distinct and even conflicting ways. In particular, they suggest that mirth is associated with feelings of irreverence and permissiveness, so much so that had their study framed moral violations in utilitarian terms, people may have been more permissive of violating utilitarian rather than deontological norms. In other words, humor does not make people more utilitarian so much as it makes them more morally indiscriminate in general (Yang et al., 2019). It is no stretch to predict that greater moral permissiveness could induce people to favor more antirealist responses to metaethical probes, since antirealist moral stances are typically seen as less rigid and more permissive; indeed, that is often the very reason why people are motivated to endorse various forms of moral relativism and the very reason why moral realists object to these positions.

Of course, humor is only an element in just one study on folk metaethics (i.e., Sarkissian et al., 2011), so such concerns have little impact on the overall external validity of folk metaethical research.

Even so, the lack of experimental realism present in Sarkissian et al.'s study hints at a more general concern that participants may not be especially engaged with more mundane experimental stimuli. However, I do not believe experimental realism is a major threat to the external validity of most folk metaethical research. The primary purpose in raising objections to Sarkissian et al.'s study is that it stands out among much of this research for purportedly demonstrating that people may shift more towards antirealism (*relativism* in particular) the more salient that cultural differences become, which hints at the possibility that under the some circumstances that many (or most) people would endorse moral relativism, and purports to show that the activation of distinct psychological processes may differentially favor different metaethical positions under the right circumstances. Specifically, Sarkissian and colleagues suggest that people are not rigidly committed to realism or antirealism with respect to particular moral issues, but rather that their willingness to endorse realism/antirealism shifts in accordance with the degree to which they "engage with radically different perspectives," leading Sarkissian et al. to conclude that:

> Future research might proceed not by asking whether 'people are objectivists' or people are relativists' but rather by trying to get a better grip on the different psychological processes at work here and the conflicts and tensions these processes can create. (pp. 501-503)

This is an interesting and novel proposal that is well worth exploring. It is plausible that the output of different psychological processes could favor different metaethical judgments on the assumption that people have (or are readily capable of) metaethical judgment. Yet it is precisely *because* it is novel that I am singling it out for criticism. If the methods used in any one study are sufficiently flawed that they do not provide good evidence for a particular hypothesis, it is still possible that other studies do provide evidence for that hypothesis. Yet Sarkissian et al.'s proposed account is only supported by their findings. If these findings do not support their conclusions, then no findings (currently) do.

Unfortunately, poor experimental realism offers a plausible alternative explanation for their findings. But just what are those findings? Sarkissian et al. presented participants with three conditions:

*(1) Same-culture*

*(2) Other-culture*

*(3) Extraterrestrial*

In the *same-culture* condition, participants considered a moral disagreement between two people from their own culture.

In the *other-culture* condition, participants considered a moral disagreement between a member of their own culture and someone from an isolated tribe in the Amazon with a culture very different from their own. Like the extraterrestrial condition, the stimuli are somewhat humorous, see **Figure S2.2**:

**Figure S2.2**

*Other culture condition from Sarkissian et al. (2011)*



This looks like an image from a *Dungeons & Dragons* manual. Finally, in the *extraterrestrial* condition, participants were asked to consider a moral disagreement between a member of their own culture and

a Pentar, a member of an alien species whose primary goal is to maximize the number of equilateral pentagons in the universe (p. 488). Participants were asked to judge on a 7-point Likert scale how strongly they agreed or disagreed that, since the two people "have different judgments about this case, at least one of them must be wrong."

Across four studies, Sarkissian et al. consistently found that agreement with objectivism was highest when the two people were from the same culture, intermediate when they were from different cultures, and lowest when one of the people was a Pentar. That is, as cultural distance increased, "relativist" responses showed a corresponding increase as well. These findings are consistent with Sarkissian et al.'s explanation. Perhaps it really is the case that when we consider different cultures, this activates psychological processes that prompt us to engage with "radically different perspectives and ways of life," (p. 501) and that this, in turn, inclines us towards a relativist view of morality. Yet there are at least two plausible alternative explanations of these findings.

First, some versions of antirealism (e.g., cultural relativism) are consistent with judging that there are circumstances in which disagreements do have correct answers. This could explain why there was stronger agreement that at least one person was mistaken in the same-culture condition: a subset of antirealist participants who endorse cultural relativism may have judged (correctly, given their view) that if two people from the same culture disagreed, at least one would be mistaken. These participants would be antirealists, yet they'd be mistakenly classified as realists given the measures Sarkissian et al. use. This is a serious flaw with the method they used. Since Sarkissian et al.'s response options cannot disambiguate different forms of antirealism, and their response options require some realists to answer in a way opposite to others, the response options they use cannot serve as a valid measure of realism/antirealism (nor their own distinction between objectivism and relativism, which if anything their stimuli are even *less* capable of distinguishing). Their measures are independently invalid for this reason alone.

However, the focus of my objections in this section concerns the second alternative explanation. Experimental realism can present a general threat to the external validity of an entire set of experimental stimuli. Yet in some cases, there can be variation in the experimental realism of different portions of a given set of stimuli, such that some stimuli have higher experimental realism and some stimuli have lower experimental realism. If so, participants may be more engaged with more realistic stimuli than less realistic stimuli. As a result, any differences between these conditions could be due to differences in the degree of experimental realism of the stimuli. Applying this reasoning to Sarkissian et al.'s study, it could be that the same-culture condition has (comparatively) high experimental realism, the other-culture condition has (comparatively) intermediate experimental realism, and the extraterrestrial condition has (comparatively) low experimental realism. It is possible, for instance, that there is nothing humorous about the same-culture condition, while the Mamilon condition is slightly humorous and the Pentar condition is very humorous. If humor is a threat to experimental realism that reduces engagement and alters the pattern of responses that would be present in its absence (because it e.g., activates different psychological processes than nonhumorous stimuli), it would represent a *differential* threat across stimuli that *could produce the differences across conditions that we observe*. If so, Sarkissian et al.'s findings would be an artifact of using stimuli that vary in how realistic they are. In fact, their stimuli appear to not only vary in experimental realism, but in mundane and psychological realism as well. In other words, across all three relevant forms of realism, their stimuli all vary in the same direction (that is, increased cultural distance is associated with a concomitant decrease in realism, as discussed in section **S2.11**). This reveals that poor realism is not only a threat to external validity, but it can also be a threat to *internal* validity. Insofar as the hypothesis Sarkissian et al. propose cannot be empirically disentangled from this alternative, it is unclear whether their stimuli provide a valid measure of what it purports to measure, or instead is an artifact of internal variation in the level of realism between different stimuli.

**S.2.12 Lack of external validity**

The disagreement paradigm may also suffer low external validity. *External validity* refers to how well a study's findings generalize to circumstances outside the context of the study (Calder, Phillips, & Tybout, 1982; Campbell & Stanley, 1966; Findley, Kikuta, & Denly, 2021). External validity is not always relevant or necessary for the results of a study to be meaningful or practically useful (Mook, 1983; Stroebe, Gadenne, & Nijstad, 2018). Mook (1983) convincingly argues that the artificial conditions of the lab often pose no threat to the central aims of a study, and may even be an asset, depending on the researcher's goals.[85] However, generalizability is not only relevant to research on folk metaethics, it is the whole point. The purpose of folk metaethical research is to determine the metaethical stances or commitments that characterize everyday moral thought and discourse. If it cannot achieve this goal, then it has failed to achieve its primary purpose. In other words, if responses to the disagreement paradigm cannot tell us what people's metaethical stances and commitments are outside the context of the study, then we cannot conclude things like "most ordinary people are realists" or "folk metaethical pluralism is true." Such claims refer to and are generalizations about people outside lab contexts; we cannot appeal to empirical research on folk metaethics to support such inferences if the data does not generalize to the relevant populations. I argue that there are substantial reasons to doubt the generalizability of folk metaethical research:

---

[85] Researchers may be interested in understanding how a psychological mechanism works, e.g., memory or visual perception. In such cases, their studies need not directly reflect real-world conditions. This is because these studies are designed to test theories or hypotheses about the mechanisms themselves. The goal is not to predict events in the real world by mirroring those conditions in the lab, but to make predictions about events in the real world by generalizing from *the theory tested in the lab* to conditions in the real world. Mook gives the example of studies on how our visual systems adapt to the dark. Participants sat in dark rooms and instructed to stare at a particular location and report whether they see red dots of light. These conditions are unlike anything we experience in everyday life. Yet the controlled conditions of the lab allowed researchers to develop a better understanding of how the visual system itself worked, which in turn provided insight into the real world. As Mook observes:

"How then do the findings apply to the real world? They do not. The task, variables, and setting have no real-world counterparts. What does apply, and in spades, is the understanding of how the visual system works that such experiments have given us. That is what we apply to the real-world setting—to flying planes at night, to the problem of reading X-ray prints on the spot, to effective treatment of night blindness produced by vitamin deficiency, and much besides." (p. 385)

(1) **WEIRD populations**: Most studies sample a narrow and unrepresentative body of participants drawn almost exclusively from WEIRD populations (Henrich, Heine, & Norenzayan, 2010)

(2) **Stimulus-as-fixed-effect fallacy**: Most studies make inferences about the moral domain on limited and nonrandom stimuli that are treated as though they were randomly selected but were not, and which may not represent morality as a whole (Judd, Westfall, & Kenny, 2012)

(3) **Lack of ecological validity**: Findings elicited in experimental context may be subject to performance errors, biases, misunderstandings, and demand characteristics, or other factors that result in a pattern of response that does not reflect the kinds of moral judgments that occur in ecologically valid contexts that accurately reflect what metaethical stances and commitments are supposed to be about (Navarro-Plaza et al, 2020; cf. Holleman et al., 2020; Lewkowicz, 2001)

Any one of these issues would be enough on its own to challenge the external validity of the disagreement paradigm. Taken together, I contend that they represent an insurmountable case that current findings tell us little about what ordinary people, from college students to Bajau free divers to long-dead Roman legionnaires think (or thought) about the nature of morality. Here, I discuss only the first two of these.

### S2.12.1. Overreliance on WEIRD populations

By now, Henrich et al.'s (2010) article, "The weirdest people in the world?" has reached such wide circulation that few psychologists are unaware of its general thesis. At the time of writing, it has been cited 10,626 times![86] Their thesis is simple:

(1) Most psychological research relies on samples of people who live in "Western, Educated, Industrialized, Rich, and Democratic," or *WEIRD* societies.

(2) Researchers presume that these samples are sufficiently representative of humanity as a whole to generalize to our entire species.

(3) However, people from WEIRD societies are especially psychologically unrepresentative of humanity as a whole.

---

[86] As of July 18, 2022.

(4) Thus, researchers are not justified in uncritically assuming that findings drawn from WEIRD populations generalize to humanity as a whole.

There is little reason to relitigate the case for (1) and (2). There is no credible objection to either, but Henrich et al. summarize the results of Arnett's (2008) analysis of studies published in top tier journals across six areas of psychology that leaves little doubt about (1):

(1) 68% of participants were from the United States

(2) 96% were from Western nations

(3) 73% of the lead authors were at US universities

(4) 99% were of lead authors were at Western universities

One may quibble with these results, but there is little reason to believe the landscape has dramatically changed in the past decade, and little reason to believe that an analysis of a different set of studies in a different set of journals would reveal a wildly different picture. Not only are the vast majority of participants WEIRD, they also tend to come from especially wealthy and industrialized nations, including the United States, Canada, Australia, New Zealand, and Israel. Yet many participants are drawn from an even thinner slice of humanity. Arnett's findings revealed that two thirds of US participants and 80% of participants outside the US were undergraduates in psychology courses (p. 604). Undergraduates who happen to be taking psychology courses may be even less representative of humanity, since their age, socioeconomic status, differ from and are more restricted than those drawn from WEIRD populations as a whole. Beginning in the mid-2000s, researchers began transitioning away from an almost exclusive reliance on student participants, as more researchers conduct surveys on online platforms such as Amazon's Mechanical Turk and Prolific. However, such findings are still primarily confined to WEIRD populations and the paid workers who participate in these studies exhibit their own distinct cluster of demographic traits that differ in some ways from both WEIRD and non-WEIRD populations. Standard use of these services is thus no substitute for genuinely representative cross-cultural research.

I doubt I need to convince most readers that psychologists frequently treat their findings as evidence about people in general. However, they are rarely explicit about doing so. As Henrich and colleagues observe:

> A typical article does not claim to be discussing 'humans' but will rather simply describe a decision bias, psychological process, set of correlations, and so on, without addressing issues of generalizability, although findings are often linked to 'people.'

Discussion about the generalizability and limitations of findings may be on the rise, but the behavioral sciences have yet to enter a Renaissance of cross-cultural research or coordinated efforts to gather data from representative samples from diverse populations. It remains an unfortunate shortcoming of psychological research that it continues to be conducted by researchers from WEIRD universities on WEIRD populations, and often a narrow subpopulation of WEIRD people (i.e., undergraduates taking psychology courses), and researchers continue to presume that findings among these populations capture universal features of human psychology.

This would be irrelevant if folk metaethical research were an exception to this trend. Unfortunately, it is not. Almost all of the participants in folk metaethical research sample participants from WEIRD nations. However, like most psychological research, the vast majority of studies were conducted on student samples and people from WEIRD societies.

There would also be little reason to worry if we could be confident that folk metaethical findings generalized to non-WEIRD populations, but there is little justification for such confidence. An overwhelming array of studies reveal that people from WEIRD societies are psychological outliers (Henrich et al., 2010). Not only do they not represent humanity as a whole, they are one of the *least* representative populations, since they exhibit a cluster of traits that set them apart from most other societies. Why are WEIRD societies so peculiar? Henrich makes a compelling case that the divide between WEIRD populations and the rest of the world may be due to a revolution in education, literacy, technology, wealth, and social and political institutions that dramatically and rapidly

transformed WEIRD societies. As a result of these cultural changes, people living in WEIRD societies developed an unusual psychological profile that differs from most other societies. The differences don't end there. As Downey (2010) sarcastically states "people in industrial societies are JUST LIKE hunter-gathers," before continuing:

> [...] except for the gigantic scale, anonymous interaction, replacement of reciprocity-based relationships with market transactions, and the unprecedented-in-human-history levels of material inequality. (For the slow readers, yes, that's irony.) Oh, and the domestication of plants and animals, sedentary settlements, high technology, extended classroom education, mass media imagery, enormous social institutions, changes in family structure, decrease [sic] parent-infant contact, radically new built environment, completely different, dense social structure…

Recent human evolution occurred primarily in small scale societies, among ethnically homogenous and comparatively culturally, economically, and geographically isolated communities. People tended to live in tight-knit family groups. Although they are in decline, some of these societies exist today. And while many societies have begun to modernize, they have preserved at least some of the cultural legacy of their (and ultimately *our*) ancestors. Unfortunately, nobody from such societies has been the subject of research on folk metaethics. As such, we have, at best, a handful of studies that evaluate folk metaethics among somewhat less-WEIRD societies, which may or may not provide much insight into how people outside the ambit of WEIRD cultural forces would think about the nature of morality.

Even so, there are a handful of exceptions to the lack of folk metaethical research on non-WEIRD populations. Beebe et al. (2016) conducted cross-cultural studies using the disagreement paradigm, and found a similar pattern of responses in China, Poland, and Ecuador. Sarkissian et al. (2011) also replicated their findings among participants in Singapore. Finally, Yilmaz and Bahçekapili (2015a; 2018) and their collaborators (Yilmaz et al., 2020) have conducted several studies on participants in Turkey. Beebe and Sackris (2016) also surveyed a large number of participants ranging from ages 12 to 89. Their explicit goal was to correct for an overreliance on studies with a restricted range of ages, and to explore variation in folk metaethics across the lifespan.

Efforts to sample demographically diverse populations are laudable, and findings that draw on participants outside the traditional WEIRD populations go some way in allowing us to generalize towards broader populations. However, these studies are a small fraction of research on folk metaethics. Even if they were representative of humanity as a whole, they may represent too small a body of literature to make confident judgments about the metaethical stances and commitments of people in general. Yet even a small but genuinely representative set of studies could be extraordinarily informative about the degree to which a finding generalizes. High-powered replications in a handful of diverse communities would go a long way in providing evidence of a general pattern of psychological phenomena. However, the main shortcoming with demographically diverse folk metaethical research is that the participants in these studies are not especially diverse or representative of humanity as a whole. Recall that WEIRD reflects a cluster of traits: Western, educated, industrialized, rich, and democratic. Societies and the individuals that compose them do not exhibit all or none of these qualities. People and nations differ. While it *may* be that the distinct convergence of these five traits leads to a distinctly idiosyncratic psychological profile, it is also likely that each one of these traits, taken in isolation or interacting with one or more (but not all) of the other qualities, likewise exert an influence over people's psychology in ways that limit how representative those individuals are of humanity as a whole. Wealthy people may differ from less wealthy people. People living in industrialized urban centers with high population density may differ from people living in rural areas. A highly educated person may differ in important ways from less educated people, and so on.

Most of the participants in more demographically diverse folk metaethics samples likely exhibit one or more of the WEIRD traits. For instance, Beebe et al. (2016) report that their sample of participants in China, Poland, and Peru were all drawn from major metropolitan areas. Many of these participants were likely enculturated in comparatively industrialized regions, which already unifies

them along one of the WEIRD dimensions and distinguishes them from people living outside densely-populated urban regions.

Sarkissian et al. provide a non-Western sample by surveying students at the National University of Singapore. However, their sample consisted exclusively of college students enrolled in a philosophy course. Singapore is a wealthy nation with a higher per capita GDP than many WEIRD nations, including Australia, New Zealand, and Canada. It is the second most densely populated nation in the world[87] (Population Density, n.d.) and ranked 11th by the UN's Human Development Index in 2020 (United Nations Development Programme, 2020). Most importantly, Singapore was a sparsely populated region boasting no more than a thousand people (Swee Hock, 2012, p. 8; Rahim, 2010, p. 24). This changed only after the British established it as a trading port following the arrival of Sir Stamford Raffles in 1819 (Buckley, 1902; p. 154). Aside from Japanese occupation during World War II (Huff & Huff, 2020), Singapore remained under British dominion and a British influence has persisted to this day. In the post-war period, Singapore was declared a Crown Colony that remained under at least nominal British rule until 1963 (Abshire, 2011). Singapore endured nearly 150 years of British occupation, and its entire modern population and infrastructure were shaped by British influence in the region. English is the lingua franca in Singapore, and is the primary language used in education, business, and law (Goh, 2017). Given that the British are one of the most paradigmatically WEIRD populations of all, Singapore is not an ideal candidate for a non-WEIRD population to sample from if one's interest is in identifying universal features of human psychology that bridge the divide between WEIRD and non-WEIRD populations, especially if one does so by focusing on a student population at an English-speaking university. I draw attention to all these details to highlight one of my central issues with efforts to mitigate the WEIRDness of sample populations: such efforts

---

[87] Singapore appears third on The World Bank list, however, this is because the list includes Macau. Macau is a Special Administrative Region (SAR) of the People's Republic of China, and is not technically a country.

are often half-hearted, drawing largely on populations due to convenience and not due to how distant they are from WEIRD populations. While it is nevertheless commendable to gather *any* culturally diverse data, when such efforts are made, they may give the misleading impression that we've reached genuinely non-WEIRD populations when we haven't. Simply because Singapore isn't paradigmatically WEIRD doesn't mean that WEIRD influences haven't irrevocably changed the way its population thinks and speaks in ways that would be impossible to assess. We simply don't have access to the alternative history where the population of Singapore grew under its own auspices, without the influence of British colonialism. For instance, Singapore was originally occupied by the *Orang Laut* (Swee Hock, 2012, p. 7). Had they grown into a booming metropolis outside the orbit of the East India Company, never spoken English, and never adopted British customs or institutions, would they respond to questions the same way participants did in the sample we actually have? I have no idea, and unfortunately, neither does anyone else.

Students at NUS represent an especially elite and unrepresentative population, as well. NUS is one of the top-ranked universities in the world, and is therefore likely to attract an unusually well-educated, affluent, and ambitious student population that may be especially unrepresentative of humanity, even compared to other universities. I do not have data on the specific characteristics of students who take philosophy courses at NUS, but it is not a stretch to imagine that such students differ in significant ways from a population genuinely representative of variation in socioeconomic status, cultural background, and other characteristics of potential psychological relevance. NUS is a highly international university, with over 25% of its students hailing from over 100 different nations. While this may reflect a culturally diverse population, some of the students in the sample may have been from WEIRD nations. If that were not enough, the study was conducted in English and English is the language used for instruction at NUS (Sarkissian, personal communication). Such students plausibly watch movies and shows from Western nations, listen to Western music, and otherwise have

had substantial exposure to institutions, concepts, and ideas that overlap with or directly originated from WEIRD societies. NUS undergraduates are an interesting population worthy of study, but findings among such students may tell us little about how people in small scale societies think about metaethics, and such findings may not even tell us much about how East Asian or Southeast Asian people generally think about metaethics, since philosophy students at NUS are unlikely to be especially representative of these populations.

In other words, the students taking these courses may be culturally distinct in some ways, but they are not culturally *isolated* from WEIRD populations, having both significant exposure to and exhibiting many qualities of people from WEIRD societies. Such populations are undoubtedly appropriate subjects of research, but they still represent a narrow body of participants that overlap in many ways with WEIRD populations. Along a continuum of human diversity, Students from NUS may even be psychologically more similar to WEIRD populations on average than they are to people from indigenous foraging, horticultural, and pastoral societies that populate regions far from urban centers, such as indigenous communities in Australia, Brazil, or New Guinea (Henrich et al., 2010). Furthermore, these were students in *philosophy* courses. Students in philosophy courses may be especially unrepresentative of ordinary people, since they may be especially likely to have a prior interest in and exposure to philosophical concepts and ideas, and in particular *Western* philosophical concepts and ideas.

What about other efforts to gather data among non-WEIRD populations? Yilmaz and Bahçekapili (2015a; 2018), along with their colleagues (Yilmaz et al., 2020), conducted a series of folk-metaethical studies among participants in Turkey. However, most of their participants were undergraduates. Y&B (2015a) first set of studies consisted exclusively of undergraduates across all of their studies, while only the last of three studies recruited non-student participants. Yilmaz et al. (2020) did not rely on student participants, however, instead opting for a participant pool composed of

people approached on the streets of Istanbul. More importantly, they conducted the same study on American MTurk workers with the explicit purpose of comparing WEIRD and non-WEIRD populations using the same measures.

All of these studies represent a genuine departure from an exclusive reliance on WEIRD populations. Nevertheless, their findings still provide only a partial picture of humanity as a whole. While Turkey may be regarded as a non-WEIRD society, the WEIRD/non-WEIRD distinction is not a discrete, all-or-nothing distinction, but a continuum. While Turkey may fall on the non-WEIRD side of the continuum, it is not *that* non-WEIRD. Turkey scores highly on the Human Development Index (HDI) at 0.820 (ranked 54th in the world, United Nations Development Programme, 2020), and is labeled as a Newly Industrialized Country (NIC), indicating that economic growth and urbanization outstrip other developing countries (Doral, 2010). As such, it is at an intermediate stage between the most and least developed countries. Turkey also has a long history of exposure to Western culture, including explicit internal efforts to Westernize (Çağaptay, 2014), and this may have influenced Turkish culture. For instance, individualism is one of the most distinctive traits of Western societies. Yet Turkey has a similar degree of individualism as Brazil (Muthukrishna et al., 2020). While Henrich et al. (2010) confine their conception of WEIRD societies primarily to Northwestern Europe and the Anglosphere, Brazil nevertheless falls within the Western world, broadly construed, e.g., the primary language is Portuguese and the most common religions are Roman Catholicism and various Protestant denominations (Office of International Religious Freedom, 2021; Stuenkel, 2011). And while Hofstede's (2001; n.d.) cultural dimension theory identifies a number of other characteristics that distinguish cultures from one another. While there are undoubtedly significant cultural differences between Turkey and Brazil, they score similarly with respect to all of Hofstede's dimensions. Muthukrishna et al. also developed a method for measuring the cultural distance between different societies which they call the *cultural fixation index* (CFI). While Turkey may be distinct in some ways

from Western nations, it is *also* distinct from many other nations. For instance, Turkey exhibits a similar cultural distance from the United States as it does from China. And while some of these differences may capture a qualitative distinction from both WEIRD and non-WEIRD societies, at least some reflect a more intermediate position between them.

While the CFI is only a single metric, it points to a more general point about cultural differences. Even if we can identify society as non-WEIRD or at least less WEIRD, any particular non-WEIRD society may be culturally distinct from other non-WEIRD societies just as it is distinct from WEIRD societies. Replicating results in both WEIRD societies and Turkey does provide *some* evidence that a given trait may generalize to humanity as a whole. But it does not provide *decisive* evidence, since the conjunction of WEIRD societies and Turkey is only somewhat more representative of humanity than WEIRD societies alone.

Finally, measures of cultural distance take into account the entire population of a society, which can obscure differences between individuals and subpopulations. While Turkey is not a WEIRD society, the Turkish people who participate in psychological studies are likely to be disproportionately skewed towards being closer to WEIRD societies than the nation as a whole. Thus, the *specific* populations surveyed in Turkey may be biased towards people most culturally similar to WEIRD populations.

The distinction between analytic and holistic cognition is one of the most prominent and well-established cross-cultural psychological differences (Choi, Koo, & Choi, 2007; Nisbett, 2004; Nisbett et al., 2001; Norenzayan et al., 2002; cf. Chan & Yan, 2007). Anglophone and East Asian societies reflect the respective polar ends of this continuum, but national tendencies can obscure local differences within nations. Uskul, Kitayama, and Nisbett (2008) propose that communities that depend on social interdependence are most likely to rely on holistic cognition, while those that favor independence and self-reliance are more likely to favor analytic cognition. They found evidence of

these differences *within* Turkish communities. The success of farming and fishing communities depends on mutual cooperation and social harmony, while herding communities call for more individual judgment, since herders are often isolated and less reliant on one another. Uskul et al. found that Turkish herding communities were more analytically inclined, while fishing and farming communities were more holistic. These findings demonstrate that psychocultural differences exist within nations and not just between them, including precisely the differences that distinguish WEIRD from non-WEIRD societies. How many members of Turkish herding, farming, and fishing villages were included in folk metaethical research on Turkish people? I am not sure, but it is plausibly few or none. Participants were college students, MTurk workers, or people on social media such as Twitter and Facebook.

How representative are these participants of Turkey as a whole? It is unclear. Perhaps they are representative of most Turkish people, but the exclusion of people from subcultures or entire nations that are more difficult to access points to a more general shortcoming with much cross-cultural research, including all cross-cultural research in folk metaethics. Most of these studies rely on convenience samples that are drawn from members of communities that are the most readily accessible, such as college students, people in densely-populated urban centers, and people who are comfortable enough with technology to participate in online surveys. However, such people may be disproportionately likely to be wealthy and educated, to live in more industrialized regions, and to have exposure to Western influences. In other words, most of the samples drawn from non-WEIRD societies nevertheless sample the most WEIRD people available because WEIRDer people are more accessible. In other words, the people who participate in cross-cultural research not only tend to be less representative of members of their nations and societies than were we to truly randomly sample people within a nation's boundaries, they are all unrepresentative *in the same way*. As a result, the most accessible participants are the *least* divergent from WEIRD societies *and from each other*.

An overreliance on more accessible participants weakens the degree to which empirical findings are truly representative of humanity as a whole. Worse still, humanity is no more appropriately characterized by people alive today than people who lived in the distant past. As globalization continues, cultural differences are dissolving, and fewer societies remain untouched by education, wealth, and industrialization, Western culture and the adoption of democratic governance. Psychologists cannot readily study the minds of the deceased. But as cultural, economic, and political divides close, and as traditional indigenous communities integrate or dissolve, we will have little choice but to study an increasingly psychologically homogenous pool of participants. We may even be witnessing the tail end of the long, slow death of genuinely divergent cultures. Whatever psychological differences result from cultural, economic, and technological differences, they were at their peak prior to the advent of trade networks, trade languages, and transportation. Researchers interested in how *human* psychology should keep this in mind. How many cultures, replete with terms, concepts, and ways of thinking wildly divergent from contemporary societies have left us with nothing but potsherds or the ashes of campfires? Psychologists should recognize that generalizations about how humanity is are radically contingent on the prevailing cultural, technological, and ecological conditions in which humans live. Findings about the psychology of humans in their current state are not findings about humans as they *could be*, and we should be wary of presuming that, in the absence of culturally and psychologically divergent populations, that humans naturally converge on or are innately predisposed to exhibit a particular psychological profile. While this *could* be the case, the psychological uniformity of the world's population could result from an entrenched but contingent cultural uniformity. Globalization and the far reaching effects of missionaries, colonialism, trade, emigration, and telecommunication has radically altered the world. Almost no remaining populations remain uncontacted or outside or fully outside the cultural influence of global powers. The age of cultural isolation is, for better or worse, effectively over. What few uncontacted and thus culturally

uncontaminated populations are unlikely to participate in many studies. Yet even if we could reach such populations, we face an insurmountable hurdle: the very mindset that would prompt any of us to conduct such research, and the entire set of institutions involved in its publication, is thoroughly suffused with WEIRD influences. Anyone conducting such research is likely to be from a WEIRD population or heavily influenced by one, to study and write in English, to interact with colleagues who mostly speak English and mostly interact at conferences and universities in WEIRD nations, to publish their articles in English in a journal hosted in a WEIRD nation with WEIRD editors and WEIRD reviewers. While the *participants* may not be from a WEIRD population, every other element of the research, root and branch, is utterly WEIRD. Whatever biasing and narrowing influences this has on what hypotheses are pursued, how questions are framed, how data is analyzed, how data is interpreted, how data is presented, how data is received, and so on will all be inescapably influenced by WEIRD cultural influences. We are trapped within a cultural paradigm whose influence is pervasive, inescapable, and unknown. While this need not lead to a descent into a deep postmodern skepticism about our ability to know anything, which is *not* what I am advocating, it should serve as point of caution about generalizing to humanity as a whole, especially when one seeks to draw conclusions about human nature and not merely to a universal but transient way that contemporary humans happen to be. Whatever cross-cultural convergence we find among humans *now*, this does not necessarily entail that this is how humans had to be, that such traits aren't contingent products of enculturation, or that under different environmental conditions or given a different historical trajectory humans could have been radically different. I am also not expressing radical anti-nativism about human psychology. Far from it, I have a background in evolutionary psychology (see e.g., Liddle, Bush, & Shackelford, 2011) and am opposed to reflexive opposition to dismiss or stifle legitimate evolutionary hypotheses about human psychology. I am simply arguing that justifying any generalization about humanity based on evidence about how any given populations happen to be is

extremely difficult, and fraught with a variety of epistemic and methodological hurdles that are at best difficult to overcome.

Setting aside the lack of cultural diversity in folk metaethics research, we may still wonder how likely we'd be to find significant cultural differences. Henrich and colleagues provide evidence that people from WEIRD societies are unrepresentative of the rest of the world's population in a variety of ways, including spatial cognition and visual perception. However, many of the core psychological differences between members of WEIRD populations and the rest of the world center on individualism. Henrich emphasizes that *individualism* and a cluster of associated traits are central to the differences between WEIRD and non-WEIRD societies. Members of WEIRD societies are more "individualistic, independent, analytically-minded and impersonally prosocial," and exhibit less "conformity, obedience, in-group loyalty, and nepotism" than non-WEIRD societies (Schulz et al., 2018). Critically, many of these differences are rooted in more fundamental differences in social structure and kinship relations. According to Schulz and colleagues, differences in the "social norms, social networks, technologies and linguistic worlds" that people "encounter while growing up" all play a significant role in shaping differences in the "motivations, emotions, perceptions, thinking styles and other aspects of cognition" between WEIRD and non-WEIRD populations. Critically, many of the ways WEIRD societies have ultimately diverged from non-WEIRD societies center directly or indirectly on traits and behaviors relevant to morality. According to Henrich (2022), success in the precursors to contemporary WEIRD societies involved:

> the cultivation of greater independence, less deference to authority, more guilt, stronger use of intentions in moral judgments, and more concern with personal achievement. Success became less bound by tradition, elder authority, and general conformity. WEIRD individuals have to "sell themselves" based on their personal attributes, specialized abilities, and dispositional virtues, not primarily on their friendships, lineages, or family alliances.

In contrast, more traditional, non-WEIRD societies differ in their greater emphasis on kin-based institutions, which encourage norms that "reward greater conformity, obedience, holistic/relational

awareness and in-group loyalty but discourage individualism, independence and analytical thinking," which in turn reduces "people's inclination towards impartiality, universal (non-relational) moral principles and impersonal trust, fairness, and cooperation" (Schulz et al.). In other words, the underlying causal factors that gave rise to the psychological differences between WEIRD and non-WEIRD populations have a direct and pervasive impact on our respective moral psychology. Such differences could have implications for folk metaethics, as well. Members of weird societies may be more prone towards a broad, cosmopolitan, egalitarian moral outlook that includes a broader array of subjects in their ambit of moral concern, which could plausibly be associated with a greater tendency towards relativism or other forms of antirealism.

Yet the psychological differences between WEIRD and non-WEIRD populations do not end there. In a review of the moral differences between WEIRD and non-WEIRD societies, Graham et al. (2016) show that, unsurprisingly, members of WEIRD societies are more concerned with autonomy and individual rights, while members of more collectivist societies tend to express greater concern with spiritual purity and moral duties and to the community. People in less WEIRD societies are also more likely to take contextual information into account when judging whether it is morally permissible to push someone off the bridge in the bridge dilemma, e.g., given their (hypothetical) social status in the scenario, are they in a position where they are entitled to make life-or-death decisions? As Graham et al. observe, "This relational consideration in turn leads to less admonishment of individuals who do not flip the lever, and fewer character attributions of actions made in the absence of their broader contextual meaning" (p. 126). In contrast, people from WEIRD societies are more likely to rely on abstract moral principles that ignore social information about the relationships between the people in a situation. To someone from a WEIRD society, such relations are less important (or even irrelevant) for judging the moral status of an action, while people in less WEIRD societies may consider such information important, or even necessary for judging the moral status of an action.

Significant differences can also be overlooked if we are not careful to evaluate cultural dissimilarities at the appropriate level of analysis. While people from both WEIRD and non-WEIRD societies may endorse abstract moral concepts such as *justice*, what *counts* as justice may differ. People in WEIRD societies tend to regard distributive justice as a matter of apportioning resources on the basis of effort or desert, even at the expense of the less well-off. In contrast, non-WEIRD societies that exhibit greater collectivism are more likely to focus on equality of outcome. Rather than continue to provide an exhaustive list of examples, however, I will simply emphasize that these findings point towards the likelihood of numerous potential differences between WEIRD and non-WEIRD societies, differences that we are only beginning to reveal as psychologists respond to the growing need for more culturally diverse non-WEIRD research participants.

Preliminary efforts to explicitly address the lack of inclusion of culturally diverse participants have already revealed intriguing hints about the future of cross-cultural moral psychology. For instance, Berniūnas (2020) reports a series of studies which suggest that associations with the term "moral" do not readily translate to Chinese and Mongolian societies. Although both societies have developed translations of the Western term "moral," *daode* 道德 (Chinese) and *yos surtakhuun* (Mongolian), Berniūnas found that Chinese and Mongolians did not associate these terms with the same prototypical violations we would typically take to be exemplars of the moral domain. As Berniūnas observes, when Americans, Canadians, and Australians were asked to provide examples of immoral actions, they "overwhelmingly produced lists with harm and fairness transgressions" (p. 62) In contrast, when Chinese participants were asked to provide examples of *bu daode* 不道德 they tended to focus on behaving in an *uncultured* way, providing examples such as littering and spitting. Mongolians likewise provided examples such as littering and spitting, but also tended to emphasize respect. As Berniūnas, the way other cultures think about normative issues involves a rich panoply of concerns

that are culturally distinct, highly complex, and diverge in emphasis from the concerns central to Western conceptions of morality:

> [...] it seems that Mongolians are mostly concerned with the issues related to respect (*khündlekh*), which could mean many things. For instance, beside respect of others, respecting natural environment [sic] by not polluting, cleaning it (*orchindoo tseverkhen baikh*), could also be included. Then, there is a rather culturally specific notion of not being a burden or a nuisance to others, especially to parents (*gai/saad bolokhgüi*). Also, culturedness (*soyoltoi baikh*) appears as a general requirement, and in the *yos surtakhuungui* list it shows as concrete actions such as spitting (*nulimakh*), littering (*khog*) and cursing (*kharaal ügs*). (p. 64)

These findings dovetail with recent critiques of the notion that all societies draw a shared distinction between moral and nonmoral norms (Stich, 2018; Machery, 2018). This research is only just beginning, but there are already intriguing hints that the very notion of "morality" may *itself* represent an idiosyncratic, *sui generis* domain distinctive to WEIRD societies, and that, while other societies undoubtedly exhibit significant overlap in their normative concerns (e.g., with promoting cooperation and prosocial behavior, minimizing harm, and so on), they do not even engage in distinctively moral thought or judgment at all, or at least not do so in a way that resembles WEIRD conceptions of morality. If the moral domain is not culturally universal, and some societies lack distinctively moral terms and concepts, it would be implausible to presume that their judgments and discourse are nevertheless underwritten by an implicit commitment to distinct meta*ethical* positions. Such populations may have meta*normative* positions, but we are not entitled to extrapolate from a person's metanormative stances or commitments to their metaethical stances or commitments in particular.

To illustrate why, consider how we might think about *prudential* or *epistemic* norms. In principle, you could consistently adopt a different metanormative stance for each of these normative domains. Such positions may even be more attractive than a uniform realism or antirealism about all normative domains. For example, you could adopt a realist stance towards epistemic norms, an error theoretic stance towards moral norms, and a subjectivist stance towards prudential norms. That is, you could believe there are stance-independent facts about knowledge and epistemic justification claims, that

there are no stance-independent moral facts at all but believe ordinary moral discourse mistakenly attempts to refer to such facts, and believe that prudential norms are best understood as hypothetical imperatives that provide us with stance-*dependent* reasons to act in accordance with our interests. This a la carte approach towards different normative domains is by no means unusual, it is *already* the kind of position contemporary philosophers take. Joyce, the most prominent contemporary error theorist, endorses error theory and fictionalism about the moral domain, but not the epistemic domain (Cline, 2018; Joyce, 2001; 2011a; 2019; 2020). When Joyce makes moral claims, he is explicitly committed to making claims that *do not* genuinely refer to claims about stance-independent moral facts, but he's not a fictionalist about epistemic claims (Joyce, 2020). There is no logical inconsistency in this, and this could in principle be both the most consistent and defensible way of thinking about different normative domains. Nothing about the way we speak or think requires us to adopt the same metanormative stances and commitments for all normative domains. We have no good reason to presume that ordinary normative thought and discourse would require metanormative uniformity across all domains. Yet if people in some cultures have substantially different conceptions of morality, or lack moral concepts altogether, this further undermines the possibility of attributing any consistent metaethical stances or commitments to these people. If the moral domain turns out to be a *sui generis* culturally constructed domain (Machery, 2018), it would make no more sense to ask whether people are realists or antirealists *about the moral domain* than it would make sense to ask whether members of an alien species prefer the Yankees or the Red Sox. While it may turn out that people who lack moral terms and concepts are *normative* realists about certain violations, e.g., they think it's stance-independently "wrong" to disrespect your parents, or throw cigarette butts in the woods, it would not follow that they are expressing a *moral* realist stance towards these transgressions. They could be e.g., a *khündlekh realist* about the wrongness of disrespecting your parents or littering. It would be an error of cultural projection to characterize such a person as being a *moral* realist about such issues, for the

same reason it would be absurd to insist that contemporary moral realists such as Mike Huemer and Russ Shafer-Landau are *khündlekh realists*. Chances are they have no idea what this word means, and even if they'd heard of it, they are not sufficiently enculturated within Mongolian society to have adequately internalized what this would mean, i.e., they might have a superficial understanding of the word, but this wouldn't be sufficient to *grok* it (Rabkin, 1979).

In fact, many concepts central to Western moral philosophy may be absent from the way people in other cultures think about normativity, and this may be true whether or not they possess distinctively moral thought and discourse. Like research on folk metaethics, most research on free will has been conducted in WEIRD populations (91% according to Berniūnas et al., 2020). Yet Berniūnas et al. (2020) have recently pointed out that there are no lexical equivalents of the term *free will* in standard use among native speakers of Chinese, Hindi, and Mongolian. After examining differences in how native speakers of these languages think about translations of the term *free will*, they conclude that their findings "could be interpreted as showing that *free will* is a WEIRD notion" (p. 12). They maintain that while

> […] if Americans and Lithuanians believe in free will, and Chinese, Indians, Mongolians believe in *ziyou yizhi* [自由意志], svatantra icchā [स्वतंत्र इच्छा], chölöötei khüsel, respectively, then they believe in concepts with markedly different content. Indeed, *it could be argued that they believe in markedly different concepts.* (p. 12, emphasis mine)

The concept of free will is at the epicenter of many of the central disputes in contemporary analytic ethics. If non-WEIRD societies don't even share the concept of free will, this highlights a significant and fundamental difference in WEIRD and non-WEIRD populations, and lends support to the possibility that many terms and concepts central to analytic philosophy are culturally parochial.

Since most participants in psychological research are drawn from unrepresentative populations, we cannot confidently generalize towards humanity as a whole. This same limitation applies to research on folk metaethics. Most research on folk metaethics has been conducted in

WEIRD or semi-WEIRD populations, is usually conducted in English, and frequently draws on undergraduate populations. What little research diverges from this general pattern is rarely conducted among highly culturally divergent populations. Instead, such studies tend to include convenient, accessible populations, such as students taking courses with a member of a research team. These student populations are unlikely to be especially representative of the populations they are drawn from, much less traditional indigenous societies and therefore provide only limited generalizability. Existing evidence overwhelmingly establishes that there are significant psychological differences between WEIRD and non-WEIRD populations, and that many of these differences are relevant to moral cognition and behavior. Since research in folk metaethics focuses primarily on WEIRD populations, we know little about the folk metaethics of humanity as a whole. Many studies also draw primarily or exclusively on student populations. Yet as Beebe and Sackris (2016) have shown, people between the ages and often draws on student populations who are *even less* representative of humanity as a whole. In fact, realism drops sharply as people enter their teens, then peaks again as people approach their thirties, before stabilizing among all subsequent age groups. In other words, precisely those participants most likely to participate in research on folk metaethics are the least psychologically representative age group of the populations they are drawn from. In other words, college-age students are both (a) the most likely participants in folk metaethics research and (b) are statistical outliers compared to every other age group. Age-related differences appear to be especially significant, and researchers should keep this in mind when drawing conclusions about how people *in general* think about the nature of morality. Such findings must be qualified by an appreciation for the possibility that children, teens, young adults, and older adults respond to questions about metaethics.

Alarmingly, differences across demographic groups, including both age and culture, may be hampered by inadequate measurement invariance, as well. *Measurement invariance* refers to the degree to which a measure used in one population measures the same construct in another population (Putnick

& Bornstein, 2016; Schoot, Lugtig, & Hox, 2012; Vandenberg & Lance, 2000). Any efforts to conduct cross-cultural research on folk metaethics or to extend paradigms used in one popular to another must be mindful of the risk that stimuli used in one population may be interpreted in a way that systematically differs from another population. Doğruyol, Burak, and Yilmaz (2019) and Iurino & Saucier (2020) have already found that questionnaires designed to test Moral Foundations Theory suffer problems related to measurement invariance, e.g., Doğruyol et al. found that the item loadings for the same measures differed across different culture groups (*metric non-invariance*), prompting them to caution that:

> One of the implications of metric non-invariance is that the MFQ might not be a suitable tool if one aims to compare moral foundations across WEIRD and non-WEIRD cultures, because any difference could be due to the differences in loadings, rather than mean endorsement of a specific moral foundation. (p. 4)

Evidence of problems related to measurement invariance occurring in other research in moral psychology demonstrates that such difficulties can and do emerge in cross-cultural research. Coupled with decisive evidence that moral terms and concepts are understood differently in some cultures (or may even be absent), there are compelling theoretical grounds for suspecting that the threat to measurement invariance for measures of folk metaethics may be very high. Existing findings not only fail to generalize to humanity as a whole, they also face steep methodological hurdles that will prove incredibly daunting for researchers who wish to evaluate folk metaethics in non-WEIRD populations.

**S2.12.2 Stimulus-as-fixed-effect problems**

Another shortcoming with most forms of the disagreement paradigm is that they are subject to the *stimuli-as-fixed-effect-fallacy*. This occurs when researchers treat the *participants* in their study as a random factor, but treat the *stimuli* as a fixed factor. As Baguley (2012) observes, "[b]y treating stimuli as fixed it is assumed that we've exhaustively sampled the population of interest in our study. This limits statistical generalization to those particular stimuli." This often occurs when researchers treat the

specific stimuli used in a study (i) as though the stimuli were randomly selected from the set of possible stimuli that could have been selected in such a way that (ii) these stimuli are therefore *representative* of the relevant range of stimuli, and that, as a result, (iii) mistakenly believe they are entitled to make general inferences about the domain or class of stimuli as a whole.

The reason that (iii) is a mistake is that ignoring "systematic variation between experimental stimuli" has the potential to "contribute to statistically significant mean differences that may not replicate in studies with different stimulus samples" (Judd, Westfall, & Kenny, 2012, p. 54). More generally, it means that our inferences are restricted to the stimuli used in a study. As Baguley (2012) points out: "Any design that restricts the population sampled from (of participants or stimuli) restricts its variability and therefore restricts its generalizability to the pool of participants or stimuli being sampled from."

Judd et al. point out that methodologists have been aware of this problem for decades (Clark, 1973; see also Bonge, Schuldt, & Harper, 1992; Raaijmakers, Schrijnemakers, & Gremmen, 1999; Wike & Church, 1976). Yet efforts to mitigate the problem are rarely adequate, or are simply ignored. For instance, Westfall, Nichols, & Yarkoni (2016) document the tendency for studies employing fMRI to overgeneralize to a broader population on the basis of an impoverished set of stimuli that limit generalizability:

> Most functional magnetic resonance imaging (fMRI) experiments record the brain's responses to samples of stimulus materials (e.g., faces or words). Yet the statistical modeling approaches used in fMRI research universally fail to model stimulus variability in a manner that affords population generalization, meaning that researchers' conclusions technically apply only to the precise stimuli used in each study, and cannot be generalized to new stimuli. A direct consequence of this stimulus-as-fixed-effect fallacy is that the majority of published fMRI studies have likely overstated the strength of the statistical evidence they report. (p. 1)

This illustrates that the problem of the stimulus-as-fixed-efficacy is relevant to contemporary empirical research, and that large bodies of published research have yet to adequately address, or even demonstrate an awareness, of the problem. While researchers typically avoid generalizing based on a

single instance of a given stimuli to represent an entire category of theoretical interest, their typical solution is to "include some reasonable sample of stimuli in order to suggest generalization" (p. 54). Unfortunately, *suggesting* generalization is a far cry from *justifying* generalization, and such methods may be insufficient to warrant the generalizations researchers extract from their findings. This is especially the case when the pool of items is selected in an unprincipled and unsystematic way, e.g., relying on heuristics or intuition, or simply including "enough" stimuli to feel a superficial sense of satisfaction. For instance, suppose researchers wanted to evaluate how much people living in the United States liked eating fruit. They don't want to measure attitudes towards a single fruit, since they correctly recognize that people's attitudes towards e.g., apples won't necessarily tell us much about their attitudes towards other fruits. To get around this problem, they choose the first three fruits that come to mind: apples, peaches, and pineapples, ask people how much they like all three on a 7-point Likert scale, then average these to form a composite "fruit preference" score. Would this provide us with a valid measure of how much people living in the US like fruit? It's unclear. If they had selected a different suit of fruit, we might have obtained very different results. If we wanted to make general claims about how much people like *fruit*, it would be a mistake to think that how much they like apples, peaches, and pineapples *in particular* will be a reliable measure of how much they like fruit *in general* for the obvious reason that the mean score for the fruits researchers selected may not be very similar to the mean score we'd obtain if we exhaustively surveyed a representative sampling of fruit.

Unfortunately, the same problem applies to the majority of research in folk metaethics. Typically, researchers employ either *abstract* or *concrete* questions about moral realism and antirealism. *Abstract* questions prompt the participant to express a realist or antirealist stance about the moral domain as a whole. Since these questions operate over the entire domain of theoretical interest (i.e., the moral domain), they are not subject to stimulus sampling problems, since such stimuli by design

reflect the entire domain of theoretical interest.[88] The FMO (Zijlstra, 2019) and MRS (Collier-Spruel et al., 2019) both rely on abstract questions, and thus do not suffer from this stimulus sampling problems, as do a handful of abstract paradigms employed in other studies (e.g. Pölzler & Wright, 2020a; 2020b).

Standard versions of the disagreement paradigm *do* rely on concrete moral questions. Regrettably, there seems to be no substantive, principled basis for selecting the items used in any prominent studies on folk metaethics. Consider, for instance, the items used by Goodwin and Darley (2008):

1. *Opening gunfire on a crowded city street*
2. *Robbing a bank in order to pay for an expensive holiday is a morally bad action*
3. *Providing false testimony in court about the whereabouts of a friend who is being charged with murder (i.e., to protect that friend by offering an alibi)*
4. *Consciously discriminating against someone on the basis of race*
5. *Cheating on a knowledge section of a lifeguard exam, to obtain a job for which one is not qualified*
6. *Anonymously donating a significant proportion of one's income to charity is a morally good action*
7. *Before the 3rd month of pregnancy, abortion for any reason (of the mother's)*
8. *Assisting in the death of a terminally ill friend who is in terrible pain, and who wants to die*
9. *Scientific research on embryonic human stem cells that are the product of in vitro fertilization*

Are these items representative of moral actions in general? I don't know. More importantly, Goodwin and Darley don't know, nor do we have any good reason to think that they represent the moral domain. It may be tempting to suppose that, because they used nine items in total, that this is "good enough." But it's not good enough. Using numerous items isn't going to resolve the problem of

---

[88] This does *not* mean that if a participant expresses a realist or antirealist stance in response to questions about "moral judgments" that this rules out pluralism. Since participants are forced to respond to a question that expresses a view about *all* moral issues, they have no way to express a different view towards different moral issues. As such, abstract questions may mask underlying differences in how a participant would respond to different stimuli in the same domain. Thus, abstract questions simply trade off one methodological limitation for another, and may fail to solve problems related to generalizing about domains of theoretical interest.

generalizing to other items in the moral domain if the items weren't randomly selected, for the same reason that averaging the heights of nine NBA players isn't going to allow you to make inferences about the average height of people in general: if the items weren't randomly selected, they could be unrepresentative in a variety of ways, and could even be *systematically* unrepresentative.

First, the proportion of participants who expressed a realist stance towards particular moral issues varied dramatically. This same pattern emerged in Wright, Grandjean, and McWhite (2013) and Beebe (2015), both of whom used many of the same items as Goodwin and Darley. Beebe provides a helpful graph of the proportion of realist responses using his version of the disagreement paradigm:

**Figure S2.3**

*Proportion of "objectivism" responses reported in Beebe (2015)*
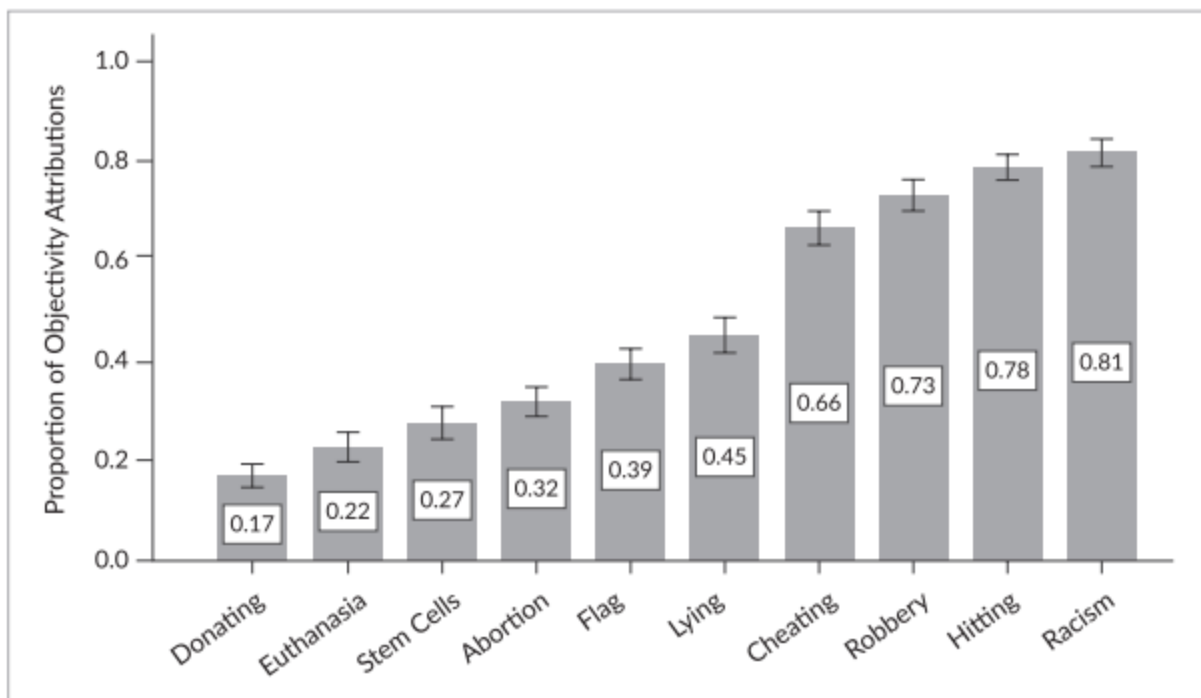


Each of these studies employed slightly different sets of measures and items, but all found the same pattern: the proportion of participants who expressed a realist or antirealist stance varied dramatically across items, from a tiny minority to a significant majority of participants. Given the high variation

between items, we clearly cannot presume that any one item is as good as any other for representing the moral domain as a whole. If we'd only chosen, e.g., donating to charity, we'd have the impression that almost everyone was a moral antirealist, while if we asked about racism, we'd have the impression that almost everyone is a realist. We simply cannot generalize from any single item to the moral domain as a whole. But we also can't generalize to the moral domain as a whole by *averaging* the realism rate for items we happened to use in our study, because we don't know how well the average of whatever items we happened to choose corresponds to the average of the moral domain as a whole.

Unfortunately, *this is exactly what Goodwin and Darley did*. That is, they used the average realism score across all items to make inferences about the moral domain as a whole. In their abstract, they state that, "Experiment 1 showed that individuals tend to regard ethical statements as clearly more objective than social conventions and tastes, and almost as objective as scientific facts" (p. 1349). They reiterate this claim in their general discussion, stating that one of their major findings "[...] was that ethical beliefs were treated almost as objectively as scientific or factual beliefs, and decidedly more objectively than social conventions or tastes" (p. 1359). While both comments are qualified by explicitly acknowledging variation across items (i.e., metaethical pluralism), Goodwin and Darley still claim that, *overall* (or on average), people are more likely to adopt a realist stance towards moral issues than other issues. Given their findings, this claim is not justified. Since we do not know how well the pool of items they use to represent the moral domain represents the moral domain, we cannot make any inferences about people's overall tendency to adopt a realist or antirealist stance towards moral issues.

To illustrate why this inference is not justified, we need simply think of the relevant analog when it comes to selecting participants. Suppose we wanted to identify the favorite pizza toppings of people living in the US. One good way of obtaining a representative sample of the US population would be to randomly select participants. Of course, in practice, you can't achieve genuine randomness

for a variety of practical reasons. So we approximate randomness by, e.g., hiring a polling agency to gather a representative sampling for us, or posting a survey online and hoping the people who fill out the survey are sufficiently similar to the general population. But what if we just chose the first few people we could think of? I know my uncle Steve likes anchovies, and that's weird, so let me ask him. And I know my sister-in-law is really into pineapple on pizza, so maybe I'll ask her. This would, for obvious reasons, be a *terrible* approach. I'd have no good reason to think the people who came to mind would be representative of people in the US.

Yet when it comes to selecting stimuli, researchers seem content to do something at best only marginally more principled than this. Sure, they may not merely select the first moral issues that come to mind. They might instead make some nominal effort to select a handful of moral issues that seem to serve as a decent spread of moral issues. Yet this is based on little more than armchair theorizing and intuitions, and is still highly vulnerable to a form of availability or salience bias: researchers constructing a pool of items to represent the moral domain may be biased towards selecting those items that most readily come to mind, or seem most striking, noteworthy, or otherwise salient (Schwarz et al., 1991; Taylor, 1982; Tiefenbeck et al., 2018; Tversky & Kahneman, 1973). That is, when researchers begin thinking about which moral issues to include in a study, they are likely drawn towards those that most readily come to mind or stand out. Some issues might stand out because they are provocative or trigger a strong emotional response. Others may come to mind because they are e.g., politically salient, or known to be central to ongoing social conflict, such as abortion or euthanasia. Even if researchers recognize this and put effort into including more diverse moral issues, for instance, by including, e.g., a statement about donating to charity, this is still achieved in an unprincipled, top-down, *a priori* method that simply relies on *whatever seems adequate*. In other words, researchers are simply relying on their *intuitions* about which items adequately represent the moral domain. Our intuitions are not a reliable guide for selecting which participants to include in a study; that's why we try to

approximate randomization. Why, then, should we think intuitions are adequate for selecting representative stimuli? The simple answer to this is: we shouldn't. Researchers are simply making an error when they do so. Consider the specific issues Goodwin and Darley chose: four are extremely serious moral transgressions: one describes what sounds like terrorism and attempted (or actual) murder, one describes robbing a bank (which can lead to hostages and deaths) for the trivial purpose of going on a vacation, one involves perjury to protect a potential murderer, and one involves deliberate and explicit racism. Then we have cheating on a lifeguard exam when one isn't qualified, donating to charity, and three controversial issues in biomedical ethics (abortion, euthanasia, embryonic stem cell research). Why should we presume that one positive moral action (donating to charity), three issues in biomedical ethics, and four or five very serious transgressions represent the moral domain as a whole? This list seems almost completely arbitrary. Why aren't participants given an equal portion of positive and negative moral actions? Why are there so many biomedical issues? Are one third of moral issues related to biomedical ethics? Do judgments about biomedical ethics generalize to morality as a whole? I have no idea what the answers to these questions are. It's plausible that the answer to many of these questions is "no." Regardless of what the answers are, we simply have no idea what they are, so why use *these particular items*?

Consider the many issues that might accompany drawing on a pool of items like this. Suppose biomedical issues tend to have higher antirealist rates than average. If so, the inclusion of such items could inflate estimates about the overall rate of antirealism. Very serious moral transgressions could have higher realist rates than average, so including too many of these could lead us to overestimate the overall rate of realism. And so on. Without knowing how well these items represent the moral domain, we are simply not in a position to make precise claims about how people think about morality in general by evaluating how they think about these particular moral issues.

This problem is further confounded by the lack of clarity about what the moral domain even is. What counts as a moral issue? What level of specificity are we referring to? Note that Goodwin and Darley's items include nonspecific instances of third trimester abortion, but provide more concrete details for some moral issues, e.g., robbing a bank with a particular motive, cheating specifically on a lifeguard exam, and so on. Do we have to count cheating on every conceivable type of exam a person could cheat on, and count responses to each of those separately when measuring how people think about issues in the domain? That seems empirically intractable, to say the least.

Researchers are also relying on their own *a priori* assumptions about which issues fall within the moral domain, *even if participants do not consider the issues in question to be moral issues*. Wright, Grandjean, and McWhite (2013) reveals that most participants did not share researcher assumptions about which issues were moral or not, nor did they agree with one another. As such, any attempt by researchers to present participants with the same set of items, all ostensibly "in the moral domain," involves the imposition of a top-down, a priori categorization scheme that participants themselves may not share. While Wright et al. (2013) still found evidence of metaethical pluralism even when focusing only on those moral issues participants themselves classified as moral, this still poses a considerable barrier to making inferences about "the moral domain" since it's still unclear what the moral domain even is, and while a general pattern may persist even when factoring in participants' own classification schemes, researchers would still face considerable barriers in making precise estimates about overall rates of realism or antirealism in the moral domain; this would especially be the case if researchers didn't consistently include the kind of domain classification task Wright et al. used and look only at those issues participants themselves classified as moral. In other words, we could consider "the moral domain" to be whatever participants *themselves* consider moral issues, or we could operationalize "the moral domain" as a set of issues included or excluded from the moral domain based on the *a priori*

categorization scheme of the researchers conducting a given study. Whatever data we gather with respect to one will not necessarily permit justified inferences about the other.

Take Goodwin and Darley's (2008) inclusion of an item related to donating to charity: In their first study, only 11% of participants classified donating to charity as a moral issue. This was, in fact, *lower* than the percentage of participants who judged "Talking loudly and constantly to the person next to you during a lecture" (16%) and the claim that "*Homo sapiens* evolved from more primitive primate Species [sic]" (14%) (p. 5; p. 21). Yet talking loudly during a lecture would typically be classified by researchers as a transgression of social convention, while the claim that *Homo sapiens* evolved would typically be classified as something like a "scientific claim." Should we just ignore how participants themselves view these issues, and how they classify them? I see little justification for doing so. Of course, one might challenge the validity or at least the interpretation of Wright et al.'s method of assessing domain classification. Participants were asked to choose which was the "best fit" for each issue:

> *(i) personal choice/preference*
> *(ii) social conventions/norms*
> *(iii) moral issue*
> *(iv) scientific fact*

There are several problems with this classification scheme. First, these options are not mutually exclusive. Second, "scientific fact" is ambiguous, and inappropriately would suggest that the participant *agrees* with the statement in question. For instance, suppose participants were asked about the claim "the earth is flat." This should fall within the domain of scientific facts, because it is a false scientific claim. Yet even if we set aside every methodological issue with this task, or even ignore these results altogether, we still face the distinct possibility that participants would not classify issues as moral or nonmoral in the same way researchers do. There are already indications that there are substantial demographic differences in how people think about the moral domain (Levine at al., 2021;

Machery, 2018; Berniūnas, 2020; Berniūnas, Silius, & Dranseika, 2022). For instance, Levine et al. (2021) found that there are substantial differences in the way members of different religious groups classify moral issues, e.g., Hindu participants "did not seem to make a moral/non-moral distinction of the same kind" (p. 139). Such findings led Levine and colleagues to "suggest a profound relationship between religious affiliation and conceptions of the scope of the moral domain" (p. 139). We are not entitled to simply presume researchers and participants share a common conception of "the moral domain."

At present, we simply lack adequate descriptive information about how different populations think about the moral domain, whether they even have a concept of the moral domain at all, and whether their normative judgments conform with any particular nomological cluster of characteristics that would justify treating "the moral domain" as an appropriate subject of empirical inquiry. If, for instance, it turns out that there simply is no moral domain (Stich, 2018), or if, at best, the moral domain is a historical invention (Machery, 2018) and there are no principled characteristics that distinguish moral from nonmoral norms (Sinnott-Armstrong & Wheatley, 2012; 2014), then it is unclear whether researchers could readily generalize from how people think about specific moral issues to how they think about moral issues in general. Indeed, it's not even clear what, exactly, what "moral issues in general" would mean. Granted, researchers could simply *stipulate* what they mean, and operationalize their measures accordingly. Yet in practice, most research seems to be oblivious to these concerns, to ignore them, or to at best make half-hearted but ultimately inadequate efforts to address them.

We've meandered far from the central point of this section, which is a simple and straightforward criticism. Simply put, most versions of the disagreement paradigm present participants with a set of concrete moral issues, and measure levels of realism and antirealism towards each of those issues. While this approach permits researchers to make inferences about levels of realism and antirealism towards those specific issues, researchers are *not* justified in generalizing from the average

level of realism and antirealism from the items used in their sample to the average level of realism and antirealism of the moral domain as a whole. The reason for this is simple: since we do not know whether the items used in these studies are representative of the moral domain (and there's no good reason to think they are), we cannot make inferences about the moral domain. This means that all existing studies employing the disagreement paradigm permit at best highly constrained inferences that are confined exclusively to the items used in the studies themselves.

**S2.13 Forced choice obscures indeterminacy**

All versions of the disagreement paradigm present participants with a *forced choice* between a limited set of response options. Currently, almost every version requires participants to endorse either *realism* or *antirealism*.[89] A forced choice is any measure that requires the participant to respond in such a way that they must demonstrate at least some minimal preference for one or more substantive response options indicative of some measure of interest (Lavrakas, 2008, p. 289). This means that participants cannot generate their own open-ended response (by e.g., selecting "other" and elaborating, writing in an empty textbox, or speaking to the researcher and having their verbal report transcribed as the main measure), and they cannot avoid demonstrating a selection preference by choosing a neutral option or explicitly selecting a response option that expresses indifference, lack of knowledge, etc., such as *I don't know* or *I do not wish to answer*.

This is a serious problem if *metaethical indeterminacy* is the correct account of folk metaethics, for a very simple reason: even if people had no determinate metaethical stances or commitments, the structure of all existing studies would *require* them to endorse one or the other, thereby creating the

---

[89] One notable exception is Goodwin and Darley (2008), who provide an "other." Wright, Grandjean, and McWhite (2013) employ an otherwise identical method, but don't mention the use of "other," suggesting that it was dropped. "Other" may be an especially unappealing option, since it is unclear what it means, and it doesn't provide any substantive information about the participant's views. This one minor deviation from otherwise universal tendency for forced choice methods does little to mitigate the problem.

artificial appearance of determinate folk metaethical stances or commitments even where none exist. Of course, participants may be able to simply skip a question, but the fact that a participant can skip a question is a trivial feature of any study design. Rather, forced choice means that any *measurable* response must necessarily conform to a restricted set of categories or response options that necessarily express a preference for one or more substantive measures of interest over others. For instance, if you asked people *which is better: Coca-Cola or Pepsi?* The only interpretable responses to this question would *require* the participant to pick (a) *Coca-Cola* or (b) *Pepsi*. You could also ask them to express how strongly they prefer one or the other on a scale, but deny them a midpoint that experiences no preference. Forced choice paradigms may suffer from a variety of methodological shortcomings, e.g., requiring a participant to select from among a narrow set of options without the ability to opt out of deciding can induce discomfort and prompt the participant to choosing whichever option most alleviates this discomfort, e.g., a "default" option least likely to be perceived as an error or lead to negative evaluations of the participant (Dhar & Simonson, 2003). However, the most serious problem is simply that the participant's actual position may simply not be represented by any available response options, so any response they select will *necessarily* be at least somewhat (if not entirely) inaccurate.

Another reason why this is a serious deficiency with the disagreement paradigm is that even if response options for indeterminacy were included, the mere act of presenting participants with a range of substantive response options may prompt people to select those options even if they have (or held prior to participating in the study) no previously determinate stance. Take, for instance, someone who has never genuinely considered whether they endorse one of two positions. They don't have a view at all because they've *never considered it*. Furthermore, nothing about the way they speak or act would allow us to conclude that their linguistic practices commit them to one of the two positions. Once the question is asked, that person may choose one of the substantive response options (rather than e.g., "I have no idea" or "I have no position on this") for a variety of reasons *other than that this is a genuine*

*reflection of the stances or commitments they had prior to participating in the study*. Response options that don't determinately favor one of the substantive positions on offer may be unappealing for a variety of reasons, e.g., they may suggest incompetence, disengagement, or lack of motivation to comply with researcher expectations. Alternatively, participants may simply form a conclusion in response to the stimuli, a phenomenon I refer to as *spontaneous theorizing* (see **Chapter 3**). Regardless of *why* a participant chooses a determinate response when they didn't previously hold one, we cannot rule out by fiat the possibility that many participants would select determinate responses even when they were given the option to claim to have no view.

Existing versions of the disagreement paradigm almost exclusively[90] require participants to select either a realist or antirealist response.[91] For someone defending indeterminacy about folk metaethics, this is one of the most critical shortcomings. It should be obvious why: even if people had no determinate stance or commitment about a particular topic, if you required them to select from among a set of response options that reflect determinate positions, *no matter what pattern of responses you obtained, your data will always appear to constitute evidence of determinacy*. Even if we presented people with nonsensical questions that they could not plausibly have a stance or commitment about, if we forced them to choose from among some restricted set of options, it would at least superficially look like they held some determinate stance.

---

[90] The only exception I know of is Goodwin and Darley (2008). However, they only provide "other" as a response option. On the one hand, this makes their measures much less of a forced choice: if you really don't want to pick one of the main response options, you don't have to. On the other hand, "other" is so flimsy and unappealing an alternative that it would be unsurprising if almost nobody selected it, regardless of whether "other" would be the most apt account of their views.

[91] The closest is Sarkissian et al. (2011), which employs a 7-point Likert scale. However, since any response option is interpreted as evidence of *degree* of support for moral realism, this study is functionally incapable of detecting indeterminacy, since all response options are interpreted as support for or against folk moral realism. Even if participants who strongly disagreed with a realist statement did so because they had no particular stance at all, such responses would be indistinguishable from those who disagreed because they held a determinate antirealist stance. And, in any case, they *were* interpreted as antirealists (in particular, as relativists). This study thus not only presents a forced choice, but it presents a peculiarly narrow one characteristic of the weakest versions of the disagreement paradigm: participants can only express realism or appraiser relativism; no other response options are available.

Unfortunately, even if indeterminacy is correct, there is no easy way to modify the disagreement paradigm that would allow us to test for it, because there are plausible reasons why people would avoid response options that would indicate indeterminacy even if they held no determinate stance or commitment prior to participating in the study. For instance, even if studies provided responses explicitly indicative of indeterminacy (e.g. "I don't have a position about this issue") or at least consistent with it (e.g. "I don't know" or "I don't understand the question"), there are many reasons why participants may be disinclined to select these responses even if they held no determinate metaethical stance or commitment. People who participate in studies are often at least partially motivated by a desire to provide useful information or to otherwise aid researchers in their endeavors. It may seem unhelpful to select responses that don't fit researchers preselected categories. This could motivate participants who might otherwise wish to express ambivalence, ignorance, or indifference to select a more "decisive" response. Social desirability may also be a concern. People do not want to appear ignorant, indecisive, or obstinate. Participants may wish to avoid response options that signal these qualities. They may also be motivated to maintain a positive self-concept, which could be similarly threatened by an inclination to select response options that would lead them to see themselves as ignorant, indecisive, or obstinate. All of these factors may lead people to select response options that suggest a determinate stance even if they have none, which could lead to an overestimate of determinacy.

Even if we set all of these concerns aside, the inclusion of indeterminacy-consistent response options would still be inadequate. Most people with no determinate stance would ideally choose response options such as "I have no stance about this issue," or "I don't know." But would they? In other words, suppose we asked participants whether murder is morally wrong. They strongly agree that it is. Then we ask them to imagine a person strongly disagrees that murder is wrong. We then ask the participant whether:

*(a)  We are both correct*

*(b)  At least one of us is mistaken*

*(c)  I don't know / I have no perspective on this issue*

Would most people who have no determinate metaethical stances choose (c)? I suspect not. There are two primary reasons why. First, the validity of the disagreement paradigm requires participants to interpret the questions and response options in line with researcher intent. Yet there are many interpretations of what is being asked and what response options mean that could motivate participants to select seemingly-determinate responses without those responses indicating a genuinely determinate stance or commitment. People may not understand what researchers intend to ask for all the reasons discussed here; such unintended interpretations would necessarily entail a different understanding of what is being asked or what response options mean than what is intended by researchers. People may readily interpret the question in non-metaethical terms, or take (c) to have non-metaethical implications (regardless of whether they recognize its metaethical implications). For instance, they may interpret (c) to reflect that they don't have good reasons or justification for holding their first-order moral views (an epistemic stance), or that they don't believe murder is wrong (i.e., a *normative moral* stance), rather than an indication of their metaethical stance. Social desirability may also induce participants to be reluctant to choose (c), since it could signal lack of moral commitment or other negative traits. There may be *many* other, non-specific ways in which participants may interpret the question and the meaning of response options in unintended ways, or may favor determinate response options for reasons unrelated to these responses accurately reflecting their views. For instance, people may be naturally inclined to interpret difficult or ambiguous questions in ways that are comprehensible to them, even if their interpretation of what is asked does not match the interpretation intended by researchers. If so, their responses could not be appropriately interpreted as indicators of determinate metaethical stances or commitments, since these participants are effectively

responding to different, potentially unrelated questions. In short, people with no particular stance or commitment about an issue may nevertheless select a particular, determinate response for superficial reasons. In short, preexisting interpretative difficulties with the disagreement paradigm could not be remedied by including explicit response options for indeterminacy since there are many extraneous reasons to doubt people are interpreting the disagreement paradigm as intended in the first place. In the absence of substantial corrective measures to ensure participants interpret the disagreement paradigm as intended, the proportion of people who opt for or against seemingly determinate or indeterminate responses will not be diagnostic of the actual proportion of each.

These objections may seem to insulate indeterminacy against falsification. After all, if people *did* reliably select a response that supported folk indeterminacy, wouldn't this be evidence of indeterminacy? It would seem so. Yet I seem to be suggesting that if they did *not* choose such responses, this wouldn't be evidence *against* indeterminacy. This is definitely not what I am proposing. I am claiming that if people did not choose responses reflecting indeterminacy that this would not be good evidence against indeterminacy. But it would also not be good evidence for indeterminacy if they did. This is because participants who do hold a determinate stance may readily favor responses suggesting indeterminacy. For instance, they may select (c) because they find the question confusing, or aren't motivated to seriously engage with the question, want to hedge and remain non-committal, concerns about providing an "incorrect" answer, or worries about lack of anonymity and the potential repercussions that could come from providing a committed response (Denman et al., 2018; Zhu, 1996). When conducting semi-structured interviews designed to probe people's metaethical stances and commitments through more extensive dialog and interaction, David Moss reports that participants routinely vacillated or expressed uncertainty or hesitance (personal communication; see also Moss, 2017). Perhaps this is because they have no determinate stances or commitments. But it is also possible they do, but lack confidence, are motivated to give the interviewer the response they

think the interviewer wants but are not getting adequate confirmatory feedback, or are unfamiliar with articulating these views to others. After all, they are participants in our studies precisely *because* they lack formal education with the technical language used to describe these views. It is no surprise if people are not able to belt out clear metaethical stances and commitments even if they have them: they just don't have the language to do so!

Another reason why my concerns about the adequacy of the disagreement paradigm does not suggest that the indeterminacy thesis is unfalsifiable is that the disagreement paradigm is just one test; it is certainly possible that *some* tests are not capable of falsifying a hypothesis. Identifying inadequacies with one paradigm's suitability for falsifying a hypothesis does not entail that other paradigms couldn't do so. It is certainly possible other paradigms could cast serious doubt on the indeterminacy thesis.

A second problem is the possibility of *spontaneous theorizing* (see **Chapter 2**). Spontaneous theorizing is a serious problem not simply with the forced choice design of standard versions of the disagreement paradigm, but with *all* existing attempts to measure folk metaethical belief. This is not a problem for studies that are interested in people's intuitive inclinations when initially confronted with a novel philosophical consideration; however, most studies that assess folk metaethical belief attempt to describe the stances and commitments people already held before participating in the study. The possibility of spontaneous theorizing may be the most serious empirical challenge for the indeterminacy thesis. It is extremely difficult to demonstrate that people don't speak or think in certain ways prior to participating in social scientific research by *using the tools of social scientific research*. Indeterminacy faces what we could call the *pink elephant dilemma*. Although it is by no means a perfect metaphor, we are all familiar with the notion of someone saying, "Don't think of a pink elephant!" The whole point of this statement is that, of course, the very assertion itself tends to cause you to think of a pink elephant, even though you were of course almost certainly not thinking of one before that. In much the same way, it may be that people have no determinate metaethical stances or

commitments. Yet the very act of asking them about their metaethical views, and worse, providing them with pre-packaged options to choose from, can *cause* them to pick a view that they did not previously held. Indeed, it's unclear whether expressing agreement with a realist or antirealist stance is anything more than a fleeting judgment that doesn't stick with the participant after they finish the study. For all we know, our erstwhile folk realists and antirealists are like Boltzmann philosophers, who manifest a realist or antirealist stance during the course of the study, only to retreat back into the void of indeterminacy shortly thereafter.

In short, even with the inclusion of response options for indeterminacy, *any* pattern of responses would be weak evidence of the degree of folk determinacy and indeterminacy. If people consistently selected determinate responses, this would provide *some* evidence of determinacy, but it would hardly be decisive. Conversely, if people reliably favored responses that indicated indeterminacy, this would be at best only weak evidence of indeterminacy. Regardless of whether we include response options for indeterminacy, no pattern of responses to the disagreement paradigm would serve as strong evidence *for or against* indeterminacy. The disagreement paradigm is simply ill-equipped to address indeterminacy.

Worse still, the very nature of the disagreement paradigm may prompt participants to express views they did not previously hold due to the possibility of *spontaneous theorizing*. Of course, it is possible that participants hold implicit metaethical commitments or explicitly hold particular metaethical beliefs. Unfortunately, even when a variable of interest is mostly absent from a population, studies are designed in a way that will invariably yield the superficial appearance that determinate views are present.

**S2.14 Inaccurate, biased, or misleading stimuli[92]**

Another threat to external validity occurs when participants are given inaccurate, biased, or misleading stimuli, including instructions and response options (see Beebe, 2014; Pölzler; 2018a; 2018b; Pölzler & Wright, 2019). When this occurs, participants may fail to understand questions as intended, or respond in ways that do not reflect how they would respond in the absence of these biasing or misleading instructions. Goodwin and Darley (2008) provide one example of a way that instructions could mislead participants. Participants were given a list of moral statements then asked whether each statement is a:

> *(i) True statement*
> *(ii) False statement*
> *(iii) An opinion or attitude*

Yet Beebe (2015) provides several reasons why response option (iii) could mislead participants. First, *opinion* is ambiguous. It is sometimes used to refer to propositional beliefs. For instance, someone could have the opinion that "There was once life on Mars." Such statements are either true or false. Yet "opinion" could be used to refer to nonpropositional attitudes, such as a negative evaluative attitude towards particular foods or genres of music ("Country music? Bleh!"). As Beebe observes, G&D must intend for "opinion or attitude" to be interpreted in the latter way, since if they did not, it would not represent a genuine alternative to options (i) and (ii), but would in fact be consistent with both (p. 13). If participants interpret 'opinion' in a way consistent with expressing a propositional claim, then they would not be interpreting the question in the way Goodwin and Darley require for responses to be valid. Thus, the only way for option (iii) to provide a valid reflection of participants' noncognitivist moral stance is if they interpret an ambiguous term in a particular way.

---

[92] Similar phrasing was originally used by Pölzler and Wright (2020b).

As it is used in everyday language, the term 'opinion' is also unlikely to discretely and cleanly reflect either a propositional or nonpropositional claim; instead, its precise meaning may be vague or underspecified, highly context sensitive, and carry extraneous connotations incompatible with treating it as the simple semantic equivalent of a propositional or nonpropositional claim, as philosophers might use these terms. For instance, Beebe points out that as it is commonly used, "opinion" often carries epistemic connotations.:

> [T]here is a common, nonneutral use of 'opinion' that is generated when someone's point of view is said to be 'merely an opinion,' implying that the judgment in question is not based upon good reasons or evidence. And there is a colloquial sense of 'true' and 'false' (to which philosophers strongly object) that can serve as a foil to this sense of 'opinion'—viz., one that takes 'true' and 'false' to be equivalent to 'well-confirmed' or 'disconfirmed.' On this epistemic interpretation of 'true,' 'false,' and 'opinion,' the answer choices represented in (1.1) through (1.3) are asking participants to say something about the evidential merits of the ethical judgments in question (p. 14)

If participants understand the distinction between (i), (ii), and (iii) in line with this meaning of 'opinion,' this exacerbates the risk of interpreting the question as an *epistemic* question rather than (as intended) a *metaethical* question.

This problem is compounded by the instructions G&D initially gave participants.[93] Recall that standard versions of the disagreement paradigm first ask participants to rate their level of agreement with moral statements before proceeding to questions about metaethics. Before giving these statements to participants, they were instructed to "indicate your *opinion* about the status of each statement, whether it is true, false, or an *opinion*" (p. 1343, emphasis mine). Their first use of opinion in this sentence is obscure, but the most natural reading is that participants are instructed to express what they think is *true or false* about the moral issue in question; that is, whether it is true or false whether the moral claim in question is true, false, or an opinion. In other words, the very instructions

---

[93] Most of these objections were first raised by Beebe (2015, pp. 12-17).

themselves begin by using the term "opinion" to refer to propositional claims, right before they present participants with a response option that requires them to interpret "opinion" in exactly the opposite way.

This problem is not alleviated by the inclusion of "attitude." First, the inclusion of both in a single response option is a problem all by itself. Since the option "an opinion or attitude" includes two distinct possibilities, it involves the use of a double-barreled response option. There is a general consensus that researchers should avoid using response options that conjoin two or more distinct claims since doing so presents methodological problems (Menold, 2020). For instance, in G&D's study (a) we cannot distinguish participants who think the claim is an opinion but not an attitude from those who think it is an attitude but not an opinion and (b) people who agree that it is an opinion but not an attitude or vice versa may not want to select this response option because it could imply endorsement of the other arm of the disjunct.[94] More generally, such responses are simply hard to interpret. Do participants who choose this option think that the moral claim in question is an opinion, or an attitude, or both? Do G&D intend for 'opinion' and 'attitude' to be understood as synonymous? If so, why? If not, why collapse them into a single response option? Finally, the term "attitude" is in little better a position as "opinion" for clearly and unambiguously representing a noncognitive state distinct from statements that are true or false. Just like "opinion," "attitude" can refer to both cognitive and noncognitive states. Thus, *neither* disjunct is an appropriate response option for G&D's purposes.

---

[94] For a clear example of why people would wish to avoid agreeing with a disjunctive statement, imagine asking someone if they "enjoy eating pizza or human feces." Many people who enjoy eating pizza will say "no" for obvious reasons: such statements are ambiguous between inclusive and exclusive reasons, and their response could readily be interpreted to imply that they enjoy eating feces. As for people who enjoy eating feces, they would have entirely reasons to avoid affirming this fact.

**S2.15 Questionable *a priori* theorizing**

One concern with research on folk metaethics (and moral psychology in general) is the tendency for researchers to rely on top-down, *a priori* approaches when developing hypotheses and measures. As a result, researchers begin their empirical investigations with certain assumptions about what psychological phenomena they expect to find, and what measures would be appropriate for measuring those phenomena. While there is nothing necessarily inappropriate about this approach, researchers who rely on *a prioristic* assumptions run the risk of importing whatever biases, idiosyncrasies, and parochial preconceptions into the questions they ask and the methods they use to answer those questions.

Contemporary analytic metaethics is a recent, highly insular academic field that only matured in the past century, and is largely confined to obscure publications written and read almost exclusively in the Anglophone world by a demographically narrow group of elite scholars. Such researchers aren't simply WEIRD, they're *extra* WEIRD. In their description of philosophers engaged in contemporary analytic epistemology, Bishop and Trout point out that such work is:

> [...] written primarily for and by people who have received idiosyncratic educations and who have a highly specialized set of skills. This education significantly affects the concepts, categories, and inferential patterns one uses in thinking about the world [...] One needn't be a sociologist to recognize that philosophers as a group are a relatively small and idiosyncratic sample of folks. Philosophers' median education and intelligence are surely well above average. We speculate that philosophers' median scores on various MMPI scales (e.g., social alienation, hypersensitivity, and social introversion) might be above average as well. (pp. 703-704)

Their description of the methods used in standard analytic epistemology echo, in many ways, concerns I have raised with the approach researchers have taken to folk philosophy. As Bishop and Trout observe of epistemology, its "primary tools" furnish us with the reflective epistemic judgments of a group of idiosyncratic, non-representative people who have been trained to use highly specialized epistemic concepts and patterns of thought (p. 704). In a parenthetical, they note that by "highly

specialized" they mean that "people who have not received the relevant training would find at least some of those concepts and patterns of thought strange, foreign or unfamiliar" (p. 704).

This is precisely what I suspect is the case with respect to realism and antirealism. People use terms like "knowledge" without any familiarity with the vast philosophical literature on foundationalism, coherentism, reliabilism, and so on. Just the same, people use moral terminology without any familiarity with the concepts and distinctions discussed in metaethics. In both cases, philosophical theorizing takes ordinary terms and their allegedly ordinary content as starting points. This practice is distinct from the common practice among scientists to coin jargon that has no colloquial analog and isn't intended to reflect or describe ordinary thought or speech. There is no folk analog to *quantum harmonic oscillator* or *photothermal microspectroscopy*. As a result, the jargon and technical concepts and distinctions devised by philosophers are projected back onto colloquial terms like "knowledge," "reason," and "bad," with philosophers believing that they can see their favored theories are reflected in the way ordinary people speak and think. Unfortunately, in the absence of evidence, such beliefs may turn out to be little more than a philosophical mirage.

Folk metaethics is not the first or only example of research on moral psychology taking a top-down approach. Research on the moral/conventional distinction (MCD) likewise involved a presumption on the part of Turiel and colleagues that moral concerns were specifically associated with harm, justice, and rights (Turiel, 1983). Yet as Machery and Stich (2022) observe, "Other researchers, notably Richard Shweder and Jonathan Haidt, argued that Turiel's definition 'does not travel well', because people in non-Western cultures treat a much wider range of transgressions as moral" (e.g., Currey et al., 2021; Graham et al., 2011; 2013; 2016; Graham, Haidt, & Nosek, 2009; Haidt, 2012; Haidt, Koller, & Dias, 1993; Henrich et al., 2005; Shweder et al., 1997). Turiel and others presumed that moral concerns were confined to a narrow set of normative considerations that incidentally corresponded to precisely those normative concerns that are moralized in *their* culture. It seems

reasonable to presume that they presumed morality just was, by its very nature, concerned with the kinds of moral concerns most familiar to them. Although there is ongoing dispute about the extent of cross-cultural moral diversity (Gowans, 2021; Saer, 2019) Turiel and others would at best be vindicated by chance, and not because their presumptions were reasonable or the most methodologically appropriate way to study moral psychology.[95]

This top-down, *a prioristic* approach is a serious shortcoming in a great deal of research on moral psychology, and plausibly afflicts much of the research on folk metaethics as well. This *a prioristic* approach may be a symptom of a broader methodological blind spot in psychological research. In a critique still relevant today, Rozin (2001) argues that social psychology leapt prematurely into the methodological deep end, focusing on experimentation without first building a solid foundation in (among other things) solid descriptive research. This is no less true of folk philosophical research. A great deal of research on folk philosophy presumes that ordinary people think and speak in ways that conform to traditional philosophical categories. As a result, many studies simply operationalize philosophical concepts, then conduct research on the presumption that these concepts can serve as psychological constructs.

I believe this top-down approach to folk philosophy is a serious methodological error: it relies on the presumption that we can intuit, from the armchair, how ordinary people are disposed to think, rather than simply going out and engaging in the challenging, bottom-up process of finding out in a way that remains neutral and open to the possibility that folk philosophy doesn't conform to the concepts and distinctions that dominate academic philosophy. Research on folk philosophy should make no pretense that people *must* speak and think in accordance with the categories and concepts that interest philosophers. Many of the problems with folk philosophy, and folk metaethics in

---

[95] I suppose we cannot rule out the possibility that Turiel or proponents of the MCD could be extremely prescient and have very good intuitions about morality while I and those who agree with me lack these qualities.

particular, could have been avoided by a deeper engagement with how ordinary people think about the relevant topics without presumption or pretense. As Asch (1952/1987) observed:

> In their anxiety to be scientific, students of psychology have often imitated the latest forms of sciences with a long history, while ignoring the steps these sciences took when they were young. They have, for example, striven to emulate the quantitative exactness of natural sciences without asking whether their own subject matter is always ripe for such treatment, failing to realize that one does not advance time by moving the hands of the clock. Because physicists cannot speak with stars or electric currents, psychologists have often been hesitant to speak to their human participants. (pp. xiv-xv, as quoted in Rozin, 2001, p. 2)

Fortunately, some researchers have taken up this challenge of speaking to their parents, if not directly, then by proxy. Curry, Mullins, and Whitehouse (2019) drew on the Human Relations Area Files (eHRAF), which they describe as "an archive of thousands of original, full-text ethnographies from hundreds of societies of varying complexity, from simple hunter-gatherer bands to kingdoms and modern states" (p. 52). They settled on studying *sixty* societies drawn from around the world with an emphasis on selecting societies that are culturally independent of one another. Data on each society consisted of "at least 1,200 pages of reliable, well-rounded cultural data" gathered by ethnographers who'd lived in the communities they studied for at least a year with "working knowledge of the native language(s)" (p. 52). They extracted portions of the texts related to ethics then conducted a search for the terms consistent with the seven types of moral concerns proposed by their account. This isn't an ideal or completely bottom-up approach. There is still an emphasis on identifying terms and phrases consistent with a particular set of theoretical assumptions. Yet it is at the very least based on identifying patterns across culturally diverse populations by examining a rich body of ethnographic data based on studying the actual daily interactions of real human beings, not the parochial armchair assumptions of a culturally homogenous group of academics.

I am not convinced the approach Curry and colleagues take is adequate, either. This is not the place to provide an extended critique of their own methods, so I'll make two brief notes. First, I'm skeptical of attempts to identify "morality" in an external way that operationalizes morality without

consideration for whether the norms within a given society are thought of and spoken about as a unified normative domain. While many people in WEIRD populations may be disposed to see *moral* norms as, well, *moral*, it's not clear people in other societies do, or that subcultures rarely studied by researchers treat norms in a way that conforms with any particular, shared set of metanormative characteristics. There is some indication that people from distinct religious backgrounds (Levine et al., 2021), or people from non-WEIRD nationalities don't distinguish moral from nonmoral norms in the same way as people frequently do in general WEIRD populations (Berniūnas, 2020; Berniūnas et al., 2021; Berniūnas, Silius, & Dranseika, 2022; Machery, 2018).

Curry's account focuses on an external account of what "morality" predicted on a unified evolutionary account that binds the relevant norms together in accordance with a proposed shared functional role, i.e., to facilitate cooperation. Yet even if people did have an evolved predisposition to develop cooperative norms, it's unclear why we should regard these as *moral*. The very term *moral* may be a parochial, culturally idiosyncratic notion distinct to particular populations, as Machery (2018) and Stich (2018) suggest. If so, it would be misleading to describe a species-typical psychological trait as a capacity for distinctively *moral* cognition, rather than a culturally neutral form of normative cognition. Is *morality* not what we think it is? I, for one, don't think morality is reducible to a set of tools for promoting cooperation; indeed, very few of my moral concerns are distinctively about cooperation, and none are reducible to a concern for cooperation. Speaking for myself, I care more about increasing wellbeing and reducing suffering, and for me, *that's* what morality is about. For others, morality is about respecting rights, acting in accordance with God's will, complying with specific moral duties and obligations, cultivating virtues and acting virtuously, or some combination of these. Proponents of these views might even insist that this is what morality is about as an *analytic* matter that is known *a priori*; psychological theories about why morality evolved and what adaptive functions it allegedly serves may seem fundamentally misconceived: sure, maybe natural selection endowed us with a host

of psychological mechanisms that prompt us to develop norms and institutions that foster cooperation. But why is *that* morality?

There's a great deal more to be said about Curry's morality as cooperation (MAC) theory (Curry et al., 2019). However, I want to end by briefly noting a methodological concern that reinforces my concern about describing the theory as an account of *morality* as cooperation. Consider how coders were asked to code the data gathered in the eHRAF files:

> Please read through the following paragraphs. Your task is to decide, for each paragraph, whether it contains evidence that any of seven behaviors explained in table 1 is considered morally good or bad. 'Moral goodness' may be indicated by comments to the effect that the particular behavior is good, right, moral, ethical, or virtuous, or that it is an obligation, duty, or moral norm, and so on. It may also be indicated by morally-valenced words. For example, the mere mention of 'family loyalty,' or 'property rights' would suffice. Moral goodness can also be indicated by evidence that not performing the particular behavior is bad, wrong, immoral or unethical, etc. Similarly, moral badness maybe indicated by comments to the effect that the particular behavior is bad, wrong, immoral, unethical, or sinful, or that it is taboo, shameful, prohibited, and so on. (p. 53)

Suppose Stich (2018) is correct that there is no moral domain, and Machery (2018) is correct that morality is a historical invention. Suppose Sinnott-Armstrong and Wheatley (2012; 2014) are correct that there are no principled ways to distinguish moral from nonmoral norms. All of this would be obscured by this procedure: the coders would be *imposing* their own, culturally idiosyncratic notion of "morality" onto the texts they're coding, effectively smuggling in a top down, *a priori* notion of what "morality" is covertly, in the coding of the data, rather than having the data speak for itself. In other words, a great deal of psychological research does not generalize because most of the participants are from WEIRD populations. One remedy is to conduct research in culturally diverse populations. Yet if our methods of analyzing what people in these societies say and do explicitly, and by design, *forces* their words and actions through a categorization scheme that draws on potentially parochial WEIRD terminology and concepts, we'll simply be projecting the very biases and idiosyncrasies that made

WEIRD populations unrepresentative in the first place onto the data. This could result in the false impression that we're capturing universal patterns and regularities that may not be there.

Imagine if a group of time traveling knights wanted to create a theory of "chivalry." To do so, they transcribed interviews from people all over the world, then recruited a group of knights and nobles to code the transcripts for instances of people "discussing chivalry." Naturally, they see chivalry everywhere. Just the same, when people from a particular cultural and educational background that explicitly employ culturally distinctive terms and concepts when coding what others say, they will inevitably come away with the impression that other people speak and think in much the way they do. Avoiding WEIRD conclusions will require more than studying non-WEIRD populations. It will require researchers taking off their WEIRD-tinted glasses when coding and analyzing data.

# SUPPLEMENT TO CHAPTER 3

## S3.1 Advantages of metaethics scales

Scales have a number of advantages over the disagreement paradigm. Multiple items that each tap into the same construct can reduce noise and may be necessary to capture all facets of a multifaceted construct. Researchers may also use scale validation procedures to assess the degree to which different items appear to measure the same construct (Brown & Moore, 2012). A diverse array of distinct measures that appear to capture the same construct can provide mutually corroborating measures of the construct, which isn't possible for single-measure paradigms. However, there is at least some reason to doubt whether traditional scale validation procedures provide robust evidence of validity (Maul, 2017).

Subscales can also be used to represent distinct metaethical positions that participants are free to endorse or reject independently of their stance towards items that reflect other metaethical positions. This allows degree of belief in different forms of realism and antirealism to vary independently of one another, unlike the disagreement paradigm, which forces participants to choose only one metaethical stance towards each moral issue. This provides scales with the potential to capture metaethical pluralism or inconsistency towards the moral domain (and could, in principle, provide evidence for Loeb's incoherentism. This provides a decided advantage over the disagreement paradigm. The disagreement paradigm cannot do this, since participants must express a particular realist or antirealist stance towards each moral disagreement. As a result, participants with conflicting attitudes towards a particular moral issue are unable to express such mixed responses.

Scales are also more efficient than paradigms that take longer or are more cognitively demanding. More complicated or lengthy scales are more expensive and can take longer, which could limit their use. This isn't simply a practical matter, as participants may become fatigued, drop out, or

not understand instructions. More demanding tasks may also be unsuitable for some populations (e.g., children), and more prone to unintended interpretations, which could introduce methodological problems or limit generalizability by limiting use to only some populations. It may also be easier to assess the face validity of individual scale items than the face validity of more indirect methods. This is because the disagreement paradigm takes a more indirect approach that relies on assumptions about how participants interpret the source of the disagreement and the meaning of the response options, while scale items can more directly reflect a particular metaethical position.

Since researchers can use a pool of different items to represent a single metaethical position, scales could provide greater content validity than the disagreement paradigm (Allen, Iliescu, & Greiff, 2022). Using only a single measure may fail to capture all the relevant characteristics of a particular position, while a broader range of items that do not perfectly overlap with one another could pick up on facets of a construct (e.g., "belief in relativism") that the disagreement paradigm doesn't capture. Collier-Spruel et. al (2019) capitalized on these advantages by presenting a large initial pool of items to a panel of 11 experts in moral philosophy and psychology[96], asking them to rate how well an initial pool of 60 items represented relativism, then eliminating items that were not rated as highly representative.[97] While I doubt these methods (or the other methods they employ) are sufficient to

---

[96] Collier-Spruel et al. state that "Experts included professors, postdocs and graduate students of psychology and philosophy, all of whom were researchers of morality" (p. 4).

[97] This appears to be an explicit motivation for Collier-Spruel et al. (2019). They state that in constructing their initial pool of items, they removed "[...] items potentially lacking robust content validity" (p. 4). Their goal in recruiting experts to assess the validity of the items is likewise explicitly motivated by a concern for ensuring the content validity of the items.

However, I have some reservations about their procedure. Collier-Spruel et al. state that a panel of experts were asked to rate "each on a 7-point scale reflecting the accuracy with which it represented moral relativism" (p. 4). While this may be a fine way to assess the quality of each item as a generally adequate representation of relativism as a whole, proper assessment of content validity would require assessing *all* of the items holistically, to ensure that the pool of items taken together cover all relevant aspects of the construct in question. Does relativism have multiple facets? At the very least, cultural relativism and individual subjectivism were both collapsed into a single construct, so it would appear to have at least two facets. It is unclear to me whether it has any others. Yet whether or not a set of items broadly capture all facets of a construct is itself a matter that should be subjected to direct empirical evaluation. Item-by-item evaluation is simply incapable of providing the relevant evaluation.

establish validity, this approach at least has the advantage of starting from a large set of candidate items and whittling them down into what are perceived to be the best of the lot. Whatever their advantages, metaethics scales also suffer from a number of disadvantages as well.

## S3.2 Disadvantages of metaethics scales

Since scale items provide more direct measures of metaethical positions, they may be more vulnerable to demand characteristics (McCambridge, De Bruin, & Witton, 2012; Nichols & Maner, 2008; Orne, 1962). Metaethics scales in particular risk prompting participants to draw on prior knowledge of terms, concepts, and associations with particular metaethical stances that are inaccurate or prompt concerns about self-presentation. For instance, people may conflate relativism with tolerance, or believe that relativism implies or signals tolerance, open-mindedness, or other qualities people perceive to be desirable, which could encourage participants to inappropriately express agreement with these items. Others may associate relativism with undesirable qualities. Some participants may associate items that directly convey relativism or other forms of antirealism with nihilism, a lack of moral commitment, or debauched or libertine standards, and could prompt associations with progressive or politically left-leaning values that some participants may find undesirable. Conversely, participants may associate items representing realism with rigidity, close-mindedness, authoritarianism, conservatism, and a stodgy, dogmatic religious mindset.

---

For example, suppose I recruited a panel of experts and ask them to assess a set of items intended to measure one the Big Five personality trait *extraversion*. I provide them with five items that all accurately reflect extraversion, yet all five items specifically capture the *assertiveness* facet. Unfortunately, they might judge all five items to be excellent representations of extraversion, but it is not enough for each item, taken in isolation, to accurately reflect extraversion. The five items must, taken together, capture all the contours of extraversion in a way that provides an overall balanced measure of the construct. Five items exclusively emphasizing assertiveness would not be adequate, since this would exclude other facets of extraversion e.g., impulsivity and gregariousness. Any set of items that does not capture *all* facets of extraversion lacks content validity. Thus, the only way to properly assess the content validity of a scale is to consider all of its items *together*, not individually.

However, relativism may represent a much simpler construct that doesn't exhibit a variety of distinct subtraits. If so, this procedure may be unnecessary. If so, this concern is largely moot, and assessing the quality of individual items would be adequate. Even so, researchers only asked to rate individual items may not have been given the opportunity to express whether they considered the overall pool of items to exhaustively represent the domain.

Researchers who develop these items and only solicit feedback from experts may also be susceptible to the curse of knowledge (Camerer, Loewenstein, & Weber, 1989). When people possess knowledge of a particular topic, they may have difficulty recognizing and suppressing that knowledge when predicting other people's thoughts or actions. An expert at math, for instance, may have a harder time explaining mathematical concepts than someone with less expertise because the expert cannot fully extricate themselves from thinking in terms of their superior knowledge of a subject.[98] This same problem can influence the construction of scale items, especially for complicated, subtle, unfamiliar, technical concepts like metaethical positions (Bush & Moss, 2020). There is some risk that both the researchers who develop items, and the expert evaluators who judge how well those items reflect the relevant metaethical positions, suffer from a shared curse of knowledge: both researchers and expert evaluators know that the items are intended to reflect a particular metaethical position; they are not blind to the purpose of these items, unlike participants, who must interpret scale items without this context or background knowledge. Second, those of us who study metaethics are so familiar with the terms we use to convey relativism that we may be overconfident in thinking that non-experts will interpret these statements in the way we do. There may be ambiguity that we cannot see; paradoxically, our knowledge of metaethical accounts may blind us to ways people could interpret statements ostensibly intended to reflect these accounts that have nothing to do with the construct of interest.

To illustrate the problem, I'll draw on a topic that will hopefully prove unfamiliar to most readers—*Magic: The Gathering* (*MTG*). *MTG* is a collectible card game in which players take on the role of dueling wizards who summon monsters and cast spells.[99] If I designed a survey about *MTG*, I could easily fill it items like the following, and ask people to rate how much they agree or disagree with each:

---

[98] I believe Brian Tebbitt mentioned this example, and floated the possibility that non-experts may sometimes be better at educating novices in a particular topic for just this reason.

[99] For those of you who know the rules of the game, this example won't work very well. Hopefully you can substitute what I say here for the rules of some imaginary game with inscrutable rules, and this will suffice to make my point.

*(1) You may play a legendary permanent even if you control another permanent with the same name. However, you must then sacrifice one of them.*

*(2) Indestructible permanents may be exiled or sacrificed.*

*(3) Creatures with vigilance can block creatures with flying.*

Any player familiar with the rules of *MTG* would find these items intelligible.[100] Yet to anyone unfamiliar with the game, these items are complete gibberish. I suspect most *MTG* players would recognize that most people are unfamiliar with the rules and would have no chance of understanding these items. I am less confident that researchers studying metaethics are as aware of the degree to which their knowledge of the subject biases the construction of items.

There is no presumption that people who don't know the rules of *MTG* would be able to understand the rules of the game. Yet researchers *do* seem to presume that they share enough in common with ordinary people that participants will understand the items in their scales in precisely the way researchers intend, as though the items had a single unambiguous meaning. It's not obvious this assumption is warranted, and the structure of metaethics scales may be misleading. This is because items on metaethics scales don't use jargon or technical terms, but instead draw on a pool of everyday terms. This can create the misleading impression that these items represent the kinds of everyday, ordinary things someone might say or think, even if this isn't the case. As I will show, many items are ambiguous in ways that may not be obvious to people familiar with metaethical accounts and know that these items are intended to represent those accounts.

To illustrate the general problem, consider the statement, "The truth of moral facts does not depend on our preferences or desires." This sentence does not invoke any specialized jargon. Yet to anyone with training in metaethics, this sentence will often convey a specific notion: stance-

---

[100] The answers are as follows (1) Yes, you can play a legendary permanent even if it's a copy of one already on the battlefield. While you must choose to put one in the graveyard, this does not technically count as sacrificing it. (2) Yes, indestructible permanents can be exiled or sacrificed. (3) No, creatures with vigilance cannot block creatures with flying. Only creatures with reach or other creatures with flying can block creatures with flying.

independence about distinctively moral facts. We cannot simply presume that ordinary people will interpret it this way precisely *because* of the lack of jargon. It would be clear to anyone with the relevant experience what this sentence means, if they were attending a metaethics conference or read the sentence in a journal article on metaethics. But once we step outside a context that would serve to fix the meaning of otherwise ambiguous or unclear sentences, we've jettisoned all the contextual information that would flesh out the meaning of these sentences via pragmatics. Someone who studies metaethics may be inclined to interpret it as a statement about stance-independence even outside the contexts in which this sentence would ordinarily occur, but this could be due in part to their prior experience with sentences of this form reflecting this specific interpretation in the contexts the metaethicist is familiar with. It's unclear why someone without this training would necessarily interpret the remark in the same way, unless we have good reason to believe that there are few plausible alternative interpretations and that pragmatics play little role in how the statement is interpreted. While this may be a safe assumption for many conventional items used in social science, it's not obvious that such confidence is warranted for metaethical positions.

In short, testing face validity by having experts evaluate items is not sufficient to ensure that scale items are valid. The fact that experts judge a given set of items to be an accurate representation of a psychological construct is good evidence that *other experts* would interpret those items as intended. But it does not ensure that non-experts would. If non-experts consistently interpret items in unintended ways, then their responses will not reflect measures of the construct of interest, regardless of how well the items reflect that construct. When there are legitimate doubts about whether participants are interpreting questions as intended, the only way to ensure validity is to provide evidence that participants *themselves* interpret the items as intended. Indeed, it is even possible in principle that experts would interpret items in unintended ways but lay people wouldn't. Perhaps experts in a given field are so pedantic or so concerned about ambiguity that they consider items that

have a stable meaning among ordinary people to be highly ambiguous and impossible to interpret. If this occurred, it would even be possible for a set of items to be a valid measure of a given construct among a lay population even if most experts *denied* that the items in question were face valid. This might post serious epistemic challenges, but one could even test this by using comprehension checks or open-response questions and demonstrating (to the satisfaction of these same experts) that people are interpreting the questions as intended. Thus, while expert evaluation of a set of items may play an important auxiliary role in assessing their validity, it is far from decisive. Direct evidence of how the population you are sampling interprets items provides more direct evidence of validity.

In some cases, it may be that everyday language is not sufficient to adequately specify the construct of interest in a way that participants reliably interpret as intended. Could a simple scale assess beliefs about quantum mechanics or levels of selection in evolutionary biology in a way ordinary people would understand? I have my doubts. But even if it could be done, would it be appropriate to eschew actually assessing how people interpreted these questions? Conventional validation methods alone seem to fall short in circumstances where the items in question are difficult to interpret. Rather, direct evidence of adequate levels of intended interpretation among populations of interest seems like a critical step in establishing scale validity.

Finally, there is the question of whether there even is a psychological construct to be measured. Simply because one can devise a set of items, present them to participants, and obtain some pattern of responses, does not entail that these responses reflect any particular psychological construct, much less the one a scale is designed to measure. Conventional scale validation procedures may bolster confidence that when sets of carefully constructed items cluster together, that one has identified sufficiently coherent clusters of items to represent constructs of interest. Yet it is possible for such patterns to reflect superficial similarities between items, and for these superficial similarities to yield factor structures that appear to legitimize the putative constructs researchers are interested in, even

when further study would reveal that response patterns were not genuinely tracking the construct of interest. To illustrate this possibility, Maul (2017) constructed a series of progressively absurd studies that demonstrate how patterns can emerge from studies that aren't plausibly measuring any particular construct. Maul adapted a set of measures for capturing *growth mindset* (Dweck, 2006). Maul swapped out the key nouns in the items, e.g., "talent," or "intelligence," with nonsense terms, such as *gavagai* and *quintessence*, resulting in items such as "you have a certain amount of gavagai, and you can't really do much to change it" (Maul, 2017, p. 3). Yet a scale consisting of such items still exhibited the properties one might expect of a well-validated scale, such as a high Cronbach's alpha ($a = .91$), a two-factor solution (which seems to consist of positively and negatively-worded items), and significant (though weak) positive associations with agreeableness and openness. None of these results would be out of place for ostensibly "validated" measures. Yet this isn't too shocking, since much of the meaning of the items could still be gleaned from context. However, a similar pattern held even when Maul used completely nonsensical sentences consisting of nothing but lorem ipsum (a high Cronbach's alpha, reasonable loadings for a one-factor solution, and a significant negative association with agreeableness). In fact, this pattern persisted even when Maul used *completely blank items*. Participants were simply presented with:

> 1.
> *[1 = strongly disagree, 6 = strongly agree]*

As Maul observes, in all three cases "items were written in the complete absence of a theory concerning what they measured and how they worked" (p. 7). Nevertheless, Maul concludes

> Prima facie, it would seem difficult to take seriously the claim that any of these sets of items constituted a valid measure of a psychological attribute, and if such a claim were made, one might reasonably expect any quality-control procedure worthy of the name to provide an unequivocal rejection. To state this in Popperian language: If ever there were a time when a theory deserved to be falsified, this would appear to be it [...] Yet, this is not what occurred. In all three studies above, reliability estimates for the deliberately poorly-designed item blocks were quite high by nearly any standard found in the social sciences. (p. 7)

Maul's findings point to the possibility that traditional validation procedures may be insufficient, on their own, to conclude that a given scale captures the construct it purports to capture. This is too broad an issue to adequately address here. Yet I draw attention to it to illustrate that the mere fact that the items on a particular scale appear to cohere well with one another and to exhibit modest associations with other constructs is not sufficient to demonstrate that those scales are valid.

At least some metaethics scales perform better than this, however. Items on the MRS exhibit stronger associations with other constructs than the items used in Maul's scales, and more importantly exhibit *predicted* associations with other constructs (e.g., relativism scores are positively associated with tolerance scores). In addition, responses to the MRS are closely associated with other metaethics scales, such as Forsyth's EPQ.[101] These findings suggest that the MRS exhibits predictive validity and convergent validity, which Maul's findings did not show.

Unfortunately, these are also insufficient evidence of a scale's validity. If the items on the MRS, EPQ, and other metaethics scales are invalid for the reasons I propose, it is not only consistent with their invalidity but expected that items on different scales intended to measure metaethics to correlate with one another: they simply recapitulate similar mistakes prompt similar patterns of unintended interpretations! For comparison, it would be absurd to insist that a coin minted at a particular facility wasn't defective by showing another coin from the same facility that looked the same (or very similar), for the obvious reason that whatever caused the first coin to be defective could also cause the second coin to be defective. Just the same, since items on different metaethics scales employ many of the same terms and phrases, and consist of similar remarks, they may prompt similar conflations and thus result in similar patterns of unintended interpretations. While corroborating one measure against another is a sensible practice, it cannot provide evidence of validity in the absence of extraneous

---

[101] Two samples showed strong correlations between the MRS and EPQ-relativism at $r = 0.73$ and $r = 0.61$.

evidence of validity. Otherwise, the conclusion that both scales are invalid for similar reasons is equally consistent with the data as the conclusion that they're both valid.

**S3.3 Critiques of metaethics scales**

In the sections below, I present an extended critique of existing folk metaethics scales.

*S3.3.1 The Ethics Position Questionnaire (EPQ)*

The earliest attempt to directly assess folk metaethical belief may be the Ethics Position Questionnaire (Forsyth, 1980). While it is possible to find earlier references to metaethics in the work of Piaget (1932/1997), Kohlberg (Beebe & Sackris, 2016; Kohlberg & Kramer, 1969; Quintelier & Fessler, 2012), and Turiel (1974; 2012; Shweder, 1990) among others[102], the EPQ appears to offer the first explicit and direct effort at capturing a construct that, at least at first glance, seems to roughly correspond to *relativism* as it is understood in contemporary research on folk metaethics. The original EPQ included 20 items, but only ten items comprise the relativism subscale, so I will confine my assessment exclusively to these items. The 10-item relativism subscale of the EPQ may be seen in **Table S3.1** below.

However Forsyth and others conceive of the construct *relativism* measured by the EPQ, my only interest is in assessing its suitability as a measure of antirealism.[103] Yet the description Forsyth offers of *relativism* measured by the EPQ (hereafter EPQ-relativism) does not seem to match its more narrow technical meaning in contemporary analytic metaethics. Recall that relativism (as it is used in metaethics) is the view that moral claims can only be true or false relative to the standards of

---

[102] Piaget (1932/1997) used the term "moral realism" as "the tendency which the child has to regard duty and the value attaching to it as self-subsistent and independent of the mind, as imposing itself regardless of the circumstances in which the individual finds himself (p. 106, as quoted in Medinnus, p. 127). This is remarkably consistent with my use of the term *moral realism*. However, Piaget's notion of moral realism may have been more robust, including additional conceptual and psychological content that metaethicists would tend to exclude.

[103] I have corresponded with Forsyth about the degree to which the construct *relativism* used in the EPQ corresponds to the concept of relativism in contemporary metaethics, but he noted that he is not a moral philosopher and could not definitively confirm whether the two terms referred to the same concept (Forsyth, personal communication).

individuals or groups. This allows the truth status of moral claims such as "abortion is morally wrong" to because such claims have an indexical component that allows their truth to differ depending on which standard they are relativized to (Joyce, 2015). Consider the following claims:

Alex: *"Abortion is morally wrong."*

Sam: *"Abortion is not morally wrong."*

According to subjectivism, such claims indexically refer to the standards of the speaker, and are best interpreted as follows:

Alex: *"Abortion is morally wrong **according to my moral standards**."*

Sam: *"Abortion is not morally wrong **according to my moral standards**."*

Since each claim refers to a different set of moral standards, Alex and Sam don't contradict one another, just as they wouldn't if they said, "My name is Alex," and "My name is not Alex," respectively.

Despite common misconceptions that would suggest otherwise, relativism has no further normative implications, e.g., relativism does not guarantee that people actually have different moral standards, nor does it entail that we have any moral obligation to tolerate or respect people or cultures with other moral beliefs (Bush, 2016). Yet Forsyth and others appear to conceive of *relativism* as having a broader and more robust ideological stance towards morality, and to incorporate normative moral attitudes, rather than exclusively metaethical ones. I am not the first to observe that EPQ-relativism does not correspond to relativism. West (2016) notes that:

> Although Forsyth's typology does include a dimension for relativism, his descriptions of moral relativism refer to emphasising situational or contextual factors over moral principles and thus correspond more to moral particularism than to moral relativism. (p. 400)

West supports this characterization by pointing to Forsyth's own characterization of EPQ-relativism. According to Forsyth (2008), EPQ-relativism:

> [...] pertains to one's emphasis on moral principles as guides for determining what is right and wrong. Highly relativistic individuals' moral judgments are configural, for they base their appraisals on features of the particular situation and action they are evaluating (p. 815).

West points out that this more closely matches *particularism* (Dancy, 2004; Hooker & Little, 2004). Particularists are people who "believe that the moral status of an action is not determined by moral principles; instead it always relies on the particular configuration of its contextual features (Tsu, 2011, p. 388, as quoted in West, 2016, p. 200).[104] At least half of the items on the original EPQ are consistent with this claim, in that they explicitly describe variation in the moral status of an acting depending on the situation or context, or imply that insensitivity to context would be undesirable (emphasis mine):

1. What is ethical varies from one *situation* and society to another.

2. Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their own individual codes.

3. *Rigidly codifying an ethical position that prevents certain types of actions* could stand in the way of better human relations and adjustment.

4. *No rule concerning lying can be formulate*d; whether a lie is permissible or not totally depends on the *situation*.

5. Whether a lie is judged to be moral or immoral depends upon the *circumstances* surrounding the action.

Item (1) refers to the moral status of an action varying based on the "situation." Item (2) implies that the complexity of moral situations doesn't allow us to adhere to a shared set of moral principles, but instead calls for individuals to take initiative in deciding for themselves. (3) expresses rejection of the idea of a rigid approach to moral judgment that would prohibit sensitivity to situational factors, while items (4) and (5) describe how a particularist would regard lying, since particularists reject the notion that general moral rules can dictate whole categories of action. The same is not true of a relativist. A relativist might believe that lying is *always morally wrong*, but also recognize that, according to someone

---

[104] Strictly speaking, one need not be a particularist to be sensitive to situational factors when judging the rightness or wrongness of an action. Particularism could reflect the extreme end of a continuum between those who believe that there is at most one or a handful of very general moral principles that determine the rightness or wrongness of all moral action, to those who believe there are *no* general moral principles and that "moral thought does not consist in the application of moral principles to cases" (Dancy, 2017). Since EPQ-relativism is already treated as a continuous variable that admits degrees, this, if anything, makes particularism an even more appropriate analog in moral philosophy than relativism.

else's moral standards, lying might be morally wrong in some situations but not in others. In short, sensitivity to situational factors is completely orthogonal to relativism.

The mismatch between EPQ-relativism and relativism does not end with the former's association with particularism. Several of the items on the EPQ could be plausibly associated with other philosophical positions, and in some cases could reflect two or more positions. Unfortunately, many of these interpretations have little to do with relativism. As a result, the EPQ is simply not an appropriate measure of relativism, understood as a metaethical position regarding the indexicality of moral claims. However, in many cases it is unclear whether the EPQ is intended as a measure of relativism, but the items in question conflate relativism with other positions, or whether it isn't intended to measure relativism in the first place. Either way, none of the items in the EPQ could serve as a measure of relativism, regardless of whether they are intended to, without evidence that participants consistently interpret items in a way that reflects beliefs about relativism. Rather than address each in the main text, I present each item on the relativism subscale of the EPQ in the left column of **Table S3.1**, and a corresponding set of reasons why that item is not suitable as a measure of relativism. Notably, *none* of the items on the EPQ could serve as unambiguous measures of relativism.[105] Without exception, every item on these scales could either be plausibly interpreted in multiple ways, or is most appropriately interpreted as a measure of a belief or attitude orthogonal to relativism. In short, any attempt to use the EPQ as a measure of relativism would suffer from the simple fact that, at least for this purpose, none of the items are face valid. This will be a recurring theme as I assess the items included in other scales used to assess folk metaethics. Most scales simply fail to provide items that unambiguously represent metaethical statements.

---

[105] A more recent version of the EPQ has dropped some of the items discussed here, resulting in a reduced, 5-item relativism scale that maintains many of the properties that traditionally characterize robust measures (O'Boyle & Forsyth, 2021). This reduced scale has the added side effect of eliminating some items with troubling qualities, but the remaining items were not modified, and thus remain inappropriate as measures of relativism.

**Table S3.1**

*The relativism subscale of the Ethics Position Questionnaire (EPQ) (Forsyth, 1980)*

| Items | Reasons for invalidity |
|---|---|
| *1. There are no ethical principles that are so important that they should be part of any code of ethics.* | (1) Includes normative considerations ("important", "should")<br>(2) Could be interpreted as universalism<br>(3) Anti-generalism<br>(4) Could be interpreted as a practical question about what moral principles should be formalized (e.g. written down into a literal code of ethics) |
| *2. What is ethical varies from one situation and society to another.* | (1) Could be interpreted as descriptive<br>(2) Could be interpreted as particularism |
| *3. Moral standards should be seen as being individualistic; what one person considers to be moral may be judged to be immoral by another person.* | (1) Includes normative considerations ("should")<br>(2) Individual prerogative is not the same as relativism<br>(3) Could be interpreted as descriptive<br>(4) Double-barreled<br>(5) Vague (it's unclear what it means to say morality is "individualistic") |
| *4. Different types of moralities cannot be compared as to "rightness."* | (1) Could be interpreted as incommensurability<br>(2) Vague: it's not clear what "types of morality" are, what it means to compare them, or what is meant by "rightness."<br>(3) Unusual use of quotes around rightness |
| *5. Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.* | (1) Could be interpreted as universalism ("everyone")<br>(2) Could be interpreted as epistemic ("can never be resolved")<br>(3) Could be interpreted as descriptive/practical (whether people could resolve disagreements in practice is independent of whether moral facts are relative or nonrelative)<br>(4) Individual prerogative is not the same as relativism |
| *6. Moral standards are simply personal rules which indicate how a person should behave, and are not to be applied in making judgments of others.* | (1) Personal rules may seem like subjectivism, but it is not. Subjectivism is the view that the truth of moral claims depends on individual standards, but those moral claims are not confined to personal conduct and do not prohibit judging others<br>(2) The part about not making judgments of others could be interpreted as normative, in that it implies tolerating others by abstaining from moral judgment |

<div align="center">(3) Double-barreled</div>

| | |
|---|---|
| *7. Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their own individual codes.* | (1) Could be interpreted as epistemic ("complexity")<br>(2) This is a practical solution to a practical problem, and its implications are independent metaethical considerations<br>(3) Individual prerogative is not the same as relativism<br>(4) Includes significant descriptive content. You can be a subjectivist regardless of how complex you think ethical issues are |
| *8. Rigidly codifying an ethical position that prevents certain types of actions could stand in the way of better human relations and adjustment.* | (1) Biased: "Rigid" has negative connotations<br>(2) Could be interpreted as being about formalization moral rules by writing them down ("codifying")<br>(3) Seems to concern practical questions about how to improve welfare<br>(4) Could be interpreted as anti-generalism or in favor of particularism<br>(5) Vague (it's unclear what is meant by "better human relations" and especially "adjustment") |
| *9. No rule concerning lying can be formulated; whether a lie is permissible or not totally depends on the situation.* | (1) This is most naturally interpreted as an expression of anti-generalism or particularism<br>(2) Double-barreled |
| *10. Whether a lie is judged to be moral or immoral depends upon the circumstances surrounding the action.* | (1) Could be interpreted as descriptive<br>(2) Could be interpreted as anti-generalism or particularism |

*Note*. **Particularism**: A normative view which holds that there are at least some moral facts whose truth does not depend on an appeal to any general moral principles (Dancy, 2017).

**Descriptive**: Non-normative facts about what is or isn't the case.

**Anti-generalism**: The rejection of generalism, the view that there are general moral principles. Conceptually similar to particularism, though one could reject general principles while denying particularism as well.

**Universalism**: The view that there is one correct set of moral facts.

**Double-barreled**: An item that contains two or more distinct claims. Such items require one to agree with all or none of these distinct claims, which limits the ability of participants to express distinct attitudes towards each claim.

**Epistemic**: Related to how we acquire knowledge or justified belief.

**Practical**: Roughly, this refers to norms, actions, and policies that would promote one's goals.

**Formalization**: This refers to some (unspecified) process of formally enshrining some set of principles, e.g., listing them in numbered form, presenting them for public display, incorporating them into the law, etc.

Individual prerogative: This refers to the moral permissibility of making decisions in some domain that may differ from the decisions someone else might make without either person necessarily doing something morally wrong.

**Incommensurability**: The inability to compare two things in terms of a shared standard of evaluation (see Hsieh & Andersson, 2021).

Given how unsuitable these items are as measures of relativism, it is surprising that researchers have treated the EPQ as a measure of relativism. What's even more surprising is that these items are *clearly* not appropriate measures of relativism. Nevertheless, researchers have used or described the EPQ as a measure of relativism, including Colebrook (2018), Collier-Spruel et al. (2018), Goodwin and Darley (2010), Lam (2020), Quintelier & Fessler (2012), Rai and Holyoak (2013), Sarkissian and Phelan (2019), Uttich, Tsai, & Lombrozo (2014), and Yilmaz & Bahçekapili (2015a).[106] Collier Spruel and colleagues, along with Goodwin and Darley, raise numerous concerns about the suitability of the EPQ as a measure of relativism, noting similar concerns to some of those raised here. But aside from these exceptions, most researchers have uncritically treated the relativism subscale of the EPQ as a measure of relativism.

For instance, Rai and Holyoak (2013) refer to the EPQ as an "individual differences measure of relativist attitudes," and it is quite clear given the rest of their paper that they have metaethical relativism in mind (p. 996). In their first study, Yilmaz and Bahçekapili (2015a) use a version of the EPQ translated into Turkish as a measure of relativism.[107] Sarkissian & Pelan (2019) also use an adapted 5-item version of the EPQ as a measure of relativism in their first study. However, they express concern that the EPQ "did not capture metaethical views in a precise way" and adopt a new set of measures for subsequent studies (p. 4). Similar treatment of the EPQ appears throughout the references above.[108, 109] In short, there is an unmistakable tendency for researchers to refer to the EPQ

---

[106] Zijlstra (2019) provides a notable exception, explicitly recognizing the unsuitability of the EPQ as a measure of relativism. Zijlstra appears to have been convinced by concerns raised by Goodwin and Darley (2010). Collier-Spruel et al. (2018) go further, explicitly arguing that the EPQ conflates relativism with "tolerance" and "situationism" (p. 3). However, they still describe EPQ-relativism as though it were merely a flawed measure of relativism, rather than a measure that may not even be intended to capture "relativism" as it is understood in metaethics. Although they raise laudable concerns about the EPQ's suitability as a measure of relativism, their critique does not go far enough.

[107] They use the term "subjectivism." However, they use the term to describe contemporary research in folk metaethics and they explicitly contrast subjectivism with objectivism.

[108] Collier-Spruel et al. (2018) ran their relativism scale (the MRS) alongside the EPQ and appealed to a correlation between the two as evidence of the MRS's convergent validity. This only makes sense if the EPQ and MRS are presumed to measure similar constructs.

[109] Colebrook (2018) states that "Some of the items in Forsyth's questionnaire seem to be getting at subjectivism ("Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their individual

as a measure of relativism, or to even use it as a measure themselves, and this tendency has persisted until very recently.

Why would so many researchers describe the EPQ as a measure of relativism? One factor contributing to the inappropriate extension of the EPQ to research in folk metaethics is a simple instance of the *jingle fallacy* (Aikins, 1902). The jingle fallacy occurs whenever people mistakenly regard two or more psychological constructs as the same simply because they are referred to by the same name (Hawley, Stump, & Ratliff, 2010). I suspect that the jingle fallacy plays a substantial role in the EPQ misuse in research on folk metaethics. If so, this confusion would be somewhat understandable. First, Forsyth's own characterization is not framed by or for a philosophical audience, and much of the description presented may, as a result, give the misleading impression that there is significant overlap between Forsyth's conception of EPQ-relativism and relativism.

Even if there *were* overlap, it still appears that EPQ-relativism refers to a more robust and multifaceted construct than relativism. It isn't appropriate to use a set of items that are intended to measure a complex construct with multiple components as a measure of one specific facet buried within that construct, unless there were specific items specifically designed to exclusively capture that construct. Yet this is not the case with the EPQ. Numerous items run multiple concepts together, or are unrelated to relativism. This may not be a problem at all for the EPQ's intended purpose of capturing a particular cluster of psychological traits, but the mere fact that one of those traits might be an endorsement of relativism is not sufficient to warrant regarding the relativism subscale as a whole as an appropriate measure of relativism as it is understood in metaethics. After all, we wouldn't

---

codes"), whereas others are better seen as measuring relativism ("What is ethical varies from one situation and society to another"' adding that "Relativism as it is used by Forsyth and others appears to be a blended measure, capturing anti-realist attitudes generally" (pp. 46-47). Neither of these items is an especially good measure of relativism. And while I agree that the EPQ appears to be a blended measure, I do not agree that it captures anti-realist attitudes generally. There is little in the way of any item that would clearly capture an antirealist stance, including those items that stand the best shot at capturing relativism. For instance, no items are even close to serving as measures of error theory or noncognitivism, despite these serving as unambiguous and paradigmatic forms of antirealism.

use a measure of attitudes about conservatism as a political ideology in general as a measure of someone's attitude towards religion.

This apparent instance of the jingle fallacy is understandable. The term "relativism" has been and continues to be thrown around with little regard for consistency. Different people mean different things when they use the term, and over time the term has come to have many semi-overlapping uses and connotations. Even philosophers mischaracterize relativism or conflate it with other distinctions. This varied and inconsistent use has rendered the term especially amenable to the kinds of confusions that would result in the jingle fallacy. This risk was further aided by Forsyth's characterization of the EPQ invoking terms and descriptions that do appear, at times, to verge on capturing aspects of relativism as a metaethical position, and perhaps Forsyth *is* gesturing towards this to some extent. The problem is that genuine relativism would, at best, represent only one facet of EPQ-relativism among an array of other facets that are distinct and to a great degree don't even concern metaethics at all.

Some researchers may also simply be unfamiliar with relativism as a metaethical concept. This is unfortunate, and could be rectified by moral philosophers and psychologists working more closely and reviewing one another's work. In other cases, researchers may be genuinely interested in a notion of relativism distinct from the meaning I am appealing to here. In those cases, there may be no jingle fallacy occurring, but if so, such efforts can probably be dismissed if the intended construct isn't in some way related to the metaethical distinctions of interest here. Of course, some researchers may simply disagree with me about the suitability of the EPQ as a measure of relativism. They might, for instance, argue that my objections are pedantic and may be practically irrelevant if it turns out people do reliably interpret relativism items on the EPQ as expressions of relativism as a metaethical position. To support this claim they could appeal to the reasonably high correlation between the relativism subscale of the EPQ and Collier-Spruel et al.'s (2019) moral relativism scale ($r = 0.73$ and $r = 0.61$ in their two samples that ran both on the same participants). However, you may only establish the

convergent validity of two or more scales if there is independent reason to believe that at least one of those scales is a valid measure of the construct of interest. Since there are also independent reasons to doubt the validity of the MRS, the correlation between the two scales is not by itself compelling evidence of the validity of the EPQ as a measure of relativism. One would first have to provide compelling evidence of the validity of the MRS. Furthermore,

To illustrate why, suppose I have an instrument that I claim can measure the temperature outside. *If* I provide compelling evidence that my instrument can do so, and *if* you produce another instrument that consistently yields similar measurements of the temperature under the same conditions, then you would have evidence that the instrument you produced is also a reliable instrument for measuring the temperature. Yet without that crucial step of first providing independent evidence that establishes my instrument can tell what the temperature is, there would be no reason to believe that your instrument could also measure the weather. Suppose my instrument could not measure the weather at all, and you made an exact physical copy of it. Now, whatever is causing our two instruments to provide the same readings may be the same, but if it doesn't correspond to the temperature, then even perfect correlation between our instruments would be irrelevant. Likewise, if it turns out that the MRS is not a valid measure of relativism, then its correlation with the EPQ is not by itself strong evidence for the validity of the EPQ as a measure of relativism.

It is also possible some researchers simply disagree with me, and regard the EPQ as a valid measure of relativism. If so, I am open to that possibility, but given the concerns I have raised, I am skeptical they would be able to provide convincing evidence for its validity.

One reason to suspect that at least some researchers have an incomplete or inadequate conception of relativism as a metaethical position is the tendency to regard relativism as one end of a continuum anchored by relativism on one end and "objectivism" (understood roughly as moral realism, i.e., the view that there are stance-independent moral facts) on the other, as though each were

the negation of the other. In fact, although the EPQ is often cited as a measure of relativism, some researchers have referred to it as a measure of "metaethical views" (Sarkissian & Phelan, 2019, p. 4) "anti-realist attitudes generally" (Colebrook, 2018, p. 47) and of "objectivity" (Uttich, Tsai, & Lombrozo, p. 189), which the latter define as the belief that "some moral claims are true in a way that does *not* depend on people's decisions, feelings, beliefs, or practices" (p. 189, emphasis original). In other words, they imply the EPQ is a measure of *realism/antirealism*.[110] Yet the EPQ is even less suitable as a measure of realism than as a measure of relativism.

In addition to potential for relativism and stance-dependence to come apart conceptually (e.g., for some forms of divine command theory or other relation-designating accounts), relativism is simply not the only form of antirealism. There is error theory and noncognitivism as well. And someone could deny that there are stance-independent moral facts without committing themselves to any of these views. After all, the belief that there are no stance-independent moral facts doesn't *require* that we take a positive stance which antirealist account is correct, and is even compatible with the view that all of the conventional positions are incorrect. This is, after all, *my* position, and positions like mine represent alternatives that cannot be reduced to the traditional antirealist categories that one might include in a scale were one directly testing for them. Thus, one cannot simply take the rejection of relativism as an endorsement of realism, since it is possible for someone to reject both relativism and realism, or to reject relativism but not endorse realism. In other words, it is a mistake to treat relativism and realism as mutually exclusive *and* mutually exhaustive.

It is also possible to endorse a hybrid position that permits some forms of relativism within the context of a broader realist framework, e.g., one might believe that a range of possible moral systems can be correct relative to their respective standard, but that some other moral systems could

---

[110] Oddly, they also refer to these efforts as "recent" despite Forsyth (1980) being nearly 35 years old at the time of publication.

not be correct relative to their moral systems (Wong, 1995; 2006). Conventional relativist scales could not identify people who held such views. People may also be metaethical pluralists who adopt relativist stances towards some moral issues but not others. The EPQ relies primarily on abstract questions about the moral domain as a whole, which would be incapable of detecting such pluralism, and while it does include two items with concrete normative content, those items are only about lying. It is not plausible that we could generalize from attitudes about lying to one's attitudes about morality as a whole, since this would run up against a standard and neglected problem: generalizing about members of a category by appealing to the characteristics of nonrandom members of that category. Judd, Westfall, and Kenny (2012) describe this as a "pervasive but largely ignored problem" in social psychology: researchers will treat specific stimuli as random factors that are implicitly presumed to be representative of the category they represent *even though these stimuli were not randomly selected and are at risk of not reflecting members of the category from which they are drawn*. For instance, researchers may use images of particular individuals to represent members of a particular group without taking into account the possibility that judgments about these images may be distinctive to those particular images, and not reflective of general attitudes about the group as a whole. This can result in mistaken inferences, e.g., if participants exhibit a particular reaction to a particular image, this may be mistakenly taken to reflect a general attitude about members of that individual's group, even if this is not the case. Likewise, one cannot make general inferences about people's attitudes towards issues within the moral domain by evaluating their attitudes towards specific moral issues (such as lying) *unless* we have solid grounds for believing that attitudes about lying are representative of attitudes about morality in general (which we don't have).

I can also see little principled rationale for including eight abstract items and two concrete ones. Other studies suggest that there may be some systematic differences in how participants respond to abstract versus concrete questions in metaethics (Pölzler & Wright, 2020a; 2020b); it would stand

to reason that if one wanted to construct tools for measuring metaethical belief that one would not throw a pair of concrete measures in with a host of abstract ones without some reason for doing so.

Furthermore, relativism itself doesn't even comprise a single position. Items that capture beliefs about cultural relativism may not be appropriate as measures of subjectivism, and vice versa. Thus, at the very least, a *unidimensional* scale of relativism may not make much sense, unless items were devised to be neutral between different forms of relativism. The items on the original 10-item EPQ do not meet this condition, since some stress group differences and others emphasize individual differences. An individual/group neutral scale would also lack the resolution to distinguish between the two positions, so it would at best offer a low resolution or incomplete picture of folk relativism, as well.

In light of these many deficiencies, there is little reason to believe the EPQ could serve as an appropriate measure of relativism or of realism or antirealism more generally. Whatever its merits as a predictive tool for assessing individual differences in a business context (the purpose for which it was designed), it is not a good tool for assessing metaethical stances or commitments as they are understood in the psychology of folk metaethics. This is not surprising. The EPQ was not designed by (or in conjunction with) philosophers for the specific goal of assessing folk metaethics. However, subsequent scales *were* designed or adapted for this purpose. While these scales suffer fewer problems, the best that I can say about them is that they are marginally less unsuitable. Many of the reasons that render the EPQ inadequate apply to these scales as well. As such, my criticism of the scales that follow will be comparatively brief.

### S3.3.2 The Objectivism-Subjectivism scale (TOS)

The objectivism-subjectivism scale (TOS) emerged shortly after the EPQ (Trainer, 1983). The scale was devised by Trainer for the explicit purpose of assessing whether ordinary people are moral "objectivists" or "subjectivists," terms that roughly correspond to realism and antirealism,

respectively. Trainer telegraphs this goal in the very title of the paper: "Ethical objectivism-subjectivism: A neglected dimension in the study of moral thought." Unfortunately, and with a touch of irony, the article itself has been neglected. The EPQ has seen continuous use since its inception, while to my knowledge nobody has ever used Trainer's scale in subsequent research.[111] This is unfortunate. Emerging in the early 1980s, the EPQ and TOS predate contemporary research on the psychology of metaethics by several decades. Had anyone used either as a starting place or inspiration for studying folk metaethics, we might be a few decades ahead of where we are now. Yet neither managed to spawn a literature specifically dedicated to the study of folk metaethics in a way that is well-integrated with the philosophical literature in the way contemporary research on the psychology of metaethics has accomplished.

Like the fall of the Roman empire, any official start date to the present era of folk metaethics research will be largely a matter of convention, but if I had to choose a date, it would be with the publication of Goodwin and Darley (2008). While several articles preceded it, including the aforementioned Forsyth (1980) and Trainer (1983), as well as Nichols and Folds-Bennett (2003), Nichols (2004), and Wainryb et al. (2004), none of these attracted the attention that Goodwin and Darley have. Their 2008 article seems to have drawn enough attention to reach critical mass, and establish psychology of folk metaethics as a sustained area of research. Sadly, this literature does not include Trainer's TOS. Trainer's work has never been cited in the more recent literature, and I suspect nobody is even aware of his article.

The TOS is an ambitious effort to assess folk metaethical belief. Trainer drew on an impressive number of participants for the time, with samples just breaching 400 participants in some cases and with a total of 2,300 participants. Trainer even began by conducting exploratory research that included 140 interviews that included children, college students, and adults. This is remarkable. Few researchers

---

[111] Forsyth (1980) has been cited 2,389 times as of March 25, 2023. Trainer (1983) has been cited *five* times.

would have the time to immerse themselves in discussions with a wide range of participants to better

calibrate the measures they devised these days, given the omnipresent demands to publish or work as

a barista.[112]

One of the most laudable elements of the TOS is Trainer's efforts to clarify what, exactly, the

TOS is intended to measure. Trainer begins with a characterization of objectivism and subjectivism:

> The most fundamental division in ethics is between those who see the realm of morality as involving nothing more than things like the preferences or desires of individuals, the experienced consequences of action and the way groups choose to organize and regulate social behaviour towards desired ends, and, on the other hand, those who see morality as involving moral 'facts' which exist in addition to or irrespective of human opinion, preference and desire. Ethical objectivists assume the existence of a Moral Law of nature whereby some actions or judgements are in fact Morally right and some are in fact Morally wrong regardless of what humans think or prefer. The ethical subjectivist believes that all moral issues are completely reducible to considerations of desire, consequence and man-made rules. He sees moral codes and principles as no more than human creations developed to regulate behaviour and therefore as being continually open to revision. He regards moral argument as possible but it can only take the form of attempting to show someone what pursuit of his ultimate values will entail. (p. 192)

This distinction roughly corresponds to realism and antirealism, with room to quibble about the

precise characterization of the opposing perspectives Trainer has in mind. Trainer's characterization

of objectivism seems a fair approximation of my use of realism. Subjectivism, as I understand the

term, is more narrowly construed as the view that moral claims are true or false relative to the moral

standards of individuals. Trainer seems to have a more inclusive notion in mind that seems to capture

a range of antirealist positions. The notion that morality involves "nothing more than things like the

preferences or desire of individuals" could be construed in subjectivist terms, but also seems like it

could encompass noncognitivism. Trainer also adds seemingly contractualist (Scanlon, 1998) or

constructivist (Bagnoli, 2021) notions of morality, and perhaps noncognitivism, though it is not clear.

---

[112] Typically, it's publish *or perish* but to my knowledge philosophers and psychologists typically survive failure to procure an academic position.

Regardless, the items themselves provide ample indication that Trainer has a roughly relativist or antirealist conception of morality in mind.

Another unusual feature of the scale is that it does not use Likert scales. Instead, participants are presented with pairs of statements and asked to choose the one they "favour" or to select "Don't know — Not sure." There are 17 such pairs, resulting in 34 items in total. All items are featured in **Appendix B**. Like the EPQ, the items on the TOS exhibit a host of features that raise doubts about their value as measures of realism and antirealism. Although there are a handful of issues specific to individual items, there are five central, recurring problems with items on the TOS. First, several items conflate epistemic and metaethical considerations. Consider the first pair of items:

| EPQ 1A *realism* | You can say without any doubt in some situations that something is right or wrong, and you say that people who don't agree with you are wrong. |

| EPQ 1B *antirealism* | It isn't possible for anyone to be really sure what is right or wrong. You can only say that others with different opinions are wrong. |

Including phrases such as "without a doubt" and "to be really sure" invoke notions of what we are in a position to *know*, which inappropriately entangles questions about the nature of moral facts with questions about moral epistemology. These items are the clearest instance of unambiguously *epistemic* statements:

| EPQ 10A *realism* | It is possible to know that your basic moral principles are the right criteria for evaluating things. |

| EPQ 10B *antirealism* | Your basic moral principles can only be your best guess at the criteria for evaluating things; you can never know whether yours are the right criteria. |

While some moral realists may regard access to moral facts to be a requisite feature of moral realism, this is at best a controversial position that many would regard as unnecessary. Note that because both items include substantive epistemic elements, they are not face valid measures of realism or antirealism.

A second recurring problem for items on the TOS is the inclusion of substantive normative content or considerations. Consider these items:

| EPQ 14A *realism* | We can say much more than that we do not like this. We could say cruelty to animals is in fact morally bad and should not be done whether or not anyone likes to do it. |
|---|---|

| EPQ 15A *realism* | Honesty is in fact morally better than cheating. There is more involved here than my liking for one and my dislike of another. |
|---|---|

Both items prompt participants to consider their normative stance towards the specific issues presented in the items, which could prompt interpretations unrelated to metaethics, or lead people to favor one or another of a paired set of items for inappropriate reasons, e.g., social desirability or demand characteristics.

Another problem with many items on the scale is that they are double-barreled, include multiple, distinct components, or are drawn out and complicated in ways that may be cognitively demanding, confusing, or otherwise difficult to process.

| EPQ 2A *realism* | Some things are wrong no matter what anyone thinks and people should be told this if they don't know. |
|---|---|

This item is double-barreled. In addition, one could believe that some things are wrong "no matter what" even if they don't necessarily think people should be told this if they don't know. Finally, the

second conjunct (that we should tell people who don't know) is normative claim, and thus does not reflect a metaethical position.

| EPQ 16B *antirealism* | I cannot condemn even murder as being a morally bad action, because it breaks no moral law. There are no moral laws of nature, there are only laws men make up. All I can say are things like, I don't approve of murder and most people don't so they make laws against it. |
|---|---|

This item is an even more egregious example, exhibiting all three of these concerns. It expresses multiple, independent claims. People who agree with one or more, but not the rest, have no way to express this. But it is also confusing and complicated. What does it mean to break a moral law? What are "moral laws"? Even if participants interpreted this as intended, i.e., some kind of stance-independent moral fact, why should this mean that one couldn't condemn murder? One might be a realist for reasons other than the belief that there are moral laws, e.g., one might be a realist and a particularist, and believe moral transgressions that meet these other criteria may be condemned. Or one might be an antirealist but still believe we may condemn moral actions. The item goes on to include a host of distinct claims:

(1) You can't condemn serious moral transgressions if there are no stance-independent moral facts
(2) There are no stance-independent moral facts
(3) There are only laws created by people (moral laws, or "laws" in the conventional sense?)
(4) We can only say that we don't like an action and that other people also don't like it, so they (why not we?) make laws against it

There are too many components to this item. I have no idea what we could infer if someone were to agree with this more than the alternative it was paired with.

A related issue, and one that also applies to the previous item, is that many items are simply *unclear*. That is, they lack face validity for the simple reason that they are vague, ambiguous, or worded in ways that make them hard to interpret. Consider this item:

| **EPQ 8B** **realism** | It is possible to claim that your moral principles are better than those of some other people, such as a thief or a sadist. |
| --- | --- |

Strictly speaking, of course this is *possible*. Interpreted literally, it is trivially true. Presumably, this is intended to be interpreted as the claim that one can *correctly* claim that their moral principles are better than other people's. Yet it is not clear whether agreeing with this would entail a realist perspective towards morality. First, what is meant by "better"? It could be interpreted in a practical sense, e.g., that one's moral standards promote societal welfare more than other people's. It could even be interpreted as a claim about what kind of attitude we're entitled to have towards our moral standards. That is, participants may judge that each of us is entitled to consider our own moral standards to be superior to others.

Another problem with this item is the potential entanglement of normative considerations. I'm an antirealist, and I certainly think my moral principles are better than those of a thief or sadist; I just don't think they're better because they're correct and the thief or sadist's moral standards are incorrect. I think they're better because I favor my normative standards over other people's normative standards. I'm sufficiently self-aware to recognize the stance I take towards issues like this, but this may be a struggle for ordinary people. They may simply interpret this question to be asking them whether opposition to stealing and sadism is good, to which their answer will presumably be "yes." Yet this normative stance is simply not diagnostic of any particular metaethical position. Other items are similarly difficult to interpret. Consider this item:

| **EPQ 11A** *realism* | My moral principles are not just things I prefer. They are things that ought to be preferred because they are morally important. |
|---|---|

This item isn't terrible. It's getting at an important aspect of how many realists conceive of realism. Namely, the implication here isn't simply that there are descriptive facts about what people's moral preferences are. Rather, there are facts about what moral standards they *ought* to have, i.e., there are facts about what our moral standards ought to be that don't depend on and aren't determined by our preferences (or goals, standards, attitudes, etc.). However, what does it mean for something to be "morally important"? This phrasing has no obvious or standardized meaning. It could be interpreted, like many other items on the scale, as a practical question. Even an antirealist could agree that it's *important* to have rules against murder and theft. And since these are moral rules, it's important to live by some set of moral rules, not because they are *true* but because they are *useful*. Without knowing how participants interpret "morally important," it's unclear what agreement with this item would mean.

Aside from issues of clarity and complexity, many items are also too *specific*. Imagine I wanted to know if you were a moral realist, and the way I asked was, "Do you believe that there is a single objective standard of moral truth created by Yahweh on October 22$^{nd}$, 4004 BCE?" If someone says "no," this does not mean they are a moral antirealist. It simply means they reject that specific realist position. No item on the TOS exhibits this degree of specificity, but the example is a deliberate caricature to illustrate that rejection of specific claims does not entail rejection of a more general claim. When your goal is to measure something general, such as moral realism, your items must be presented at the appropriate level of generality. Unfortunately, the TOS frequently fails to do this. For example, the following represents an antirealist position:

| **EPQ 7B** *antirealism* | Moral laws are made up by humans as ways they choose for regulating behaviour. |
|---|---|

I'm a moral antirealist, and I don't agree with this. Many antirealists would reject this. This item describes a position on the status of moral antirealism. Namely, it's not just that there are no stance-independent moral facts, but the moral rules we do have were designed by people for a specific purpose: to regulate our behavior. While I am confident that *one* reason *some* moral rules were created was as a means to regulate behavior, this is at best one among many causal factors contributing to the construction of moral standards. It is also specifically focused on some shared set of moral standards that presumably dictate behavior within communities, which ignores individual moral judgment.

More importantly, this item could plausibly be seen as a descriptive claim about the origins of moral standards within communities. Such an account is compatible with being a moral realist. One might think that there are stance-independent moral facts *and* that people created moral laws to regulate behavior. One might even think that moral facts *just are* facts about what promotes cooperation, and that the construction of moral laws that regulate behavior involves mechanisms that lead to knowledge of the moral facts (Sterelny & Fraser, 2016). In short, moral realists could readily agree with this item, while antirealists could just as readily reject it. Another item exhibits a similar problem:

| **EPQ 9B** *antirealism* | Rights are entirely created by man. |
|---|---|

Once again, antirealists do not have to believe that rights were entirely created by people to endorse moral antirealism, nor do moral realists have to deny that rights were entirely created by people to endorse moral realism. A moral realist might, for instance, recognize that there are both stance-

independent moral facts *and* that there are institutionalized rules and principles reflected in our laws, and regard the latter as "rights." The construction of and respect for rights can exist as a matter of descriptive fact within a realist framework. Like the EPQ, some items also focus on concrete moral issues:

| EPQ 15A *realism* | Honesty is in fact morally better than cheating. There is more involved here than my liking for one and my dislike of another. |

In addition to inappropriately entangling the participant in normative considerations, and being double-barreled (which only exacerbates the problem of having a normative reading: the first sentence is *exclusively* normative and has nothing to do with metaethics), this item also faces the same problem the two items about lying on the EPQ face. Namely, the presumption that a person's attitudes about this specific moral issue would generalize to their attitudes about morality as a whole. Many researchers presume, without justification, that participants would regard all moral norms as having the same metanormative characteristics. Yet it remains a possibility that people could be metaethical pluralists. As such, it is not obvious that attitudes about honesty would generalize towards the moral domain as a whole.

In addition to problems with individual items, there is also a problem with some of the item pairings. For each item, participants are forced to choose between one realist and one antirealist item. Ideally, these pairs would be mutually exclusive and mutually exhaustive. Yet they are not. In some cases, you could reasonably believe *both* statements. Forcing participants to choose one or the other doesn't allow us to capture ambivalence, lower confidence, pluralism, or weaker attitudes than would be possible were participants permitted to express level of agreement with each item in the pair separately, which lowers the resolution of detail the scale is able to provide. Since many of the items

are double-barreled or contain multiple parts, it is also possible for participants to agree with parts of each statement, but they have no way to express this. Another problem is that participants may reject *both* positions. Consider the first pair:

| EPQ 1A *realism* | You can say without any doubt in some situations that something is right or wrong, and you say that people who don't agree with you are wrong. |
|---|---|

| EPQ 1B *antirealism* | It isn't possible for anyone to be really sure what is right or wrong. You can only say that others with different opinions are wrong. |
|---|---|

Participants must either express that they have *no doubt* whether something is right or wrong, or that it *isn't possible* to be really sure what is right or wrong, but that one can *only* say others have different opinions. Participants must choose either certainty or skepticism, with nothing in between. Yet it is possible, even likely, that many (or even most) people would hold a modest, intermediate position between these positions. Such an option is not available, so participants must choose between one of two extreme positions they may not agree with, and opt for the lesser evil. In such cases, it would be incorrect to infer the participant actually held the belief they selected, and we have little way of knowing how frequently this occurs.

One final problem with the scale is that the response options don't neatly distinguish different realist and antirealist positions. We cannot tell, for instance, whether a participant who selects one of the antirealist positions endorses noncognitivism, error theory, or some form of relativism. This by itself does not *invalidate* the scale, but it does mean that it could at best offer only limited information about folk metaethical positions. It is possible Trainer only intended, by "subjectivism" to capture some form of relativism. It is not clear from Trainer's description of the term, which appears to capture a cluster of related but distinct views, some of which are not even clearly metaethical. Whatever its purpose, the items included in the scale do not seem capable of neatly distinguishing

those who endorse different antirealist positions, e.g., we cannot distinguish participants who think moral claims are truth-apt but false or stance-dependent from those who think moral claims are not truth-apt at all.

In spite of these shortcomings, the TOS does include a surprising number of items that come far closer to the mark than any scale before or since. For a first attempt, this is impressive. And, given the difficulties with constructing valid items, and that to my knowledge Trainer was not trained in philosophy, this is a remarkable achievement. Consider the following items:

| **EPQ 4A** *realism* | Some values or actions are objectively right, they are right in fact, whether or not individuals think so. |
|---|---|

| **EPQ 4B** *antirealism* | All judgements about right and wrong state nothing more objective than the ideas or attitudes of individuals. |
|---|---|

| **EPQ 1A** *realism* | Human beings can only discover moral laws; we can't make them. Just as we can't make up true laws of science to suit ourselves neither can we make up true moral laws. |
|---|---|

None of these items are perfect. My findings show that participants struggle to understand "objective" in the way intended by researchers. However, in this context, the rest of the sentence in the first item clarifies that it is meant in a stance-independent respect. The second item likewise attempts to convey stance-dependence. While the third item includes unfortunate epistemic terminology ("discover") and continues with the strange phrasing ("moral laws"), it uses an analogy to natural laws in an attempt to illustrate what moral realism is *roughly* like. This item could confuse some participants into thinking that moral facts must also be *natural* facts, in the way the laws of science are, and that might dissuade anyone who thinks otherwise. And it is also fairly complicated. Nevertheless, each of these items properly attempts to frame realism in terms of stance-independence, rather than in terms of

universality, or indisputability, or some other characteristic that may correlate with, but not directly entail realism.

These items *may* prompt greater rates of intended interpretation. But there are far too many problems with the rest of this scale for it to serve as an appropriate measure of realism and antirealism. Modified versions of a handful of these items could make their way onto a scale that attempts to circumvent the challenges I've presented against these items, but without additional evidence about participant interpretation, there would be little reason to be confident even very well-designed items will serve as valid measures.

Another suspicious feature of the items is the high degree of variation in participant response across items. For instance, among a sample of 76 high school students, 8% chose the objectivist response in set 1, but 92% chose the objectivist response for set 15. This is unusual if these items were intended to reflect measures of the same stance. There was considerable variation in the overall proportion of realist responses participants favored, as well. However, Trainer claims that "Almost all people show some objectivist tendencies" (p. 200). This seems true enough, but then again, most show subjectivist tendencies as well. Coupled with the inconsistency *within* participant responses, Trainer could just as readily cite these findings as evidence of both interpersonal and intrapersonal variation, and presented these findings as evidence of metaethical pluralism, though I'd have been far more concerned that what we're seeing is a noisy response pattern due in part to unintended interpretations and interpretive variation.

In a prescient turn, Trainer also observed that participants who were comfortable offering an antirealist response for abstract items about morality were far less willing to do so when they were presented with especially evocative concrete moral items. As Trainer observes:

> It is one thing to agree that 'There are no moral laws of nature; they are all man-made', but to be asked whether a specific and disturbing case of infanticide or cruelty is solely a matter of

preference which breaks no moral laws of nature is to face a much more searching test of whether or not one is a thorough-going subjectivist. (p. 200)

Trainer devised a new set of twelve "unpleasant concrete cases." Unfortunately, I was not able to obtain these items. Notably, participants shifted far more towards a more consistent pattern of realist responses for the concrete items. This is unsurprising. However, Trainer concludes that the higher rate of realist responses for concrete items "give the more realistic indication of the extent to which people tend to the objectivist position in their moral thinking" (p. 200). Trainer reasons that "On this evidence very few people endorse the subjectivist view so clearly or confidently that their allegiance holds up when tested by confrontation with specific and extreme problem cases" (p. 200).

I don't think the TOS is a valid measure of metaethical stances or commitments to begin with, and I am skeptical about the validity of these concrete items as well. Unfortunately, they were not presented in the article, and I was unable to obtain them. However, we may still consider whether we ought to conclude that when people lean more towards realism when presented with concrete moral issues than abstract ones that this is a better indication of their genuine metaethical stance. This *may* be the case. It could be that when the implications of a position are salient that we reflect more carefully and offer a response that is more probative of what we really think. On the other hand, it is also possible that the greater rate of realist responses for concrete items is *less* probative of genuine metaethical stances and commitments.

An experience I routinely encounter as a moral antirealist illustrates why such questions may not be diagnostic of their true metaethical position (assuming they have one). One of the most predictable moves I encounter when discussing metaethics with moral realists is the following: The realist will describe some horrific action, such as genocide or sexual assault, e.g., then ask "Are you saying *genocide* isn't objectively morally wrong?!" This is a rhetorical trap, and it is especially effective

when there is an audience of nonphilosophers observing the conversation. Think of what it would sound like if the following exchange occurred:

>**Realist**: *Is genocide objectively morally wrong?*

>**Antirealist**: *No. I regard genocide as extremely immoral and repugnant and am wholeheartedly against it. However, while I regard it as morally wrong, I don't believe that there are stance-independent moral facts, so I don't believe it is a stance-independent fact that genocide is wrong.*

The antirealist needn't bother with anything after "No," because they already lost the audience. The trap has already been sprung, and the antirealist has fallen right into its gaping maw. As soon as the antirealist says "no," many people will interpret them not merely to be expressing a metaethical stance, but to have also implied that they don't take genocide very seriously, or even that genocide is permissible, or to more generally convey a permissive, indifferent, or even supportive attitude towards genocide. This is because normative and metaethical considerations are entangled in such a way that people will mistakenly interpret "no" to express a first-order (normative) stance towards genocide, rather than or in addition to a second-order (metaethical) stance towards the moral status of genocide.

The realist appears to have superficially asked a simple question that invites a "yes" or "no" response, when any response to this question will in practice readily imply a response to multiple, distinct questions, several of which have nothing to do with the antirealist's metaethical stance. It does so by exploiting (even if unwittingly) pragmatic implicature to create the impression that the antirealist is conveying one or more non-metaethical stances or attitudes in addition to or in lieu of their metaethical stance. This is achieved by playing off of two readings of the question: a hyper-literal, formal meaning exclusively reducible to the semantics of the question, and a more inclusive meaning that incorporates all of the connotations and associates pragmatically implied by a response to the question.

The formal component is a question about the truth of the conjunction of two distinct questions:

*(1) Is there a stance-independent moral fact about the moral status of genocide?*

*(2) If so, is this stance-independent fact that genocide is morally wrong?*

However, responses to this pair of questions conversationally implies one or more non-metaethical beliefs or attitudes, as well. First, (2) is readily conflated with or at least accompanied by its non-metaethical cousin:

*(3) is genocide morally wrong?*

Note that this is an exclusively *normative* or *first-order* question that has nothing to do with metaethics. An antirealist can (and most probably do) have a normative moral stance towards genocide, and regard it as morally wrong. Yet a crucial, further question is implied:

*(4) Do you personally disapprove of, condemn, or otherwise object in the strongest possible terms to genocide?*

There is no easy way for the antirealist to pull these components apart with a simple "yes" or "no." Elaboration is required to disentangle the various explicit and implied questions, and address each piecemeal. Unfortunately, that is precisely why the question is so effective. If the antirealist attempts to tackle the question head on, they are forced to acknowledge that, technically, their answer is "no," after which any efforts to elaborate and clarify that this does not mean that they don't think genocide is bad sounds like backpedaling, or at least no longer sounds very compelling. This question has a lot of rhetorical advantages. It isn't a loaded question, and by relying on pragmatics to bury the antirealist, the realist can insulate themselves against objections that there is anything formally inappropriate about the question. The realist can always lean on the hyper-literal semantic content of the question to insist that they're asking something completely appropriate. This semantic/pragmatic divide provides a ready-made escape hatch that allows the realist plausible deniability that there's anything suspect about the question.

Yet it is suspect. The realist's ploy attempts to borrow the rhetorical force that naturally accompanies our reaction to egregious moral transgressions by piggybacking questions about

metaethics on top of normative questions, and hope that the audience will conflate a response to the former as a response to the latter. And since the latter fly under the radar because they are conveyed pragmatically, rather than by the literal semantic content of the question, the realist can pretend that they're asking an innocent philosophical question. This is the most natural explanation for why the realist's ploy almost always invokes extremely serious moral transgressions, such as genocide, rape, or torture. Imagine if the realist asked the same question, but for a mundane moral violation:

**Realist**: *Is failing to report additional income on your taxes objectively morally wrong?*

**Antirealist**: *No. I regard failing to report additional income on your taxes as immoral and I object to it. We should all be honest when filing our taxes. However, while I regard it as morally wrong, I don't believe that there are stance-independent moral facts, so I don't believe it is a stance-independent fact that failing to report additional income on your taxes is wrong.*

This exchange seems rather pedestrian, and even a bit silly. This is because the antirealist now seems to be saying too much. There is no reason to bend over backwards to elaborate on why the answer is "no." The realist's question simply doesn't have the rhetorical force to make the antirealist look bad, once we swap out a serious transgression for a less serious one.

All this may be summed up simply: when asking metaethical questions about concrete questions, people have a very strong incentive to favor a seemingly realist response because doing so is the only effective way to convey that the act in question violates their own first-order normative standards, and is something that they personally condemn and find deeply objectionable. This *normative entanglement* provides a powerful incentive predicated on self-presentational concerns, e.g. a desire to signal that one isn't a horrible person.[113]

---

[113] *Normative entanglement* is a hypothesized phenomenon that occurs whenever non-normative content (such as statements, questions, or imperatives) is implicitly accompanied by normative content in such a way that they are difficult to disentangle. This can result in a stronger and a weaker reaction. The stronger reaction occurs when the normative content influences one's conception of the non-normative content. In its weaker form, a person's non-normative conception may remain unchanged, but normative entanglement nevertheless influences their incentive to react in a way that misleadingly implies a particular stance towards the non-normative content that they do not actually have.

I propose that this *normative entanglement* can explain why participants presented with metaethical questions about concrete moral issues are more inclined to favor realist responses. Even if they are not explicitly aware of it, participants may be wary of selecting options that would reflect poorly on them, and this may drive them to favor responses that would not pragmatically imply tacit approval (or at least a lack of disapproval) for egregious moral violations.

There is some precedent for this proposal in the related hypothesis of *affective biases* prompting performance errors. Experimental philosophers have spent the past two decades disputing whether ordinary people tend to regard determinism as compatible or incompatible with moral responsibility (Becklloyd, 2021; Feltz, Cokely, & Nadelhoffer, 2009; Lim & Chen, 2018; Murray & Nahmias, 2014; Nadelhoffer & Monroe, 2022; Nahmias, 2006; 2011; Nahmias et al., 2005; 2006; Nichols & Knobe, 2007). The most infamous finding in this literature is the disparity in responses to moral violations that are described in abstract or emotionally unengaging terms, versus those that involve vivid, concrete descriptions of egregious moral violation (Nichols & Knobe, 2007). For instance, participants in the more abstract and low affect cases may be asked whether "it is possible for a person to be fully morally responsible for their actions" or whether it is possible for someone to be morally responsible for cheating on their taxes. Conversely, the more evocative cases describe a man stalking and raping someone, or a husband carrying out a plot to murder his wife and children to pursue a relationship with his secretary. Participants tend to judge that people in the less evocative cases cannot be morally responsible for their actions, while they tend to conclude that people can be morally responsible in the high affect cases.

There are many explanations for these results, with little in the way of a consensus about which (if any) are correct. Yet one hypothesis that has persisted from the outset of this research is the *performance error model*. According to this model:

[...] people ordinarily make responsibility judgments by relying on a tacit theory, but when they are faced with a truly egregious violation of moral norms (as in our concrete cases), they experience a strong affective reaction which makes them unable to apply the theory correctly. In short, this hypothesis posits an affective performance error. That is, it draws a distinction between people's underlying representations of the criteria for moral responsibility and the performance systems that enable them to apply those criteria to particular cases. It then suggests that people's affective reactions are interfering with the normal operation of the performance systems.

Whether or not this explanation can account for results in research on free will, it could account for Trainer's findings. That is, it could be that participants are motivated to avoid antirealist responses to questions about concrete moral issues because doing so could pragmatically imply a tolerant or permissive attitude towards egregious moral violations.

### S3.3.3 The New Meta-Ethics Questionnaire (NMQ)

The New Meta-Ethics Questionnaire (NMQ) is an 8-item scale adapted from the EPQ. The NMQ was developed by Yilmaz and Bahçekapili (2015a) for research among Turkish participants and was conducted in Turkish. The version that appears in **Appendix C** is its English translation. The explicit purpose of the scale is to measure both "objectivism" and "subjectivism," with the former seemingly corresponding to realism and the latter representing an antirealist conception of relativism.

The NMQ suffers many of the same problems as items on the EPQ. As Moss and I noted previously, all of the items on the NMQ conflate metaethical and non-metaethical considerations, including normative, descriptive, and practical claims (Bush & Moss, 2020). Since I have already published an extensive critique of this scale, and the criticisms are virtually identical to those of the EPQ, I will simply provide the table presented in previous work summarizing the conflations between metaethical interpretations and non-metaethical interpretations (see **Table S3.2**). As that paper concludes, "Given that no item on this scale is face valid, it is not an appropriate tool for measuring metaethical belief" (p. 15).

**Table S3.2**

*Critique of the items appearing on the New Meta-Ethics Questionnaire (EPQ) (Table adapted from Bush & Moss, 2020)*

| Items | Conflations/Ambiguities with items |
|---|---|
| *1. What is moral varies on the basis of context and society.* | Conflates descriptive relativism with metaethical relativism. Conflates relativism with contextualism. |
| *2. Moral standards are personal, therefore something morally acceptable to one person might be immoral for another person.* | Conflates descriptive relativism with metaethical relativism. |
| *3. Since moral rules are not absolute, no definite judgments about them are possible.* | Conflates exceptionless rules, insensitivity to context, and/or universality with objectivism with the use of "absolute". Conflates epistemic and metaphysical interpretations with use of "definite." |
| *4. Different cultures adopt different values and no moral law is right or wrong in an absolute sense.* | Conflates descriptive relativism with metaethical relativism. Forces participant to agree/disagree with a compound statement. Conflates exceptionless rules, insensitivity to context, and/or universality with objectivism with use of "absolute". |
| *5. We can agree on 'what is moral for everyone' because what is moral and immoral is self-evident.* | Conflates epistemic and metaphysical interpretations with use of "self-evident." Conflates universality with objectivism. |
| *6. If morality were to differ from person to person, it would be impossible for people to live together.* | This is a question about the consequences of descriptive relativism. It is not related to metaethical objectivism/relativism. |
| *7. Since the moral laws I believe in are universally true, they can be applied to everyone in the world regardless of culture, race or religion.* | Conflates universalism with objectivism. Implies imposition of one's values on other people/cultures, which entangles normative considerations with metaethical ones. |
| *8. If a moral law is right and good for others, it is also right and good for us.* | Conflates normative and metaethical questions.<br><br>Conflates universalism and objectivism. Implies imposition of one's values on other people/cultures, which entangles normative considerations with metaethical ones. |

*S3.3.4 The three-item moral objectivism scale (3MO)*

Although I discuss these items in the main texts, there are other difficulties with the 3MO as well. Many issues revolve around the potential pitfalls of expecting ordinary people to associate terms that have multiple, nontechnical usages in everyday discourse with a specific, narrow intended interpretation. For instance, references to a "single" moral standard may not merely imply that there is one nonrelative standard (however broad, flexible, and context-sensitive it may be) as opposed to the multiple moral standards a relativist would endorse. Instead, a "single" moral standard could also (or instead) imply a *rigid* moral standard or a narrow set of moral principles that are general and apply to all people without being open to there being different ways to comply with the same general moral principle across cultures.

Yet it could also result from culturally contingent, learned associations between various terms, phrases, sentence structures, and concepts that people associate with particular ideological stances or identities. As people go about their lives, they acquire an extensive knowledge of the associations between particular terms, phrases, and ideas and particular cultures, ideologies, and kinds of people or communities associated with or make use of those terms, phrases, and ideas. For instance, people in the United States may associate people who listen to country music, enjoy hunting, or have a southern accent with political conservatism. In many cases, such associations can result in negative evaluations of such people based on these assumptions (e.g., Amira et al., 2018; Ash et al., 2020).

The possibility of associating between particular terms and phrases and with particular political or ideological perspectives may be a ubiquitous obstacle to valid measurement in research on metaethics. Conservative or religious people may have a desire to signal their opposition to or at least lack of association with secularism or the political left. When they are confronted with items describing "relativism" this may trigger associations with the political left, academia, postmodernism, and other institutions and ideological perspectives that conservative or religious people may wish to distance

themselves from. Indeed, in a recent collaboration with my colleagues David Moss, Andres Montealegre, and David Pizarro, we found that people who endorse moral antirealism are perceived as less religious and less politically conservative than those who endorse realism (Moss et al., n.d.).[114]

The association between is far from surprising to anyone familiar with the way antirealism is framed among Christian apologists and conservatives. The term "moral relativism" is often included in the medley of terms and phrases marshaled to galvanize opposition to rival ideologies, just one more in a long line of terminological boogeymen like *postmodernism*, *communism*, *secularism*, and *nihilism*. Consider this remark from former Speaker of the House, Paul Ryan, who apparently believes moral relativism is the *primary cause of poverty in the United States*:

> Moral relativism has done so much damage to the bottom end of this country, the bottom fifth has been damaged by the culture of moral relativism more than by anything else, I would argue. If you ask me what the biggest problem in America is, I'm not going to tell you debt, deficits, statistics, economics — I'll tell you it's moral relativism. Now is it my job to fix that as a congressman? No, but I can do damage to it. But it's the job of parents to raise their kids … But let's not ignore it. These things go beyond statistics, they go into the culture. (Pethokoukis, 2011)

Communist critic Solzhenitsyn (2009) linked communism with moral relativism, warning that:

> Communism has never concealed the fact that it rejects all absolute concepts of morality. It scoffs at any consideration of "good" and "evil" as indisputable categories. Communism considers morality to be relative, to be a class matter. Depending on circumstances and the political situation, any act, including murder, even the killing of hundreds of thousands, could be good or could be bad. It all depends on class ideology. And who defines this ideology? The whole class cannot get together to pass judgment. A handful of people determine what is good and what is bad. But I must say that in this respect Communism has been most successful. It has infected the whole world with the belief in the relativity of good and evil. Today, many people apart from the Communists are carried away by this idea [...] But if we are to be deprived of the concepts of good and evil, what will be left? Nothing but the manipulation of one another. We will decline to the status of animals." (p. 60)

---

[114] We use the terms "objectivism" and "relativism" rather than realism and antirealism, respectively. These differences are not important for the points made here and are more or less interchangeable.

Prominent figures such as Pope Francis and Pope Benedict XVI (Countering Moral Relativism, 2015) have warned of the dire consequences of moral relativism, yet these sentiments are echoed among church leaders, apologists, academics, and laypeople alike. Christian apologist and talk show host Frank Turek (2020) tweeted that "If there is no objective morality then love is no better than murder," and one can readily find articles and comments decrying relativism randomly strewn across the internet. One article warns that "Moral relativism leads to moral paralysis and indifference" (Dominici 2020) while another claims that that moral relativism has led to removing God from our political discourse and opening the door to dictatorship (Tenny, 2015) and a third insists that relativism requires us to accept rape, slavery, and domestic violence (Walters, 2020). Some are even hard to distinguish from parody, such as an article helpfully titled "Why moral relativism is dangerous," that features an image of a burning building an ominous quote from Goebbels, infamous devotee of Hitler and head of propaganda for the Nazis: "Today, there seems to be only one absolute thing: relativism" (Reudell, 2021). Far from appearing only in old and obscure publications, many of these references are recent, and continue to emerge primarily from religious or conservative outlets.

These hyperbolic screeds may overstate the harms caused by relativism, but they have accurately linked relativism with their ideological rivals. Collier-Spruel et al. (2019) found that higher relativism scores on their moral relativism scale (the MRS) were moderately correlated with endorsement of liberal values and rejection of conservative values, and were negatively associated with authoritarianism and religiosity, prompting them to bluntly conclude that "[m]oral relativism reflects anti-authoritarian, anti-conservative, and non-religious perspectives" (p. 12).[115] Yilmaz and Bahçekapili (2015a) likewise found that measuring relativism using the EPQ was associated with lower religiosity and belief in God using multiple measures, and found that priming participants with belief in God

---

[115] All of these correlations tended to hover between 0.2 and 0.5. Opposition to authoritarianism was the most closely associated with relativism ($r = -0.45$, study 5; $r = -0.42$, studies 6-9), followed by conservative political ideology ($r = -0.33$, study 5; $r = -0.34$, studies 6-9) and finally religiosity ($r = -0.24$, study 5; $r = -0.22$, studies 6-9).

decreased relativism scores on the NMQ, while priming people with arguments for relativism reduced confidence that God exists. Note that, while I have argued extensively against the validity of all of these measures, these findings may nevertheless reflect genuine associations between conservatism and religiosity with the specific phrasing used in items intended to reflect relativism, so the measures used in these studies don't need to be valid to serve as evidence of the same general hypothesis that there is a conceptual link between ideology and the phrasing that appears in items intended to represent relativism, *whether or not those items are valid measures of realism versus antirealism.*

### S3.3.5 The Moral Relativism Scale (MRS)

Collier-Spruel et al. (2019, hereafter CHJFF) have developed one of the most ambitious and rigorous scales for assessing folk metaethics, the Moral Relativism Scale (MRS). As the name suggests, the purpose of the MRS is to assess the degree to which people endorse metaethical relativism. There are many commendable features of this scale and the way it was developed.

CHJFF employed a panel of experts to assess how well an initial pool of 60 items represented relativism, and only retained items that passed a sufficient threshold. To demonstrate that the panel didn't simply greenlight any items for measuring relativism, they also had the panel assess the items used on the EPQ and found that items on the MRS scored significantly higher in terms of how well they reflected relativism.

After devising their initial pool of items, the scale was refined and validated using a pool of over 3,200 participants over the course of nine studies. CHJFF also distinguish metaethical relativism from normative relativism by devising a separate scale to assess moral tolerance, providing additional evidence for the discriminant validity of the MRS. In fact, they ran the MRS alongside *40* other constructs to assess its relation to each, providing a rich body of data for assessing the MRS in relation to other measures.

Their initial pool of items was winnowed down to just 10 items that loaded reasonably well onto a single common factor, reported high internal consistency for the items ($\alpha = 0.89$), good test-retest reliability ($r = 0.70\text{-}0.75$ after one week) and sought to ensure they included only items that at reading level below 10th grade (using the Flesh-Kincaid measure available in Microsoft Word, which showed a reading level of 8.3). Finally, they offer evidence of the predictive validity of the MRS by demonstrating that scores on the MRS are associated with variables one would expect it to be associated with, e.g., higher scores on the MRS were correlated with progressive values and tolerance for disagreement and negatively correlated with authoritarianism, conservative values, and religiosity. All of these qualities suggest that the MRS is a promising measure of relativism. I am happy to concede that it is *better* than many alternatives. But "better" doesn't entail "good enough."

There are a few straightforward limitations to the MRS. The first, and most obvious, is that the scale only attempts to assess the degree to which people endorse relativism. This is a significant limitation, since denying relativism does not necessarily entail that one endorses moral realism. Level of agreement with some items on the scale may reflect support for realism, while level of agreement with others may only reflect one's stance towards relativism in particular. For instance, denying that "People can disagree on what is morally right without anyone being wrong" could reflect realism, and agreeing that "There are moral rules that apply to everyone regardless of personal beliefs" could be interpreted as endorsement of realism (though it is best construed as universalism, *not* realism). However, level of agreement with other items cannot tell us whether the participants endorse realism, or merely reject relativism.

For instance, if someone disagrees that "The viewpoint of one's culture determines whether their actions are morally right," this could be because they endorse some form of antirealism other than cultural relativism, and in this case they could even be an individual subjectivist (a form of relativism), yet still disagree with this item. Since rejection of relativism and endorsement of realism

are not conceptually identical, nor could they reasonably be treated as a single psychological construct, the scale does not provide a conceptually clear continuum between realism and antirealism, but instead at best only allows us to judge how much people endorse relativism. While this is a viable way to assess folk metaethics, it still provides at best only limited insight into how best to interpret lower scores on the scale, since aggregating responses to items that more plausibly reflect realism with items that don't isn't appropriate.

For comparison, suppose you wanted to measure degree of belief in Christianity. To do so, you used both items that reflected belief in Christianity, e.g., "I believe Jesus was the son of God and died for our sins," and those that reflected denial of Christianity, e.g., "God does not exist." Agreeing with the second statement indicates that the participant is an atheist, while disagreeing indicates that the participant believes in God. However, disagreeing with the first statement *does not* indicate that the participant is an atheist; it only indicates that the participant isn't a Christian. It would be inappropriate to use both items on a single scale as a continuum anchored by atheism and Christianity. Rather, it would only allow us to distinguish Christians from non-Christians. While we could treat the individual items that reflect atheism as measures of atheism, we could not treat aggregate scores as a measure of atheism. In the same way, the MRS could not serve as an appropriate measure of realism without extracting inappropriate items and assessing its validity as a tool for measuring realism in their absence. Simply put: the denial of relativism does not entail realism. As such, the only way to construe the two anchors of the MRS are as agreement and disagreement with relativism. We cannot infer that low scores entail endorsement of realism, because low scores are consistent with endorsing some form of antirealism other than relativism.

A second, related concern is that several of the items represents cultural relativism (e.g., "The viewpoint of one's culture determines whether their actions are morally right") while other items represent individual subjectivism ("Each person is the final authority on whether his or her actions

really are morally correct"). One problem with this is that, since most of the items on the scale reflect individual subjectivism rather than cultural relativism, a low score on the scale is even compatible with endorsing some forms of relativism. Similarly the scale does not distinguish between agent and appraiser relativism, which are conceptually distinguished from one another, but not recognized by the authors nor incorporated into the way items on the scale are framed. Granted, one can design a scale to assess whether people agree with agent or appraiser relativism individually, but any scale that fails to do so risks collapsing the distinction and losing the ability to detect it, or misleadingly treating low responses as an indication of non-relativism rather than a distinctive rejection of agent or appraiser relativism in particular. In short, there are different types of relativism, and using a single unidimensional scale to capture agreement with "relativism" introduces a host of methodological problems. Even if the scale accurately sorted those who agreed with some form of relativism from those who rejected all forms of relativism, the scale would lack the resolution to tell us *which form of relativism* individuals or people in aggregate endorsed. The more serious issue, however, is that disagreement with items on the scale could reflect disagreement with a particular form of relativism, rather than disagreement with *all* forms of relativism. Just as disagreeing with claims about Christian theism does not entail that you are an atheist (since you could, after all, endorse some other form of theism), so too does disagreeing that "The viewpoint of one's culture determines whether their actions are morally right" not entail that you reject relativism; it only indicates that you reject *cultural* relativism.

Another problem with the inclusion of items that represent cultural relativism and other items that represent individual subjectivism is that these positions are potentially mutually exclusive: one cannot believe both that moral facts are determined by *individual* standards, *and* that they are determined by the standards of one's culture (at least not without a bit of finessing; one could think that statements could be true relative to one *or* the other, but not at the same time and in the same respect). In other words, these are two distinct constructs. Treating the scale as a unidimensional

measure of "relativism" is therefore inappropriate. One ought to measure cultural relativism and individual subjectivism separately.

This also creates a dilemma for inter-item correlations. If responses to the cultural relativism item correlate with responses to the individual subjectivism items, this points to a potential conflation among participants: it would be strange to think both that moral facts are true or false relative to *individual* standards, *and simultaneously* to group standards. Participants could in principle endorse such a view, in which case they may have incoherent or irrational moral standards (Colebrook, 2021; Loeb, 2008). This isn't to say such a view is necessarily incoherent or irrational: one could think that moral claims are true or false when indexed both to the standards of individuals *or* groups. However, we may wonder whether people are genuinely committed to unusual metaethical positions, or whether, instead, they simply don't draw a sharp conceptual distinction in the way philosophers do. If this isn't what they intend to express in endorsing both individual and cultural relativism, then it is unclear what they do mean, and this raises doubts about the validity of these items: if the cultural relativism item measures cultural relativism, and the individual subjectivism items measure individual subjectivism, but these beliefs are, by design, intended to reflect incompatible commitments, then *either* participants who agree with both are expressing incompatible commitments, *or* one or the other of these sets of items do not reflect the intended construct. It is more plausible that participants are crudely responding to items that roughly correspond to appealing-sounding positions and unappealing-sound positions without fully appreciating the meaning of these positions, which may be far richer, more sophisticated and (most importantly) narrow in meaning.

However, all of these possibilities remain live options, and at least some of them would pose challenges to the validity of one or more of the items on the MRS. While CHJFF note that their findings suggest four subdimensions of the relativism factor, including distinct factors for "culture-level relativism" and "individual-level relativism," and observe these two factors and the other two

factors were "highly correlated" (r = 0.45-0.56), this overlooks potential conceptual problems with merely interpreting factor analyses as though they straightforwardly point to the presence of distinct factors or subdimensions of some putative psychological construct. Strictly speaking, there is tension or outright conflict between cultural relativism and individual subjectivism. While we might typically presume that e.g., personality traits such as the big five may correlate with one another but still vary orthogonally because they represent distinct and at least semi-independent features of our personality, it is strange to treat one's metaethical standards as though they could be reasonably expected to vary as though they were personality traits. It's even stranger to treat ostensibly distinct and inconsistent metaethical positions as "subdimensions" of a single, higher-order construct. While both cultural relativism and individual subjectivism may be subcategories of relativism, the respect in which they are subcategories is one of conceptual similarity, not one of each being facets of a single psychological construct.

One could be ambivalent, or uncertain, or sympathetic to cultural relativism and individual subjectivism, and this could manifest as agreement with both, but one could not sincerely and simultaneously endorse two distinct philosophical positions that are mutually exclusive. Given this, if participants do tend to endorse both cultural relativism *and* individual subjectivism, *what exactly does this mean*? What is it that these people are agreeing to? This points to a more general problem with using Likert scales in contexts like these: it is not entirely clear what they are measuring about the individual. Do higher scores reflect degree of confidence in both positions, or in a kind of hybrid pluralistic endorsement, whereby one believes moral claims can be true or false relative to both cultural standards and individual standards, but in ways that do not overlap or conflict with one another, or do people hold conflicting metaethical standards? This may be a nitpicky concern that could broadly apply to a wide range of psychological scales, but if so, I stand by my willingness to pick nits: I don't know what

scores on this scale mean exactly, and it seems reasonable to worry that there is substantial room for interpretative variation that could threaten the validity of at least some of the items on the scale.

Yet another issue with the MRS is that many of the reverse-coded items do not reflect realism, but instead reflect universalism. I have discussed the difference between realism vs. antirealism and universalism vs. relativism previously, so I will not repeat that concern here. Here are the items that reflect universalism:

| **MRS #1** *universalism* | There is a moral standard that all actions should be held to, even if cultures disagree |
| --- | --- |

| **MRS #2** *universalism* | There are moral rules that apply to everyone regardless of personal beliefs |
| --- | --- |

| **MRS #3** *universalism* | The same moral standards should be followed by people from all cultures |
| --- | --- |

Each of these items reflects the position that all people are or should be subject to the same moral standards as everyone else. Unfortunately, this is *not* the same thing as moral realism, since such claims concern the scope of moral standards (i.e., who they apply to) not what makes them true. Unfortunately, two of these items also make reference to stance-independence, with the first stating that moral standards are independent of cultures and the second stating that they're independent of personal beliefs. One obvious problem with both such items is that the first only involves a rejection of cultural relativism, not individual subjectivism, while the second involves a rejection of subjectivism, but not cultural relativism. Technically, a subjectivist thinks there are moral rules we are subject to even if cultures disagree, namely, our own subjective standards. Likewise, cultural relativists think we are subject to certain moral standards regardless of our personal beliefs. Both of these items thus

conflate objections to specific forms of relativism with a rejection to relativism as a whole. More importantly, both inappropriately incorporate stance-independence into the item alongside universalism, resulting in hybrid items that appear to convey both stance-independence and universalism simultaneously.

Yet the more general problem is the emphasis on universalism. This further compromises the utility of the MRS as a measure of realism versus antirealism, since it frames the contrast between the items to be between relativism and universalism, rather than relativism and realism. The universalism/relativism debate concerns whether there is one or multiple moral standards. The realism/relativism debate would consider whether moral claims are non-indexical, or categorical in nature, such that they are true independent of the moral stance of any person or group, whereas relativism would hold that moral claims do contain an indexical element such that their truth can vary relative to the standards of the person making the claim, or that person's culture (or the standards of a person or the culture of a person assessing the claim). These are orthogonal distinctions. By presenting participants with the former contrast, any participant who interprets the set of items holistically will be encouraged not only to interpret items about universalism in a way irrelevant to the debate between realism and antirealism, but to also interpret the relativist in terms of this orthogonal distinction. Thus, items prompting attitudes towards universalism not only cannot serve as direct measures of attitudes towards realism or antirealism, but may prompt participants to interpret *other* items on the same scale to not be measures of realism or antirealism, even if they would have in the absence of the items about universalism. In short, the inclusion of items reflecting universalism threatens the value of the MRS as a measure of realist and antirealist views.

Another problem arises when using items that reflect universalism as reverse-coded indicators of relativism is that such items *directly contribute* to a participant's aggregate relativism score. That is, anti-universalism is simply treated as the same thing as relativism. This effectively embeds anti-

universalism into our indices of "relativism," which dilutes the MRS's value as a measure of relativism-as-opposed-to-realism, further compromising its potential value as a measure of the realism/antirealism distinction. The whole scale, in other words, is designed in such a way that it forces a distinction between relativism and universalism, and bars itself from serving as a suitable measure of a form of antirealism in contrast to realism more generally.

In addition, the MRS presents participants with a forced choice between relativism and non-relativism. Yet participants who do not endorse relativism have no way to express a positive metaethical stance. Their only recourse is to deny relativism. It is unclear from the armchair how folk antirealists would respond to such items. They may agree with relativist items because these express the closest available position to their own view, or they may disagree with these items, since relativism does not strictly-speaking reflect their metaethical stance. For instance, suppose a participant endorses noncognitivism or error theory. How would they respond? A reasonable case could be made for expecting them to express agreement, or disagreement, or even ambivalence by selecting the midpoint. It's entirely plausible that different antirealists would react differently, introducing considerable noise into the measure. And even if they did reliably respond in one or another way, this could lead to interpretive difficulties. If most agree with relativism, this could result in an overestimate of folk relativism, while if most disagree, this could (if interpreted carefully) pose little problem, but if researchers inappropriately interpret non-realist responses to reflect universalism or realism, they could also overestimate the proportion of realists. Even if this problem could be circumvented, the MRS would still lack the resolution to identify distinct metaethical positions.

Imagine, for instance, a scale that only determined whether people believe in the Christian God. This scale might be fine for determine whether people are Christians, but it could not distinguish atheists from people who endorse non-Christian religions. Just the same, even if the MRS could tell us whether people are relativists, it cannot tell us whether non-relativists are realists or endorse some

other form of antirealism. This doesn't mean the scale is invalid, but it does mean that, at best, it would still be limited in what it could tell us about folk metaethics. This is especially worrisome given that Davis (2021) found that noncognitivism was the modal response among participants using his paradigm, and a very small proportion endorsed error theory. If any significant number of participants endorse such views, they are invisible to this scale, and they cannot be distinguished from relativists or realists.

And, of course, the scale cannot tell us what form of realism any particular realist respondents endorse. Again, this does not mean the scale is invalid, but it does mean it cannot provide much information about the specific metaethical stance of non-relativist participants, and may provide an inaccurate estimate of relativist participants insofar as non-relativist antirealists agree with relativism because it most closely reflects their views. Yet another problem with the limited resolution of the MRS, and a forced contrast between relativism and universalism is that it permits us to interpret participants as varying in degree of agreement with relativism. By its very nature, assessing aggregate scores on the MRS does not allow us to detect ambivalence or conflicting metaethical standards, which in turn bars us from detecting metaethical pluralism. This is reinforced by the exclusive inclusion of *general* and *abstract* moral items. That is, the MRS does not include any specific, concrete moral issues. Yet research that does utilize concrete items (e.g., distinct questions about the participant's metaethical stance towards abortion, murder, stealing, and so on) consistently reveals that most participants express different metaethical stances towards different issues. *If* participants are interpreting these questions as intended, this would suggest that most participants are metaethical pluralists. Yet the MRS is incapable of detecting this type of pluralism. Since most folk metaethics research supports metaethical pluralism, this means that the MRS is incapable of detecting one of the most prominent working hypotheses in contemporary research on the psychology of folk metaethics. Metaethical pluralism could, in principle, be a consistent position to hold, but it could also indicate that people

have ambivalent or inconsistent metaethical commitments that are most readily prompted when confronted with distinct moral issues, rather than considering morality in the abstract.

The MRS cannot readily pick up on such pluralism, ambivalence, or inconsistency. Perhaps researchers could examine the degree of internal variation in how participants responded to individual items, but the scale is unlikely to be used in this way, and such data would be unlikely to provide any direct and unambiguous insights into the degree to which participants have pluralistic or conflicting metaethical commitments. This would also be unlikely to occur due to the way scales are constructed. After all, the very process of selecting the ten items that appeared on the MRS was predicated in part on how well responses to one item correlated with the other. Furthermore, scale design by its very nature can inflate intrapersonal consistency in item response, which would result in the exaggerated appearance of consistency.

In short, the very nature of a scale, and the scale construction procedures used to validate a scale, will tend to self-select for highly correlated items and to inflate reliability, creating the artificial appearance of greater consistency in participant responses. To test this, researchers could run the MRS scale on a pool of participants who are also given classic versions of the disagreement paradigm. Whatever scores they provide on the MRS, I predict that they would also judge some concrete moral issues to be "objective" and others to be "relative." If so, MRS would appear to give a single measure of their agreement with relativism, while the disagreement paradigm would indicate that the exact same subject endorsed metaethical pluralism, or had an inconsistent or ambivalent metaethical stance. In short, the MRS and other unidimensional scales that aggregate scores into a single measure per participant are poorly suited for detecting some potential ways that people may think about the topic of study.

In principle, a scale may be able to circumvent this problem by designing items for different concrete moral issues. For instance, instead of asking participants "Different people can have

opposing views on what is moral and immoral without anyone being wrong," you could instead present them with multiple versions of this problem, each addressing a different moral issue:

(1) Different people can have opposing views on [whether abortion] is moral [or] immoral without anyone being wrong

(2) Different people can have opposing views on [whether stealing] is moral [or] immoral without anyone being wrong

However, the MRS does not do this, and even if it did, if participants responded differently to these items, it's unlikely they could be expected to load onto what CHJFF proclaim to be a "robust single factor." On the contrary, the MRS relies exclusively on abstract and general statements about the moral domain as a whole, thereby prohibiting piecemeal judgments about individual moral issues. By design, the MRS treats each participant's metaethical stances as uniform across all moral issues. And, of course, all participants must select *some* response, so everyone will be treated as having some determinate metaethical stance. Thus, the MRS may serve to reinforce the UD assumption, even if the UD assumption does not accurately capture folk metaethics.

Three of the ten items are variations on the disagreement paradigm. While there may be some differences in the methodological shortcomings of items that describe a generic moral disagreement that does not specify any particular moral issue, and the items that are used in standard disagreement paradigms, which describe concrete moral issues such as abortion and stealing, all three of these items may inherit many (perhaps most) of the methodological problems associated with the disagreement paradigm, e.g., conflating relativism with sensitivity to context and realism with rigidity and absolutism, interpreting the question in epistemic terms, thinking that different cultures could each find appropriate ways of conforming to the same abstract moral principle, and so on. Given the legion of problems with the disagreement paradigm, the fact that nearly a third of the scale is composed of such items is a threat to the validity of the MRS.

In addition, items on the scale make use of the term "right" and "wrong" both to refer to something being *morally right* or *morally wrong* and to refer to something being *correct* or *incorrect* in the conventional sense, i.e., without distinctively referring to morality or to normativity. We may refer to the former as the *normative* usage of "right" and "wrong" and the latter as the *non-normative* usage of "right" and "wrong" (perhaps most closely reflecting a truth-correspondent notion of the terms). In fact, one item uses both the normative and non-normative versions of these terms: "People can disagree on what is *morally right* without anyone being *wrong*" (emphasis mine).

This is potentially confusing since it could cause some participants to confuse normative and non-normative usages with one another. For instance, someone may interpret this item to indicate that it isn't *morally wrong* for people to hold opposing moral views. It is unclear whether any participants in fact did interpret this item in this unintended way, but there is little justification for using the same terms to refer to distinct concepts when interpreting them in precisely the intended way is critical for the validity of the item. In fact, they even use the phrase "morally correct." It is not completely obvious whether they mean "correct" in a normative or non-normative way, but the normative interpretation seems more plausible. However, this means that they not only use "right" and "wrong" in a non-normative context, they also use "correct," which is more commonly used in a non-normative way, as a normative term. This may serve as only a minor cognitive hurdle for participants, and it could turn out that they readily navigate this vacillation in terminology, yet it is still not ideal. However, my primary motivation in drawing attention to these terms is twofold. First, two items employ scare quotes without any obvious justification. This is troubling because other items don't use scare quotes, yet presumably intend for the terminology to be interpreted in the same way, and because they one of the items uses "right" non-normatively (to mean *correct*) while the other uses "right" and "wrong" normatively (to mean *morally* right and wrong). As a whole, there is no consistent usage of quotations to distinguish e.g., normative from non-normative use of these terms, nor are the quotes used

consistently across items. This introduces, in a seemingly arbitrary and unprincipled way, an inconsistency in the way items are presented for which no justification is provided. This is a problem all on its own, in addition to the problems associated with the two items that make use of scare quotes—a strange choice with little justification.

My primary interest in drawing attention to this shift between normative and non-normative usages of the same term is to draw attention to the broader problem that terms and phrases that appear in the MRS could be interpreted in a variety of ways that differ from the intended interpretation. For instance, what does it mean for one's culture to "determine" moral standards, or for someone to be "the final authority" on the moral status of their actions? If metaethical considerations are salient when considering these items, they could be reasonably interpreted in metaethical terms. Yet to laypeople, metaethical considerations may not be as salient as other potential considerations. Consider the statement "The viewpoint of one's culture determines whether their actions are morally right." This statement employs the term "determines," but determines *how* exactly? *What does that mean*? Even specialists in metaethics would struggle to articulate a distinctively metaethical interpretation of these terms that could successfully circumvent various misunderstandings and ambiguities that arise *even in the context of academic papers intended for specialists*.

The problem here is a big one, and it is one philosophers and psychologists attempting to adapt philosophical concepts to psychological research have completely failed to grapple with: philosophers use conventional terms and phrases, but these are often loose, grasping, and underspecified attempts to gesture at sophisticated and highly specific concepts that those conventional terms do not precisely and unambiguously refer to in everyday conversation. Does "determines" tend to mean "serves as a grounding that makes a propositional claim true in a relativistic way indexed to the person or group who made the determination"? No. This is not at all what "determines" typically means. Yet this is *precisely* how participants would have to interpret it in items

that made use of the phrase "determines" in order for the item to be valid. Few laypeople could be reasonably expected to interpret "determines" in this way, even if they were given adequate context. And they are not. I'm not even confident people with philosophical training would overwhelmingly interpret it as intended. I have yet to see any systematic study of competence with metaethical concepts among academic philosophers, but most do not specialize in metaethics, and may have a biased, distorted, or inadequate understanding of relativism as a metaethical position.

Another problem with most of the items on the MRS is that, by failing to disambiguate agent vs. appraiser relativism, and cultural relativism vs. individual subjectivism, items often require a participant to express agreement with distinct forms of relativism, while disagreement is interpreted as rejection of relativism. However, participants may reject particular forms of relativism without necessarily rejecting others. As such, one of the odd features of the MRS is that rejecting scale items is consistent with relativism, insofar as specific items only express particular forms of relativism.

Several of the items on the MRS also lack face validity, despite the fact that they were judged by a panel of experts to be appropriate measures of relativism. First, it is worth emphasizing again that expert judgment, at least by itself, may be inappropriate for assessing face validity: it does not matter how well an item reflects the construct of interest according to an expert. What matters is whether *participants interpret the item as intended.* Even if experts interpret items as intended, it does not necessarily follow that sample populations will. And since experts often have contextual and background knowledge related to the prospective construct the scale is intended to capture, experts may suffer from the "curse of knowledge": they may mistakenly infer that because the items in question appear to represent the construct *to them*, that non-experts would interpret those items in the same way.

This risk is amplified by requesting expert judgment on how well an item represents, which focuses on the expert's expertise regarding the domain. This is altogether different from asking whether the expert thinks the item in question is one that they would expect laypeople to understand

in the same way. A focus on the latter may render potential interpretative pitfalls more salient to expert judges, and induce them to consider how people who lack their training and knowledge would interpret the items. But even if the risk that participants would interpret items in unintended ways were made explicit to experts, it is unclear whether they are in a position to accurately predict how laypeople would understand the items, due in part to the inherent difficulties in predicting how a nebulous and underdescribed population would interpret the items, and because they may be incapable of fully suppressing the curse of knowledge.[116]

In short, even if every item on the MRS was the best representation of relativism that one could reasonably convey in simple sentences, this would still not ensure that participants responding to the question understood it as intended. Expert judgment isn't enough for the same reason that "neutrinos are fermions with a spin of ½" would not be an appropriate statement to present to most populations. Physicists would be correct in judging this item to be an accurate representation of particle physics, yet almost everyone else would find it uninterpretable.

This is obvious enough in the case of statements that explicitly employ technical scientific jargon, but questions about relativism can slip under the radar both because they can be framed without employing obvious jargon and because relativism (and other metaethical concepts) are presumed to be part of ordinary thought and discourse. In other words, statements about relativism employ conventional language, but this may hinder people from interpreting questions about relativism as intended precisely *because* the intended meaning could be readily conflated with a host of alternative interpretations. Most people don't already have a number of options available for how to interpret "fermion," but there may be considerable heterogeneity in people's conception of "right,"

---

[116] For comparison, imagine an expert chess player attempting to think through how a novice would view a position in chess. Would chess experts be able to readily predict what moves any given novice would make? Perhaps, but their expertise may *impair* their ability to make poor moves, since their awareness of superior moves may be involuntary, and could impede their ability to devise rationales for moves that they recognize to be blunders.

"wrong," "disagree," and other terms and phrases featured in the MRS, opening these items up to interpretive variation.

CHJFF not only employ conventional language, but also suggest that ordinary people are familiar with the notion of moral relativism, claiming that moral relativism "[...] is a frequent topic of public discourse" (p. 2). However, their only sources are Gowans (2021) and Merritt (2016). Gowans simply states that moral relativism is "widely discussed outside philosophy (for example, by politicians and religious leaders)," but provides no further comment on its prevalence among nonphilosophers or discussion of *how* it is understood outside philosophy. The reference to Merritt is an article in *The Atlantic* that quotes former United States Speaker of the House Paul Ryan, who stated that "If you ask me what the biggest problem in America is, I'm not going to tell you debt, deficits, statistics, economics—I'll tell you it's moral relativism." I'd be willing to bet Ryan's conception of relativism would not match the narrow conception of the term as it is used in contemporary metaethics that the MRS is intended to operationalize, but even if Ryan did understand relativism in the same way as academic philosophers, the mere fact that he occasionally references is not compelling evidence that relativism is a common topic of public discourse. The degree to which relativism, understood as the notion that moral claims are true or false relative to the standards of individuals or groups, is an empirical question, and cannot be assumed without adequate evidence. While it would be easy to gather instances of public figures making references to "relativism," it is less clear that such references are intended to convey the specific academic conception of metaethical relativism the MRS is intended to measure, nor is it clear how audiences interpret these references, or how much they have permeated popular consciousness to a sufficient extent that we could be sure people were familiar with the concept of relativism.

In drawing attention to the shortcomings with these references, I don't mean to suggest Gowans or others are mistaken. Rather, I mean to indicate that the claim that a particular concept is

a common topic of public discourse is *vague* and *imprecise*. How confident are we that the particular concept in mind is actually the one featured in public discourse? *How* prominent is the topic? *How* well understood is it? Is it so well understood that any vague statement intended to describe relativism would be readily interpreted by a random sampling of survey participants as intended?

We do not need substantive evidence to demonstrate that *some* notion of "moral relativism" is a part of our cultural zeitgeist.[117] But whatever people have in mind by this idea, it may be muddled or confounded with notions irrelevant to relativism as a narrow metaethical stance about the indexicality of moral claims. And it may be prone to prompt contingent associations with particular cultural and ideological motifs that survey participants want to identify with or distance themselves from for reasons unrelated to having any kind of genuine understanding of and commitment to a coherent metaethical position. The mere fact that we can identify references to "moral relativism" among public figures is meager evidence at best that participants would have an easy time understanding questions on a relativism scale.

References to "relativism" could prompt conservatives to think of postmodernism, the "far left," amorality, debauchery, an utter disregard for the lives of others, a lack of commitment to justice or virtue, or other notions they find repugnant, prompting them to disagree with anything that has a whiff of the opposing tribe to it. Conversely, people with progressive political perspectives may associate relativism with their own political identity, and with values they endorse, such as tolerance and sensitivity to cultural differences. Ideological differences could cause people to hold different metaethical standards, and this could in turn be reflected in their responses to the MRS. Yet ideological differences could just as readily prompt people to associate survey items with particular ideological

---

[117] The notion that metaethics items could prompt ideological associations that influence survey response emerged from a discussion with Tyler Millhouse (personal communication), who presented this possibility as an additional challenge to the validity of folk metaethics paradigms.

stances, which could prompt them to respond in ways that serve more as a signal of their ideological stance or political leanings than their views towards metaethics.

Even if we were confident participant's ideological associations were not influencing their responses, it would still not be clear that the references politicians and religious leaders make to "moral relativism" closely match relativism *as a specific technical term in contemporary analytic metaethics*. Such public pronouncements may characterize "relativism" in ways that are underspecified or ambiguous, or are framed in ways that connote or explicitly incorporate a broader range of content in the notion of "relativism" than is strictly entailed by metaethical relativism. Also, the mere fact that politicians and religious leaders sometimes refer to moral relativism or use the term "moral relativism" does not provide us with much information about how familiar ordinary people are with metaethical relativism or how salient or relevant the concept is to everyday thought and behavior. In short, it's not at all obvious whether the term "moral relativism" or concepts of relativism circulating in public discourse reflect relativism as it is understood by researchers. We are not entitled to assume that people would understand questions about relativism in the same way as researchers intend without evidence that they in fact interpret them this way.

In sum, without empirical evidence there is little reason to presume widespread familiarity with relativism among lay populations nor, if there is, that such understanding closely approximates the specific conception of relativism presupposed by the MRS. Furthermore, the specific terms and turns of phrase employed by the MRS may be familiar to specialists familiar with metaethical distinctions, but such phrasing is subject to a variety of interpretations that differ from the intended interpretation. In the next section, I address each of the items on the MRS, and argue that all items on the MRS could plausibly be interpreted in a variety of ways unrelated to the intended interpretation.

| **MRS #1** *relativism* | Different people can have opposing views on what is moral and immoral without anyone being wrong. |
|---|---|

This item is an abstract variant of the disagreement paradigm. It inherits many of the problems associated with the disagreement paradigm (see **Chapter 2**).

| **MRS #2** *relativism* | People can disagree on what is morally right without anyone being wrong. |
|---|---|

This item is an abstract variant of the disagreement paradigm. It is also almost identical in structure to MRS #1, and is therefore subject to the same criticisms (see **Chapter 2**).

| **MRS #3** *relativism* | Two different cultures could have dissimilar moral rules and both be "right." |
|---|---|

This item is an abstract variant of the disagreement paradigm that is framed in terms of intercultural disagreement rather than interpersonal disagreement. In addition to inheriting many of the shortcomings associated with the disagreement paradigm, there are a few other shortcomings to this item. First, the use of scare quotes around *right* is strange. What does it mean for dissimilar moral rules to both be "right," rather than right (without quotes)? Scare quotes are used for a variety of reasons, but almost all of those reasons are inappropriate for the MRS. Scare quotes are often used to signal that the word or phrase originated with someone other than the author, e.g., *According to Alex, it was "the best party ever."* (Trask, 1997). Such usage involves neither endorsement nor opposition to the terms, but is simply used to indicate that the precise phrasing did not originate with the author. In other cases, scare quotes are used to pragmatically imply disapproval, or a mocking attitude towards the term or phrase. There are a variety of ways people employ scare quotes:

Scare quotes are quotation marks placed around a word or phrase from which you, the writer, wish to distance yourself because you consider that word or phrase to be odd or inappropriate for some reason. Possibly you regard it as too colloquial for formal writing; possibly you think it's unfamiliar or mysterious; possibly you consider it to be inaccurate or misleading; possibly you believe it's just plain wrong. Quite often scare quotes are used to express irony or sarcasm [...] It is important to realize this distancing effect of scare quotes. Quotation marks are not properly used merely in order to draw attention to words, and all those pubs which declare We Sell "Traditional Pub Food" are unwittingly suggesting to a literate reader that they are in fact serving up microwaved sludge. (Trask, 1997)

These ways of employing scare quotes are not only inappropriate for items on the MRS, but could have a decidedly detrimental effect on interpretation, in that they could signal that the people who designed the study wish to distance themselves from the terms and concepts used in the scale. Yet a far more troubling concern is that putting the term "right" in scare quotes could encourage participants to interpret the notion of rightness as one that is *itself* relative to the different cultures. Far from facilitating the correct understanding of cultural relativism, this would in fact undermine it. Cultural relativism holds that a given moral standard is morally right or wrong relative to the standards of each culture. While this means that the truth status of *moral claims* is relativized to different cultural standards, it *does not* mean that the concept of truth is *itself* relativized to each culture. Yet by putting "right" in scare quotes, the scale could signal that the sense in which cultures could have different moral views yet still be right could be right or wrong not merely relative to their moral standards, but relative to their standards of truth, as well (or instead). If so, this is *not* metaethical relativism. To illustrate, when Alex says, "I am Alex," and Sam says, "I am Sam," both statements are true, because each statement is indexicalized in such a way that its truth status is determined by the person making the statement. Yet it is not necessarily the case that the fact that Alex is Alex and Sam is Sam is only true relative to a particular standard of truth, and could be false relative to some other standard of truth; that would be relativism about different truth standards, rather than relativism about the meaning of indexicalized assertions of one's name. Just the same, if Alex thinks that murder is wrong,

and Sam thinks that murder is not wrong, then it would be true for Alex to say "murder is wrong," and false for Sam to say "murder is wrong," but the sense in which these claims are true or false, respectively, does not itself need to be relativized (though it *could* be; metaethical relativism is compatible with truth relativism). In short, metaethical relativism holds that moral claims can be true or false relative to the standards of different cultures, it does not hold that truth is *itself* relative to the truth standards of different cultures. By conflating the two, the MRS may actively encourage a straightforwardly unintended interpretation of the item.

This is compounded by a related, and more plausible possibility. Rather than the item indicating endorsement of truth relativism, it could be interpreted as the *descriptive* claim that different cultures can both *consider* their respective moral rules to be "right," whether or not the participant themselves agrees. In other words, scare quotes could be used to indicate that dissimilar cultures consider themselves right, *even if the authors of the MRS or the participant disagree*. After all, when scare quotes are used in conventional settings, signaling one's disagreement or at least lack of endorsement of the term or phrase is frequently *the whole point of using the scare quotes in the first place*. The use of scare quotes is inappropriate for the MRS, and is probably inappropriate for most scales.

Finally, "dissimilar" is a bit obscure of a phrase, and not necessarily adequate for the purposes of measuring moral relativism. Dissimilar does not necessarily mean conflicting, and only conflicting moral beliefs are appropriate candidates for the kinds of moral claims that could be relevant to realism. A realist who believes there is a single standard of moral truth could still believe that dissimilar (but not necessarily conflicting) actions are consistent with conforming to the same moral standard. For example, if one society punishes thieves by forcing them to do community service or work to pay back their victims, while another society uses jail time, both of these could be considered just punishments for the same crime: different rules, both morally acceptable.

Moral realism does not require extraordinary rigidity in the way we conform to moral rules. Such permissiveness does not entail that there are two distinct standards of moral truth regarding the just punishment for theft. Rather, the moral rule could be more abstract, e.g., it could be that we have a moral duty to punish thieves in accordance with the standards and conventions mutually agreed upon by the society in which the crime occurs, within parameters for what would be excessive or inadequate punishment (which could, in principle, be specified). Such a moral rule could be invariant and true independent of the standards of any person or culture, and could entail that execution and letting the person free without punishment would be unjust, but that a wide range of potential punishments are all just.[118]

In short, moral realism is consistent with different cultures having "dissimilar" moral rules. Moral realism does not allow for conflicting moral claims to be true, but moral rules can be *dissimilar* without *conflicting* with one another.

| **MRS #4** *relativism* | One's own culture determines whether that person's actions are "right" or "wrong." |
|---|---|

Like MRS #3, this item employs scare quotes. Since this item employs the normative rather than non-normative version of right and wrong, it does not suffer from the problem of potentially implying relativism about standards of truth. Even so, the use of scare quotes is potentially misleading and shares most of the same concerns as MRS #3. When asked to explain what this statement means, a

---

[118] There is something puzzling about the need to explain that moral realism does not require rigidity. After all, we have little trouble recognizing how one might achieve an outcome via different means in other domains. Suppose you agreed to make a dessert for a party. It clearly does not follow that there is one, and only one possible dessert that you would have to make, and that any alternative to making that one specific dessert would be a failure to uphold your responsibility to make dessert. There are many ways to comply with the duty to make a dessert: you could bring cake, flan, baklava, or even croquembouche. None of us are confused about this, or inclined to think that a duty to make dessert requires making a *specific* dessert. Yet many people seem to conflate moral realism with the notion that we have *specific* moral duties. Moral realism has *nothing to do* with the level of specificity of moral rules; the generality or specificity of moral rules is a matter of normative ethics that is orthogonal whether those norms are stance-independently true.

few participants made explicit use of the scare quotes in their response, and this did appear to influence their interpretations, which were not consistent with the intended interpretation:

> *Culture defines the set of rules which establishes the social hierarchy of a civilization. Those things which advance an individual in the hierarchy are considered "right" while those things which reduce the individuals status are "wrong".*

> *the society around you decides the morality of your actions. so what might be "wrong" in kenya could be perfectly acceptable in cambodia.*

Both interpretations are descriptive rather than metaethical, and are consistent with the typical way scare quotes are used to signify another's endorsement of the terms, rather than the participant's. But cultural relativism isn't the descriptive claim that one culture may consider a moral rule to be right (in the normative sense), while another considers it wrong (in the normative sense); it is the metaethical thesis that the rule *is in fact right* relative to one culture's standards, and *in fact wrong* relative to the other's standards. In other words, if I said that Alex believes X is true, and Sam believes not-X is true, I am merely making a claim about their beliefs. This is not what relativism holds. Rather, relativism holds that in virtue of Alex believing X is true, X *is true*…relative to Alex's standards, while in virtue of Sam believing not-X is true, not-X *is true* relative to Sam's standards. MRS #4 does not clearly distinguish this from the mundane descriptive claim that different people have different moral beliefs, and this conflation is reflected in the way participants interpreted the item.

Another problem with this item is the use of "determines." This was one of the phrases I used as an example earlier to illustrate how the MRS employs terms that appear appropriate because they are conventional, everyday words. Yet these terms are inappropriate precisely *because* they are conventional, everyday terms. Such terms have mean a variety of different things depending on the context in which they are used, and most of these meanings have little to do with their use in a specifically metaethical context. To determine that an action is right or wrong could mean to make a *normative judgment or decision* about that action's moral status. If so, this item could express little more

than the descriptive claim that people are morally judged by their societies. *Determines* could also be used to refer to discovering, or figuring out what actions are morally right or wrong. If so, it could be interpreted to convey an *epistemic* claim, not a metaethical one. Coupled with the scare quotes, the notion that one's culture *determines* the moral status of their actions could also be understood to convey the descriptive claim that one's moral standards are shaped by their culture. None of these interpretations have anything to do with metaethics, and yet they seem more consistent with how participants described this item. Here are several examples of what participants said, when asked to explain what this item means:

> *(i) This means that right or wrong is influenced by culture, not a natural born belief.*

> *(ii) This statement means beliefs that have been passed down for generations that a person has been taught by their parents that's specific to their culture. For example, some cultures believe that eating pork is forbidden.*

> *(iii) Different cultures have different beliefs of right and wrong.*

> *(iv) The environment someone is raised in determines their moral compass.*

All of these participants interpreted this item in descriptive or etiological terms, *not* metaethical terms, and such descriptive interpretations were more common than metaethical ones. I don't know if the specific use of "determines" contributed to these interpretations, but it probably didn't help.

Finally, recall that the MRS does not distinguish between agent and appraiser relativism (Quintelier, De Smet, & Fessler, 2014). This would be fine if the items used in the MRS were consistent with both forms of relativism, but item #4 makes more sense as an expression of agent relativism rather than appraiser relativism. By stating that one's own culture determines whether *that person's* actions are right or wrong, this implies that moral facts are relativized to whichever standards the agent is subject to. However, appraiser relativism holds that the truth of moral claims is relativized to the person judging the action, which could be someone who is *not* a member of their culture. By implying that whether the person's actions are right is fixed by the standards of their culture, MRS #4 would

imply that if someone from another culture judged their actions to be wrong, that person would be mistaken, which conflicts with what appraiser relativism holds. In order to reflect appraiser relativism, the item would have to say suggest that the person's culture determines the truth status of their moral judgments, *including their judgments of others*, while leaving open the possibility that other people judging that person's actions according to their own culture's moral standards could also be correct relative to their own culture. I have no data regarding which of the two forms of relativism are more common among philosophers or nonphilosophers. With respect to ordinary people, Quintelier, De Smet, and Fessler (2014) found that the moral standards of both the agent and the appraiser influenced participants' judgments, suggesting yet another way participants may be confused or interpret questions about metaethics in unintended ways, or may have pluralistic metaethical standards. With respect to philosophers, Gowans states that "Appraiser relativism is the more common position," and even assumes an appraiser account of relativism in the Stanford Encyclopedia of Philosophy entry on the topic. This provides at least some indication that appraiser relativism may be the norm among philosophers, yet this item presupposes agent relativism. Regardless of which of the two forms of relativism is more common among philosophers, this item is more consistent with one than the other. Even so, whether ordinary people are appraiser or agent relativists may be of interest, but if they were one or the other they'd still be relativists and antirealists, which would still be of theoretical interest. Yet imprecision and conflation between the two hints at least some degree of oversight on the part of researchers, raises at least some minor concerns about the validity of the scale items and how best to interpret responses to them, and insofar as items on this scale may be interpreted in some cases in appraiser terms and in others in agent terms, this may confuse participants or lead to results that are less precisely than in the absence of such inconsistency. Finally, note that if someone disagrees with this item, this does *not* entail that they aren't a relativist. It only means that they don't endorse cultural relativism in particular.

| MRS #5 *relativism* | The viewpoint of one's culture determines whether their actions are morally right. |
| --- | --- |

MRS #5 has many of the same problems as MRS #4: it uses the term "determines," disagreement with this item is consistent with forms of relativism that do not relativize moral claims to cultural standards (e.g., individual subjectivism), and it focuses specifically on agent rather than appraiser relativism. It is also almost identical in meaning to MRS #4, with the pair serving as reasonable candidates for redundancy.

| MRS #6 *relativism* | There is a moral standard that all actions should be held to, even if cultures disagree. (R) |
| --- | --- |

There are a few initial oddities about this item. First, the notion of cultures disagreeing is a little strange. A much more serious problem is an ambiguity in the way the question is worded. It states that "There is a moral standard that all actions should be held to," then follows this with "even if cultures disagree." But *what are these cultures disagreeing with*? There are at least two possibilities:

(1) They have normative moral disagreements with other cultures

(2) They disagree with the *metaethical* claim that there is a moral standard that all actions should be held to

Only the former would be consistent with the item serving as a valid measure of relativism. The second, on the other hand, would effectively render this a question about something else entirely: whether there are universal moral standards even if some cultures disagree that there are universal moral standards. Even if one believes there are universal moral standards, it is strange to express this by suggesting that there would be universal standards even if some cultures deny this. Yet insofar as relativism is contrasted with universalism for the purposes of this study, agreement with this item could be a valid indication that the participant rejects relativism.

A more serious problem emerges when considering what disagreement with this would mean. Someone who disagrees that there are universal moral standards, even if some cultures disagree, is not necessarily endorsing any form of moral relativism. An error theorist or noncognitivist may also deny that there are universal moral rules without thereby endorsing relativism, yet this item is reverse-coded, such that disagreement is treated as an indication of relativism. This is not appropriate, since disagreement with this item cannot disambiguate people who endorse relativism from those who simply deny universalism. In fact, even a realist could disagree with this item, since people could believe that there are stance-independent moral rules, but that these rules are not universal. This highlights a general problem with the MRS that applies to all reverse-coded items on the scale: all three reverse-coded items express universalism. Yet the negation of universalism does not entail relativism, but instead simply entails non-universalism, which is compatible with virtually all forms of antirealism and with some forms of realism.

In addition, the fact that cultures disagree that there are universal moral standards may be irrelevant to many moral realists. Cultural relativism is a view about whether the *normative* moral standards are true relative to different cultures; it has nothing to do with the *metaethical* views of those cultures, so whether cultures disagree that there are universal moral standards is completely irrelevant to whether moral relativism is true. As such, it makes no sense to frame this question as a conditional, as though whether cultures disagreed that there were universal moral standards would be relevant to relativism. To do so is to saddle the item with cognitively demanding and irrelevant considerations.

This item may also be mistakenly treated by anyone who uses the scale as a reverse-coded item indicating endorsement of cultural relativism, yet this is not actually the case. Suppose you are an individual subjectivist, so you believe moral standards are true or false relative to the standards of individuals. If you were asked whether there is one moral standard that all actions should be held to, you would say "no." And you would still say "no," even if cultures disagreed that there was such a

standard, and even if they disagreed with one another (since the normative and metanormative positions different cultures hold are irrelevant to you). Thus, the negation of this item cannot disambiguate individual subjectivists from cultural relativists, but generates the illusion of doing so.

In sum, one of the central flaws of this item is that it ends with the phrase "even if cultures disagree" without specifying what these cultures disagree about. Typically, when one follows a statement by saying "even if group X disagree" this would imply disagreement with the preceding statement. Yet the preceding statement is a metaethical assertion, so if item were interpreted in this way, it would not be a valid measure relativism. Thus, the only way this item could serve as a valid measure of relativism is if participants systematically interpreted it in an unconventional way. In practice, participants may actually do this, since the alternative interpretation is strange. Yet items should not be constructed in such a way that the phrasing used would conventionally be used to convey something other than what is intended. To do so needlessly burdens participants with the added cognitive demand of figuring out what they're being asked. For a topic already so fraught with interpretive difficulties, this is an especially undesirable outcome.

Another potential shortcoming of an item like this is that the notion that there is "a moral standard" that all actions could be held to could conflate moral realism with moral absolutism, specifically of a form where there is only one way to properly conform to a moral rule. For instance, one could hold that everyone should be held to the standard that *you should never lie*. Such a position has a metaethical component (its *scope*: it applies to everyone) and a normative component (the *content* of the moral rule: there is only one proper way to conform to the rule). Participants consistently conflate realism with absolutism, rigidity, insensitivity to context, or the denial that there is more than one way to conform to the same moral rule in other paradigms, so it is plausible that some participants would do so when responding to this item as well.

Finally, stance-independence and universality are orthogonal distinctions. By contrasting relativism with universalism (i.e., that there are multiple moral standards or only one moral standard), items on the MRS that contrast universality with relativism present a continuum of belief orthogonal to stance-independence and that instead focuses on the number of moral systems. Yet other items emphasize the role relativism plays in determining moral truth, which seems to put emphasis on relativism as a form of stance-dependence. In other words, some items treat relativism and its negation as a choice between whether moral standards are made true by stance-dependent or stance-independent facts, while other items treat relativism as the claim that there is more than one standard versus the claim that there is only one standard. Yet these are orthogonal distinctions. There is no problem in principle with treating relativism as the conjunct of the claim that moral standards are stance-dependent *and* that there is more than one standard of moral truth, the authors do not seem to recognize this or treat these views as distinct concepts.

A more serious problem, however, is that one's endorsement of stance-dependence does not entail belief in a plurality of moral standards, since one could endorse a relation-designating account of morality (such as ideal observer theory or relation-designating forms of divine command theory), while endorsement of the existence of multiple moral standards does not necessarily entail stance-dependence.[119] In practice, one or both of these disambiguated stances may be rare or absent among ordinary people, but that is still an empirical question that should be established in advance of running the two notions together by treating them as a single construct.

---

[119] It is harder to identify plausible instances of this kind. Here's one example: suppose you favor a teleological account whereby humans have an obligation to flourish, and to do so requires, in effect, being a "good human." Facts about what promotes human flourishing may be contingent on features of our biology and psychology. Yet the same facts would not necessarily apply to nonhuman moral agents. Nonhuman moral agents may have an obligation to act so as to promote whatever would constitute flourishing for their species. If so, then there could be more than one set of correct moral standards, and those standards would be species-relative, yet they would not be stance-dependent.

| MRS #7 *relativism* | Each person is the final authority on whether his or her actions really are morally correct. |
|---|---|

This problem exemplifies a common theme with many items in metaethics paradigms. It employs language that might be properly understood by specialists in a context in which metaethical considerations are salient, but proper interpretation requires them to understand conventional terms in narrow technical ways that would not be obvious to ordinary people. In particular, the problem concerns the phrase "final authority." Given this term, this item *could* be interpreted as a statement of individual subjectivism, but it would do so only if the notion of a "final authority" is understood to mean that each individual's subjective moral standards serve as a truth-maker for moral claims when they are made by that person. That is, if murder is inconsistent with Alex's moral standards, then when Alex judges that "murder is wrong," this statement is *made true by the fact that murder is inconsistent with Alex's moral standards*. Alex is the "final authority" precisely in the sense that Alex's own standards are what make the claim "murder is wrong" true when indexed to Alex's moral standards.

Yet is this the best, or only plausible interpretation of what it would mean to by the "final authority" about whether our actions are right or wrong? Not at all. This statement could plausibly be interpreted to reflect the *normative* view that those who perform an action are the best or only appropriate judge of their actions, though their privileged status in judging their actions could also (or instead) be grounded in their privileged epistemic access to their actions and the motivations that prompted them. If so, this statement could be interpreted partially or even exclusively as a normative or epistemic statement, not a metaethical one. Of course, I am simply speculating about potential ways participants could interpret this item and the phrase "final authority." How people interpret this item is an empirical question, and unfortunately the item would require so specific and sophisticated an

interpretation of this that there is little reason to be confident that nonphilosophers would interpret it as intended.

| **MRS #8** *relativism* | An action is only morally wrong if a person believes it is morally wrong. |
|---|---|

An appraiser relativist holds that moral claims are true or false relative to the standards of the person expressing a moral judgment about an action, while agent relativism holds that moral standards are true or false relative to the standards of the agents performing those actions.

> Suppose Alex and Sam both steal from a store. Alex believes stealing is wrong, and Sam believes stealing is not wrong. Clarke and Jesse see Alex and Sam steal. Clarke believes that stealing is wrong, but Jesse believes that stealing is not wrong. Clarke claims that Alex and Sam both did something morally wrong, while Jesse claims that neither Alex nor Sam did something morally wrong.[120]

As Quintelier, De Smet, & Fessler (2014) point out, there are at least two distinct frames of reference we could appeal to when evaluating the moral status of Sam's and Alex's actions. Appraiser relativism holds that a moral judgment is true or false relative to the standards of the person expressing the judgment. Since Clarke believes stealing is wrong, but Jesse does not, then Alex and Sam's actions are both wrong relative to Clarke's standards, while neither are wrong relative to Jesse's standards. To an appraiser relativist, the fact that Alex thinks stealing is wrong, but Sam does not, is irrelevant to the truth status of moral judgments about their actions. This is a bit tricky, however, since Alex and Sam could express judgments about their own actions, and one another's actions, as well. For instance, Alex may steal, and then judge their own actions as morally wrong. Now, Alex serves as both the agent and the appraiser. Yet stealing is wrong relative to Alex's standards vis-a-vis Alex's role as an appraiser of her own actions, and not as the agent who performed the act.

---

[120] This example is simply my own version of a functionally identical example presented in Quintelier et al. (2014). I do not know if prior examples like these appear in the literature, so I am happy to give them credit for constructing examples of this form.

An agent relativist, on the other hand, would hold that the appropriate frame of reference would be the standards of the people forming the actions. Thus, it is morally wrong for Alex to steal, but it is not morally wrong for Sam to steal, because Alex believes stealing is wrong but Sam does not. Recall that Clarke judged both Sam and Alex's actions to be wrong, while Jesse judged neither's actions to be wrong. According to the agent relativist, both Clarke and Jesse are mistaken. This is because the proper frame of reference is the standards of the agents themselves, and not the people judging them.

Note that appraiser and agent relativist are not mutually exclusive. One could endorse both simultaneously. If so, then an action could be right or wrong relative to the standards of the agent and also right or wrong relative to the standards of whoever expresses a moral judgment about the actions of those agents.

Now consider the statement "An action is only morally wrong if a person believes it is morally wrong." Unfortunately, the item leaves underspecified *whose* moral action we're referring to, so it is ambiguous between agent and appraiser relativism. It could mean "my actions are only wrong if I consider them to be wrong," which may imply a form of agent relativism. Yet the notion that "an action" is only wrong if "a person" believes it could refer to a person's judgments about any moral actions, including both that person and anyone else's actions. This may seem like the more natural interpretation, but there is something decidedly worrisome about the phrase "a person." *Which* person? Is it the person performing the act, or someone else? When asked to explain what this statement means, most participants presumed it was the agent performing the action:

> *Only if the person that is committing the act thinks the act is immoral, is it actually immoral. Otherwise its moral.*

> *that would be like saying murder is only wrong if the person who is committting it thinks Your intentions and beliefs can determine the morality of your actions.*

Others seem to interpret it more in line with appraiser relativism:

> *That whether one believes an action is morally wrong or not lies in the in who is judging the action.*

*an action is wrong to one person does not mean it is wrong to another so it is wrong or not wrong depending on whose perspective is being considered*

Although a handful of examples are hardly evidence of a robust pattern, these responses provide compelling evidence that *some* people interpret the statement to relativize moral claims to the standards of the agent, while at least some other people interpret it as relativizing moral standards to those judging the actions. Indeed, the first of the two examples of seemingly appraiser-relativist interpretations is an almost textbook expression of appraiser relativism. This is clear evidence of the potential of interpretative variation, and illustrates, if nothing else, that these concepts are distinct, that participants are capable of recognizing these distinctions when prompted to do so, and that they do not consistently interpret items on the MRS in a way that reliably conforms to one or another interpretation. And since agent and appraiser relativism are conceptually distinct, and this item fails to disambiguate them, judgments about this item could in principle express disagreement with one of the concepts but not the other, or agreement with either, or agreement with one but not the other.

Another way that participants could interpret this item in an unintended way is if they interpret the question to express a normative stance about the moral relevance an agent's knowledge has on their culpability. Suppose Alex gives Sam a glass of water. However, the glass of water was poisoned, and Alex had no reasonable way of knowing that it had been poisoned. Sam drinks the glass of water, and dies. Did Alex do something morally wrong? Many of us would think that Alex did nothing wrong because Alex did not intend to kill Sam, and Alex was not negligent in giving Sam the glass of water. Note that the item states, "An action is only morally wrong if a person believes it is morally wrong." Critically, Alex did not believe that giving Sam a glass of water was morally wrong. Suppose Alex *did* know that the glass of water was poisoned. If it is morally wrong to poison Sam, and Alex knows this, then Alex would have done something wrong in giving the poisoned glass of water to Sam. Critically, the moral status of Alex's actions depends on Alex's beliefs.

Of course it is possible that Alex knows that the glass of water is poisoned, but does not believe it is morally wrong to poison people. If so, we may still just think that Alex's actions are wrong, even if Alex does not believe that they are. However, this simply illustrates that the relationship between one's beliefs and the moral status of one's actions is complex. Beliefs about certain nonmoral facts may be sufficient to determine whether one's actions are morally right or wrong, rather than the moral beliefs themselves. Yet this item is ambiguous. *Why* someone does not believe their actions are morally wrong is typically relevant. Whether a person's actions are wrong may depend on certain *nonmoral* beliefs, but they may not depend on that person's personal moral standards and values. Unfortunately, this item simply describes a person who does not believe an action is morally wrong. It does not say *why* they think it is morally wrong. There is little indication that many participants interpreted the item in this way, but again, I have only sparse data ($n = 16$) on how people interpreted individual items on the MRS and why they answered the way that they did ($n = 16$). However, one response is suggestive of the potential inclusion of this interpretation among the range of possible interpretations that deviate from researcher intent. When asked what this item means, one participant stated:

> *If someone does something that he knows is morally wrong before he does it, it makes him morally wrong.*

Note that the participant suggests that if the person *knows* an action is wrong, that they do something wrong when they perform the action. The participant seems to take for granted that there is some standard of moral wrongness, and that whether a person's action is wrong or not depends on whether they know what that standard is. On the one hand, this could imply a realist standard, since this suggests that a person could have knowledge of what is morally right or wrong. Note, however, that moral knowledge is consistent with various forms of antirealism, including relativism, subjectivism, and constructivism (and perhaps even quasi-realism), and would be potentially inconsistent with noncognitivism and error theory, which may be the least plausible candidates for folk antirealist views.

On the other hand, this interpretation could be consistent with precisely the kind of unintended interpretation I am suggesting, since they could be expressing a *normative* stance rather than a *metaethical* one. The metaethical interpretation of this item is, roughly, that an action is right or wrong relative to the standards of the person performing that act, so if a person believes the act is wrong, then the moral status of their actions is judged relative to this belief. The beliefs of the agent performing the act serve as the standard of evaluation by which the action is to be judged. But the normative interpretation of this item would hold that evaluating whether an action is right or wrong involves (at least in part) consideration of the mental states of the agent performing the action, e.g., their knowledge, beliefs, intent, etc. We may believe, for instance, that an action is only morally wrong if the person performing that action is morally culpable. And if a person lacks knowledge of morally relevant nonmoral facts, that person may fail to believe that an action is morally wrong when knowledge of those nonmoral facts would cause them to believe the action is morally wrong. If so, the notion that an action is only morally wrong if the person performing the action believes it is morally wrong could simply convey that an action is only morally bad when the person performing that action is morally culpable, and insofar as moral culpability requires them to have moral knowledge, and moral knowledge entails moral belief, then a person could not actually commit a morally wrong action unless they knew they were doing so. Such an interpretation would reflect a stance about moral epistemology and normative ethics, but it would not entail realism.

| **MRS #9** *relativism* | There are moral rules that apply to everyone regardless of personal beliefs. (R) |
| --- | --- |

The most obvious shortcoming with this item is that it could be interpreted as a descriptive claim. As a matter of descriptive fact, we apply moral rules to people regardless of whether they agree with those rules. We lock up thieves and murderers, even if they think their actions were justified. And we

condemn liars and cheaters, even if they think lying and cheating is okay. This item could be straightforwardly interpreted as the recognition that we do hold others responsible for their actions and punish them. When asked to explain what this item means, some participants appeared to interpret it this way, e.g.:

> *You can believe that you have a right to steal, but the majority of people will believe that stealing is something that is morally wrong, no matter what you choose to do.*

And when asked to explain why they agreed or disagreed with this statement, some participants likewise offered a descriptive rationale:

> *The moral belief that humans should not murder seems pretty universal. Stealing tends to be seen as morally wrong regardless of demographic as well.*

> *Some things like killing someone are morally wrong in every culture. There are some things that no matter where you are in the world are not acceptable, like violence.*[121]

People could also interpret this as the descriptive claim that there are moral norms that everyone *knows* or *believes* or *is aware of.* This may not be obvious given the way the question is worded, but several participants interpreted it this when asked to explain what it means, or to explain why they answered the question the way that they did:

> *Even someone that kills someone else knows deep down it is wrong.*

> *It means that there is some things that everyone knows is immoral if done.*

---

[121] Note that the latter is extremely hard to interpret: what do they mean when they say that some things "are morally wrong" in every culture? That it really is wrong in that culture, or that it is *regarded* as wrong according to that culture's moral standards? The latter remark, that certain things "are not acceptable," may likewise convey the *participant's* judgment that the action is unacceptable in that culture, but interpreting this as a metaethical expression is strained: does it convey the participant's own judgment? That is, is it an expression of appraiser relativism? If so, the phrasing used here is a rather clumsy way to convey this. If someone were to say, "Some things like chocolate are good in every culture," does this mean that the participant considers chocolate good? This is implausible, since a person is most likely expressing a generalization about those cultures, i.e., something along the lines of "people *in general* find chocolate good in all cultures." Such claims do not require thinking that literally *every* person in all cultures likes chocolate. The person expressing a generalization about what people believe is always capable of not endorsing the belief conveyed in the generalization. This response is also consistent with an expression of universalism about moral standards, which is orthogonal to the realist/antirealist distinction, or a descriptive claim that certain actions are right or wrong according to every culture's standards, which does not necessarily express relativism, since recognizing that cultures share similar moral standards does not indicate that those standards are *correct*.

*because we all have them ingrained of us we just know when soemthing is wrong*

*It means that people should know from right to wrong no matter what their beliefs are.*

*Everyone shares a few ideas in common concerning morals.*

In fact, such responses were common. Unfortunately such responses are consistent with antirealism or indeterminacy, even if they allude to or are consistent with realism. More importantly, they indicate that these participants did not interpret the question as intended in a way that inadvertently seems slightly more consistent with realism than antirealism (e.g. the notion that someone "knows deep down" that something is wrong) is a poor foundation to mount a defense of the validity of these scale items. One would still have to show why this scale item is valid in spite of considerable interpretive variation, and, at best, what researchers will have achieved is "accidental validity" a person's response reflecting realism *in spite of* their interpretation of the item, and not because of it. Strangely, in spite of its wording, some even appeared to interpret the item in normative terms:

> *The are underlying guidelines that should be followed by humanity that should not be violated based on the beliefs of different groups.*

> *I agree that their are morals that people should respect no matter what their beliefs are.*

Nothing about this item provides any obvious indication that it is intended to convey a normative claim about what we *should* do. Nevertheless, it is possible that no matter how carefully you word an item about morality, that the normative nature of morality is so overwhelming and intrinsic to moral consideration that people will use discussions of morality as an opportunity to convey or express their moral standards, and may interpret non-normative questions about morality in normative terms. If so, this constitutes additional, albeit potentially minor noise researchers should be aware of.

This item could also be readily interpreted as an expression of universalism, rather than realism. Indeed, all reverse-coded items on the scale contrast relativism with universalism. Yet the universalism-relativism distinction concerns the scope of moral concerns, not what makes them true.

Thus, this item is consistent with the MRS as a whole failing to distinguish between realism and antirealism, since its operationalization of relativism contrasts relativism via scope rather than stance dependence versus independence.

Several terms in this item may also be hard for ordinary people to interpret as intended. What exactly is a moral *rule*? Do moral rules differ from moral *principles* or *actions*? Ordinary people may regard a *rule* as an especially rigid or inflexible application of moral principles, i.e., a moral "rule" could be conflated with a moral norm that does not allow for exceptions, or that must be strictly adhered to without regard for exculpatory considerations.

It is also unclear *whose* personal beliefs this item is referring to. *Whose* personal beliefs? The beliefs of the people to whom these rules apply? The people applying these rules? Presumably, researchers intend on the former interpretation, but the latter makes sense, as well. Consider cases where a person must enforce a rule or enact a punishment even if they have misgivings: people may be hesitant to punish a friend, family member, or someone who they pity or sympathize with. Instances of people overcoming their personal beliefs to enforce a moral rule abound in popular consciousness and in parables, fables, and romanticized historical narratives. One of the first consuls of the Roman Republic, Lucius Junius Brutus, garnered the respect of his colleagues and contributed to his mythic stature in part due to his willingness to oversee the execution of his own sons, when their involvement in a conspiracy to reinstate Tarquinius as king (Dionysius of Halicarnassus, 1940).[122] Judges are expected to recuse themselves when they have a conflict of interest, and we sympathize with but respect authority figures whose role mandates that they punish friends and family with the impartiality demanded of their post. In short, the notion that we must apply moral rules to all people equally

---

[122] This reference may be obscure, but this could be due in part to a fading interest in the classics for those pursuing higher education. A reference to Lucius Junius Brutus appears in Shakespeare (1599/2020, 1.2.247-250), and the events surrounding the execution of his sons are depicted in Jacques-Louis David's *The Lictors Bring to Brutus the Bodies of his Sons*, which was used along with other Roman senatorial depictions as propaganda during the French Revolution (de Vela & Earley, 2015).

regardless of *our* moral beliefs is not an unfamiliar notion. Yet this is not what this item is asking about. Rather, it is asking whether there are moral rules that apply to people regardless of whether those rules are inconsistent with that person's personal moral standards. This is an altogether different notion, yet the wording of this item is consistent with both interpretations due to the failure to disambiguate whose personal beliefs are irrelevant to the applicability of moral rules.

Finally, the very notion of a moral rule *applying* to people is obscure, even among philosophers. Just what does that mean, exactly? I have studied this topic with an intensity bordering on madness for years, and I am still puzzled about just what it means for a moral rule to "apply" to someone. At the very least, there are different ways in which a rule could apply. For instance, if I am participating in a tournament, a rule may apply to me in virtue of some intersubjective set of shared, formalized rules I have contractually agreed to abide by. These rules apply to me in a practical way: if I violate the rules, and the relevant authorities discover these violations, I will be subject to whatever penalties are prescribed by those rules. A rule could also "apply" in the more abstract and diffuse sense that people in one's culture or community will judge you in accordance with that rule. But it is altogether unclear in what sense a moral rule applies to us in the respect moral realists believe we are subject to such norms. Philosophers will variously suggest that such applicability entails that we are rationally bound by moral rules, or that such rules have oomph, or practical clout, or provide us with decisive reasons, or they state that we cannot opt out of these rules, or that they are inescapable, or have some kind of practical authority over us. The list of ways in which stance-independent moral rules purportedly apply to us is long and obscure, with many of the relevant terms merely serving as philosophical argot that relabels the very mysteries they are purported to describe, without doing any descriptive work. Metaphors with fancy hats, as it were. With respect to the way in which moral rules "apply" to us on realist accounts, the way in which they apply may differ from one account to another, yet in all cases this term remains obscure. At best, it could be cashed out technical terms in a variety of ways. At

worst, its use is a mere promissory note for some notion or concept that has yet to be adequately articulated. In short, it's simply unclear what it would mean for a moral rule to apply to someone, especially on various accounts of realism where the way in which a rule is applicable cannot be reduced to mundane descriptive facts.

Unfortunately, researchers importing terms and phrases from academic philosophy often fail to appreciate that such terms and phrases are employed in idiosyncratic ways that may seem like they must have some fairly uncontroversial meaning among philosophers, but are in some cases obscure or highly underdeveloped. Philosophers who hold that moral rules "apply" to us *may* be able to stipulate or cash out what they mean clearly, but then again, perhaps they may not be able to do so. When pressed, philosophers will on occasion fall back on claiming that such concepts are self-evident, or unanalyzable, or not in need of explication, but that this isn't a problem because we all "have" the relevant concept. Perhaps they do. But do ordinary people have the same concept these philosophers do? How do *they* interpret the notion of moral rules "applying" to people? This could be understood as the normative claim that we should hold everyone to the same standards, or the descriptive claim that we do hold everyone to some moral standards, yet neither of these would capture the kind of applicability realists (especially non-naturalists) believe moral facts have. And if this item is intended to express universalism rather than realism, then it would be consistent with a range of potential metaethical stances about the way in which the moral rules apply, since its only contention is that— however they apply—they apply to everyone.

In short, many of the terms used in this item are only superficially appropriate. Everyday terms, like "apply" and "rule," may pass an initial sniff test by whatever experts judged how well this item reflects relativism (or, in this case, its negation). Yet such experts may fail to appreciate that these terms have highly refined and obscure uses in philosophy, and that such uses may fail to reflect the ways ordinary people are disposed to interpret such terms. And the potential for unintended

interpretations is only compounded by the inclusion of such phrases together, in a sentence. After all, people don't interpret each word as an isolated semantic unit; words are interpreted in the context of the surrounding sentences and contexts in which they appear. Perhaps the collective effect of presenting these terms with a narrow philosophical interpretation in mind together is sufficient to prompt the intended interpretation in the sentences in which they appear. Or perhaps there's a cumulative effect to so many obscure terms that makes interpretation of the sentences in which they appear even more difficult. It is difficult to emphasize just how strange it is to presume that participants would interpret items like these as researchers intend, not just some of the time, but enough of the time for these items to serve as appropriate measures of a distinct metaethical position. Students often struggle to understand relativism after an extended lecture on the topic, yet we're expecting them to pick up on the relevant concept with incredibly minimal prompting, without any explanation, background, or context. This may be appropriate for emotions or personality traits, but it's not clear that it is appropriate for a technical concept in an academic discipline.

| **MRS #10** *relativism* | The same moral standards should be followed by people from all cultures. (R) |
|---|---|

This is not a valid measure of relativism. Whether or not people *should* follow the same moral standards is a normative question. Given that it is clearly not a valid measure, there is no charitable way to put it: the experts who judged this to be an appropriate measure of relativism are simply mistaken. It isn't. I'm not a moral relativist, but I am a moral relativist. Yet I do think everyone should follow the same moral standards: *my* standards. What else would I think? Of course I'd prefer other people to share my moral standards. I'm against murder and stealing. I think others should be against murder and stealing, too. Yet this stance has nothing to do with thinking that there are universal moral rules that apply to everyone, or more specifically that they do so because these rules are stance-independent

moral facts. Agreement with this item is straightforwardly compatible with antirealism, and even relativism.

For instance, an appraiser relativist would judge that actions are right or wrong relative to each individual's moral standards. There is no inconsistency in holding this view and holding that everyone should adhere to your standards. This item could even convey a nonmoral practical desire, or a moral attitude that is independent of and consistent with agent relativism. For instance, suppose you're an agent relativist. As a result, you believe everyone is bound by their own moral standard, and should act in accordance with them. Is this inconsistent with believing it'd be better, or that for practical reasons, people *should* have different moral standards than whatever standards they have? It doesn't seem like this to me.[123] In short, believing that everyone *should* adhere to the same moral standards doesn't mean that they *are in fact subject to those standards*. Wishing something is so isn't the same thing as it being so.

Just as agreeing with this item is consistent with both relativism, various forms of antirealism, and even the denial of universalism (which is presumably what agreeing with this was intended to express), disagreeing with this item likewise does not convey endorsement of moral relativism. To disagree with this item is to express that it is not the case that the same moral standards should be followed by people from all cultures. This doesn't require that one believe people should follow different standards, such as those of their culture. If I deny that everyone from every culture should have the same dietary practices, it does not follow that I think that people should therefore adopt whatever diet is customary for their culture. I think people should eat whatever they want, within reasonable limits (i.e., I don't think people should eat babies or priceless pieces of art). Yet

---

[123] For example, someone could believe that if a particular culture wishes to engage in harmful religious practices, such as refusing blood transfusions or cancer treatment, that it is permissible (or even morally good) for them to do so, but that they would still be better off if they didn't do so, and that there is a meaningful sense in which they should follow different practices.

disagreement with this item is presumably intended to convey endorsement of cultural relativism. If it isn't intended to convey the notion that people's standards are correct relative to their cultures, or the agent-relative notion that they ought to act in accordance with their cultural standards, then it's not clear how disagreeing with this item would serve as a measure of relativism. And if that is what it's intended to convey, then it seems like a poor way to ask. Finally, this item is subject to the *left hand* conflation.

I have reviewed all of the items on the MRS. As this extensive analysis illustrates, there are reasons to doubt the suitability of the MRS as a measure of folk realism. There are multiple reasons to doubt the validity of every single item on the scale. Some of these problems are distinct to individual items. Yet there are also problems with the scale, considered as a whole, (e.g., that it appears to contrast relativism with universalism, rather than stance-independence), and significant limitations (e.g., demand effects that favor endorsing relativism, and that the scale could at best only distinguish whether people endorse relativism or not, but cannot distinguish other metaethical stances). Notably, these problems remain despite careful attempts to design items that experts regard as appropriate measures, and after extensive efforts to design a scale in accordance with conventional scale validation procedures. I do not highlight these difficulties to suggest we replace this scale with a better one, but to suggest that such efforts may be futile. It may not be feasible to use methods typically employed in personality psychology to assess lay beliefs about philosophical positions. As Kauppinen argues, "The conceptual claims that philosophers make imply predictions about the folk's responses only under certain demanding, counterfactual conditions. Because of the nature of these conditions, the claims cannot be tested with methods of positive social science" (p. 95). Kauppinen's claim may be too broad and sweeping to endorse in every case. Or perhaps not. Either way, I believe it does apply in *this* case: scale items like those on the MRS are not a reliable indicator of folk metaethical stances or

commitments, because there is little reason to be confident that participants are interpreting these questions in a way consistent with researcher intent.

### S3.3.6 The Folk Moral Objectivism Scale (FMO)

The Folk Moral Objectivism (FMO) is a recent and welcome addition to efforts to develop more inclusive measures of folk metaethical belief (Zijlstra, 2019). Like the MRS, the FMO offers an alternative to the disagreement paradigm in the form of a scale.[124] Yet the distinct advantage of the FMO is that it does not restrict its measures to only assessing a single dimension. Although earlier research never converged on any shared terms or dichotomies, the disagreement paradigm tended to focus on objectivism versus some alternative (e.g., non-objectivism, relativism, etc.) while the MRS focuses exclusively on relativism. Whatever their merits, most of these studies tend to construe folk metaethical belief along a single continuum, roughly captured by the distinction between objectivism and relativism. One of the primary purposes of the FMO is to step away from this oversimplification by evaluating the ostensibly richer landscape of folk metaethical views.[125]

Like the FMO has a number of shortcomings, both in general and with respect to the specific items employed by the scale. Like the MRS, the FMO only asks about morality in abstract terms. As such, it is unable to detect the pluralism that emerges when participants are asked to render metaethical judgments about different moral issues. In addition, there is something a bit artificial and seemingly

---

[124] Like the MRS, the FMO evaluates the relationship between folk metaethical views and tolerance, with the added bonus of also assessing the relationship between folk metaethical belief and attitudes about punishing norm violators. However, its association with these variables is not especially important in assessing its validity.

[125] Zijlstra explicitly describes the disagreement paradigm and its exclusive focus on "perceived objectivity," and frames the FMO as a method of capturing the richer terrain of folk metaethical views. I've seen too many articles where authors will claim that a paper says something, only to discover that it doesn't. So, here's a couple of remarks from the introduction of the article:

"Existing experimental research measures folk moral objectivity on a single dimension of perceived objectivity. There are, however, good reasons to regard folk moral objectivity as multidimensional" (Zijlstra, 2019, p. 1).

"The main innovation is that the FMO-scale allows for the possibility that folk moral objectivity has several dimensions." (Zijlstra, 2019, p. 2)

incomplete about the categories on offer. The Big Five emerged out of an attempt to comprehensively catalog all the ways people speak and think about morality, then develop an appropriate set of categories to capture it all. No similar process was employed for developing the contents of the FMO. The goal of the FMO is to distinguish the various reasons why people might regard moral judgments as true, and this results in universalism, absolutism, and divine command theory. Is that it? Are these the only reasons why a person might believe moral judgments are true? I doubt it. Philosophers have certainly come up with many more ways. And if we included these in the FMO, participants may have agreed or disagreed with them, accordingly, suggesting additional constructs not captured by the scale. There doesn't seem to be any particular reason why the FMO is limited to just these three, nor is it a surprise if something resembling them falls out of a set of items that was never designed to capture additional ways people could regard moral claims as true.

Yet a more serious problem with these items is that it is unclear whether *any* actually make conceptual sense as *reasons why moral judgments are true*. First, take Zijlstra's characterization of universalism:

> According to universalism, moral judgments are true only if they are based on universally binding moral norms that apply to anyone and everywhere [...] An example of moral universalism can be found, for example, in the Universal Declaration of Human Rights. Article 1 of the declaration states that "all human beings are born free and equal in dignity and rights" and according to article 3 "everyone has the right to life, liberty and security of person" and so forth. (Zijlstra, 2019, p. 2)

This does not tell us *why* those judgments are true. It merely provides a description of one of the characteristics of moral truths: that they "apply to anyone and everywhere." In other words, universalism takes a moral norm, e.g., "stealing is wrong," then addresses the question "who does this moral rule apply to?" and furnishes the answer "everyone." This only tells us who the moral rule applies to, that is, its *scope*. It doesn't tell me *why it's true*. It could be true because God grounds objective moral facts. It could be true in virtue of some constructivist account whereby we agree on some set

of standards and hold one another mutually accountable (Darwall, 2006), or it could be true relative to the standards of an individual or a culture, or with respect to some stance-independent body of moral facts. Simply put, universalism has nothing to do with why a moral claim is true. It's merely a feature of the normative content of the rule itself, and the truth of such claims is consistent with many antirealist positions. It's not even clear universalism must necessarily be cognitivistic.

In fact, universalism is compatible with noncognitivism, so it doesn't even necessarily require that moral claims be *true*! According to universalism prescriptivism, moral judgments express an imperative that commits whoever expresses those judgments to expressing the same judgment in all similar situations. For instance, if I judge that it would be wrong for Alex to steal from Sam, I am committed to it being wrong for anyone to steal from anyone else in the same situation, and this is cashed out in terms of an imperative: "don't steal under these circumstances." Such an account is noncognitive, since moral claims ultimately express an imperative rather than a propositional claim.[126] That universalism could be appended to a noncognitivist account indicates that universalism doesn't entail any particular stance on what makes a moral claim true, since it is compatible with accounts that deny that moral claims are propositional (and thus capable of being true) in the first place. Universalism concerns the scope of moral claims, not what makes them true. As such, universalism may not be an appropriate construct for inclusion in the set of constructions associated with why moral claims are true.

Another problem with treating universalism as an account of what it would mean for moral claims to be true is that merely thinking that there is some reason a moral claim is true isn't sufficient to make an account "objective" (that is, realist). This is because objectivism is inconsistent with any

---

[126] That Zijlstra appears to overlook the possibility of universalism appearing in antirealist metaethical positions is all the more surprising given that he cites Hare in the quote above.

form of stance-dependent cognitivist account, such as standard forms of relativism. Consider Zijlstra's examples of universal moral claims from the Universal Declaration of Human Rights:

> *"all human beings are born free and equal in dignity and rights"*

> *"everyone has the right to life, liberty and security of person"*

A cultural relativist or an individual subjectivist could endorse these claims, and in doing so claim that they are true in virtue of expressing claims consistent with their standards or the standards of their culture. Such claims would be true, but they would not be objective. This is a problem for Zijlstra's distinction between forms of perceived objectivity and the "no moral truth" category, since relativism is included within the latter. Unfortunately, universalism is consistent with relativism, so the distinction may be misconceived.

Zijlstra's (2019) characterization of absolutism is also a bit troubling. He states that "Moral absolutism goes beyond universalism in that it also holds that true moral judgments are derived from more basic moral truths. The underlying idea is that the core of morality is determined by a set of general rules and principles which all hold true, without exception" (p. 2). Yet this seems to pass the buck on why it is that a given moral judgment is true to some higher-order normative principle. Even if we think that claims about what is morally right or wrong in any particular situation will turn on whether a particular course of action is consistent with some general moral principle, this does not tell us why the general moral principle is true. For instance, if it is true that I should not lie to my boss by calling in sick because of a more general moral rule against lying, e.g. "it is morally wrong to lie for personal gain," this does not tell me why *this* is true. And even if the general moral principle is true, its role in serving to make any specific moral judgment true or false is a type of proximal truth relation that holds in virtue of *normative* moral considerations, *not* metaethical ones. And such normative considerations are consistent with antirealism. A relativist might hold that, relative to their standards or the standards of their culture, a given action is *always wrong without exception*. Absolutism, strictly

speaking, is a *normative* position, not a metaethical one, insofar as it relates to the application conditions of a given moral rule or principle. The same holds true for general moral principles: such principles are normative, not metaethical. To illustrate why this is a problem, see **Figure S3.1**.

**Figure 3.1**

*Levels of moral specificity*



As this figure illustrates, we may judge specific moral instances by appealing to general principles. For instance, it may be true that Alex should not lie to Sam because everyone has a duty to not lie. Yet a moral duty to not lie is a normative moral principle, not a metaethical one, and it may be true for reasons consistent with antirealism (e.g. it may be true because it is consistent with Alex's moral standards or the standards of Alex's culture, or perhaps it is true in virtue of some constructivist account). The problem is that Zijlstra presents absolutism as one of the positions people could take on why moral claims are true. Yet to the extent that a specific moral action is right or wrong is true in virtue of its conformity with a general moral rule, we're still left without an account of what makes the general moral true; without such an account, the moral rule could be made true by stance-

independent moral facts, in which case it would be a form of moral realism and could therefore serve as a form of "perceived objectivity," or it may not, e.g., the general moral rule could be true or false relative to the standards of a person or group, in which case the general moral rule wouldn't be *objectively* true. Simply put, the truth relation between general moral principles (absolute or otherwise), has *nothing to do with perceived objectivity*.

Zijlstra cites Kant as someone whose views entailed that there were certain moral absolutes. First, it's not even clear Kant's views are properly characterized as realist, as one dominant line of thought has characterized his position as a form of constructivism (Formosa, 2013; cf. Bojanowski, 2012). This may entail that, on some construals of "objectivism," that there are objective moral facts, but such facts would not necessarily entail the kinds of stance-independent moral facts conventional realists endorse. If so, Zijlstra's example would itself be a contestable instance of a realist conception of absolutism. Setting aside whether Kant is a realist, we can simply grant that even if he was, it is not at all clear that Kant's views are realist *in virtue of his moral position entailing that there are moral absolutes*. It may be that, e.g., Kant's conception or at least some Kantian conception of moral duty would hold that moral facts are a product of synthetic *a priori* judgments (Hanna, 2017; Potter, 1997; Schwartz, 2017). If so, it could be that if someone believes that there are absolute moral principles, that they conceive of them or speak about them in a way that commits them to some kind of moral rationalism. However, if so, what makes these moral judgments true wouldn't be that they are absolute, it would be in virtue of these facts following from certain necessary principles (Schwartz, 2017). In other words, to the extent that one could characterize Kantian absolutism as realist, it isn't merely in virtue of the moral judgments being absolute, but in virtue of *why* they are absolute. After all, an antirealist can endorse exceptionless moral rules, e.g., a cultural relativist could observe that, "relative to this culture's moral standards, it is never morally permissible to commit any form of violence." This culture would have an absolute moral rule, but it would be true relative to that culture, not stance-independently

true. In short: *absolutism* has nothing to do with realism, so it cannot serve as a legitimate reason why people could think moral judgments are stance-independently true. Absolutism isn't even a metaethical concept (it's a *normative* concept) and is conceptually orthogonal to the distinction between realism and antirealism. Unfortunately, this means that absolutism cannot serve as an appropriate dimension of perceived *objectivity* on the FMO.

Another problem with Zijlstra's conception of absolutism is that it runs two concepts together: absolutism and generalism. Absolutism is the view that there are no exceptions to a given moral rule. Absolutism could be restricted to one or a handful of moral rules, or one could be an absolutist about all moral rules. Absolutism does not strictly require a stance that *all* moral rules are absolute. However, generalism is the view that there are general moral principles that can be applied to specific cases. It is typically contrasted with particularism, which denies that there are any general moral principles. Nothing about believing in general moral principles requires that those principles be absolute. One could endorse the moral principle that "you should not lie," but include caveats, e.g., "unless doing so would conflict with other moral duties that take priority." Absolutism is a feature that some moral principles may have, but it is not a necessary one. This will turn out to be a problem for the items that appear on the absolutism subscale, because demonstrating that people endorse the existence of moral principles does not entail that those principles are absolute. What Zijlstra proposes seems to be a conjunction of absolutism and generalism, which is then subsumed by the label "absolutism" and treated as a single psychological construct. But there is no good reason to presume that absolutism and generalism would be psychologically conjoined this way in ordinary moral thought.

Divine Command Theory (DCT) might be the most contentious of the three. While it is often construed as a form of moral realism, I'm not convinced that it is, or at least that it *must* be. Zijlstra characterizes DCT as the view that "whether an action is morally right or wrong depends on the commands of a divine being [...] In other words, true moral judgments are based on divine commands"

(p. 2). Yet this does not tell us whether these judgments are stance-dependent or stance-independent. At first glance, this would appear to be a stance-dependent account, in that moral claims are dependent on God's stances. In that respect, it would appear to be a form of antirealism. However, there may be ways of building on this simple description that render DCT a realist account. I am happy to simply grant that there are, and that DCT is consistent with or a form of moral realism.[127] However, even if we grant that DCT is one of the conceptual foundations for why someone might think a moral claim is true, there is something deficient about presenting universalism, absolutism, and DCT as the primary (or only) ways one could think moral claims were true. First, none of these necessarily entail that moral standards are *stance-independently* true. As such, Zijlstra has not properly distinguished his conception of "perceived objectivity" from cognitivism. Second, these reasons for thinking moral claims are true are not exhaustive of the reasons why someone could think moral claims are true, nor even that they are stance-independently true. For instance, one could endorse moral realism, and believe that there are moral facts that are not universal, absolute, or predicated on God's will. Such a person would believe that there are moral truths, but their views would not be captured by the categories presented by Zijlstra. It is not difficult to imagine someone who believes there are stance-independent moral facts, that these facts don't depend on God, and that they are not absolute, in that they admit of exceptions. In fact, this is probably the norm among the majority of philosophers that endorse moral realism.[128]Universalism may be the least likely of the three characteristics to drop from a realist account, but there is nothing logically prohibitive in doing so. Indeed, Wong (2006) defends an account that explicitly cuts a middle path between an "everything goes" relativism and a universalist

---

[127] Though it seems logically possible to have an antirealist conception of DCT as well, or to be a divine prescriptivist: a noncognitivist who believes moral facts are divine imperatives that are neither true nor false.

[128] Most philosophers are atheists (66.9%) and moral realists (61.9%; Bourget & Chalmers, 2014; ms). While there is no direct data on whether they believe there are exceptionless moral rules, Bourget and Chalmers (ms) found that 33.7% endorsed particularism, suggesting that about a third of philosophers reject the notion of general moral principles, while only 54.6% favor generalism, a slim majority at best.

conception of morality, *pluralistic relativism.* According to this view, there is no single correct moral system, but a variety of moral systems all of which can be true, but the truth of these moral systems is still constrained by an overarching need to consist of facts about how people effectively live together. Such an account does allow one to say that some moral standards can be incorrect, yet it simultaneously holds that moral facts are not strictly universal. Such circumscribed pluralism may not yield the clean and start contours of a simpler account that, but there is no justification for ruling out *a priori* the possibility that it could be more in line with how ordinary people are disposed to think about moral facts.[129]In short, the categories employed by the FMO suffer significant conceptual shortcomings, and do not appear to present a comprehensive, or even adequate range of dimensions of folk metaethics.

Unfortunately, Zijlstra never explicitly articulates what "perceived objectivity" means, so it is hard to know whether it is intended to reflect moral realism, though the term *perceived objectivity* was introduced by Goodwin and Darley (2008), who describe what they mean in a way that appear to refer to moral realism as it is defined here.[130] Zijlstra states that perceived objectivity is "often probed by two different questions, namely a truth-aptness task and a disagreement task" (p. 1). This is true as far as it goes, but as critics have already pointed out, whether moral claims are truth-apt or not is an indicator of *cognitivism*, not objectivity. While one could in principle maintain that all forms of cognitivism barring error theory are a form of objectivism/realism, this wouldn't be consistent with Zijlstra's own categories, since perceived objectivity is contrasted with relativism, and relativism is

---

[129] Zijlstra's account also seems not to consider naturalism: the view that moral facts are natural facts of some kind. On such a few, moral facts would be true in virtue of certain natural facts, yet this is not captured by the three categories on offer.

[130] The term "perceived objectivity" only appears in the abstract (Goodwin & Darley, 2008, p. 1339). Goodwin and Darley clearly use "objective" in a way consistent with my use, i.e., to refer to stance-independence. As they put it, they are concerned with whether moral beliefs or standards "derive their truth (or warrant) independently of human minds (i.e., objectively), or whether instead, their truth is entirely mind-dependent or subjective" (p. 1341). They use "objective" synonymously with "independently of human minds," which is a close approximation of my use of "stance independent," though I would not restrict stance independence to *human* minds in particular. After all, morality wouldn't be objective if moral claims were true or false relative to the beliefs or standards of aliens or fantasy beings, like elves or goblins.

ultimately subsumed by the "no moral truth category." Indeed, Zijlstra explicitly states that "someone who does not regard morality as objective might regard moral judgments as true relative to a culture" (p. 1.) Yet relativism and error theory are both cognitivist positions.[131] As such, the truth-aptness task is not an appropriate method for measuring perceived objectivity. Since the only other measure on offer is the disagreement paradigm, which suffers serious methodological issues as well, the FMO seems to rely on a flawed operationalization of perceived objectivity.

This reliance on the truth-aptness task and the disagreement paradigm carries over into the studies used to support the validity of the FMO. For instance, the FMO purportedly demonstrates convergent validity with other measures of perceived objectivity in that responses to the FMO correlate with responses to Sarkissian et al. (2011), specifically the "other culture" condition (i.e. the one describing a fictional Amazonian tribe, the Mamilons). One puzzling feature of these findings is that DCT was *negatively* correlated with the "objectivist" response to the disagreement paradigm, i.e., they were more inclined towards a relativist response. Zijlstra speculates on why this may have occurred:

> It might be that people have different views on whether or not the commands of a divine entity apply to other cultures. If that is the case, people who score high on Divine Truth may respond as if morality is relative because they believe that the divine commands issued by God apply to their own culture and not necessarily to members of different cultures. Indeed, God may even have different commands for members of different cultures. Alternatively, it is possible that people recognize that other cultures have different gods and that those gods may issue different commands. As a result, moral truth is relative to those different cultures. (p. 6)

Unless the majority of participants endorsing DCT endorsed some form of polytheism or henotheism, the latter is not an especially plausible explanation. If we had asked ancient Romans or Greeks, perhaps

---

[131] Strangely, Zijlstra cites Harman (2015), whose article is titled "Moral relativism is moral realism." Harman, unsurprisingly, argues that relativism is a form of moral realism. Given that "perceived objectivity" is a term that originated with Goodwin and Darley, whose conception of objectivism is explicitly identifiable with how I construe realism (i.e., that there are stance-independent moral facts), this is a puzzling decision: why would you cite an article that relativism is a form of realism, where realism is more or less interchangeable with "objectivism," when explicitly contrasting relativism with objectivism?

they would acknowledge that members of other societies would be subject to local deities, but it is far more likely that the bulk of respondents were monotheists, e.g., Christians. It seems implausible that Christians would be inclined to think their moral standards only applied to members of their own culture, nor is there any good reason to think Christianity would entail that different cultures are subject to different moral standards. A more plausible alternative is that Christians may judge that much of the substantive content of their moral standards comes from Biblical and other religious sources that Mamilons don't possess, and that they are therefore less culpable due to their ignorance of the full scope of God's moral commands (even if, as many Christians believe, the moral law is "written on our hearts," *King James Version*, 2022, Romans 2:15). Whatever the case, this result appears to be an inconsistency that is difficult to explain were all subscales of the FMO valid. Zijlstra also suggests that proponents of DCT may "believe in a very personal form of free will and moral responsibility - that is, it is ultimately God who will judge who was right and who was wrong" (p. 6). Again, this seems like a stretch. While respondents may endorse free will, free will on a Christian view is more naturally construed as the capacity to freely choose to do what is morally right or wrong; that God is the ultimate judge of what is right or wrong in no way conflicts with this.

Study 4 also offers evidence of the convergent validity of the FMO. However, responses to the FMO are assessed alongside the first study employed by Goodwin and Darley (2008). Unfortunately, this is the study that used a composite measure of one's response to the truth-aptness task and the disagreement paradigm. Since the former is a straightforwardly invalid measure of realism/antirealism, and the latter represents one of the most methodologically compromised versions of the disagreement paradigm, the prospects of the overall measure serving as a valid measure of perceived objectivity are very low.[132] It may serve as *some* evidence of the validity of the FMO that it

---

[132] These are the specific measures used: (1) Truth-aptness: "participants were asked whether there was a correct answer to whether the moral claim was true (1: no correct answer, 6: definitely a correct answer)" (Zijlstra, 2019, p. 7).

predicts responses to other metaethics paradigms; this suggests that there is some overlap in what these measures are capturing, at the very least. Unfortunately, if there are reasons to worry about the validity of both studies, and those worries are not entirely independent, in that the invalidity associated with both paradigms results in similar response patterns, a correlation between both invalid measures may be spurious. There may be other reasons to be concerned about the findings reported by Zijlstra. However, my primary concern is the items themselves. Like the MRS, analysis of the individual items of the FMO reveals significant problems with face validity and reveal ways that participants could interpret items that are inconsistent with researcher intent. If we cannot be confident that individual items serve as valid measures of their respective metaethical dimensions, then broader considerations of the association between results on the FMO and other measures may be moot.

| **FMO #1** *No truth* | Other than what people believe, are brought up to believe, or want to believe about it, there are no facts about what is morally right and wrong |
| --- | --- |

This is supposed to be part of the "no truth" category. Yet this item begins with "other than…there are no facts." This would seem to imply that there *are* moral facts of a certain kind, i.e., facts about what people believe, are brought up to believe, and want to believe. Granted, these may not be the kinds of facts relevant to a moral realist, but the way the question is worded would seem to imply that they are a subset of moral truths that must be distinguished from the rest. Strictly speaking then, the wording of the question is loaded: agreement with the question requires one to agree that there are moral facts about what people believe and so on, while disagreement implies that there are such facts

---

(2) Disagreement paradigm: "participants were asked how they would interpret a moral disagreement with regard to the moral claim (1: Neither of us needs to be mistaken, 6: The other person is clearly mistaken)" (p. 7).
(1) measures cognitivism. (2) is a version of the disagreement paradigm that is a uniquely poor choice given the inclusion of epistemic language in the response options (e.g. *clearly* mistaken).

as well as some other types of facts. This is not a good way to frame a question. For comparison, suppose you were asked the following question:

> *Other than spiders, snakes, and scorpions, most small animals are not scary.*
>
> [1 = Strongly disagree, 6 = Strongly agree]

One interpretation does not imply anything about whether spiders, snakes, or scorpions are scary, and instead means something like "ignoring these types of animals, consider most small animals: are most of them scary?" Yet this is not the most natural reading of this question. A more natural reading is one that pragmatically implies that spiders, snakes, and scorpions are scary, and then asks, once these are set aside, whether most of the rest are scary *as well*. In other words, the pragmatic implication is something like "some small animals are scary. Besides these ones, do you think most of the rest are?" Likewise, the question Zijlstra poses is worded in a way that carries a similar implicature: "Aside from these moral facts, are there other moral facts?" This is troubling. The statement implies that descriptive moral facts are a type of moral fact. It is unclear how implying that facts about people's beliefs and desires are a type of "moral fact" would influence how they interpret the statement. It's not so much that there is any obvious conflation or straightforward problem with this. It's simply that the question is a bit strange. One normally does not present potentially arbitrary or confusing exclusion criteria when asking a question: why not ask people if there are facts about what is morally right or wrong that are true even if people don't believe they are true? That seems better than this.

Another problem with this item is that it is unclear whether it is sufficiently conceptually distinct from relativism to represent an alternative subscale. A relativist might be inclined to agree that there aren't any other moral facts aside from what people believe. It is therefore unclear whether or not there is a plausible theoretical distinction between this item and items in the relativism subscale. Yet the relation indicated here is a strange one. To state "other than…" pragmatically implies that

facts about what people believe, want to believe, and so on, are facts about what is morally right or wrong. Since the kinds of "facts about what is morally right or wrong" that a moral realist believes in concern truth-status of substantive first-order normative moral claims, descriptive facts about what people believe are either not facts about what is morally right or wrong at all, or they are facts of a fundamentally different kind unrelated to the kinds of facts associated with realism. As a result, this item inappropriately broadens the implied scope of what sorts of things are "facts about what is morally right or wrong" to include considerations irrelevant to the construct being measured. Recall that moral realism holds that there are stance-independent moral facts.

This item does not make it clear that the moral facts being denied are true independent of people's goals, standards, or values. This is because rather than make it clear that people's beliefs could serve to make moral claims true, it treats those beliefs as moral facts themselves, then suggests that, aside from these facts, there are no other moral facts. To disagree with this item is intended to suggest a belief in stance-independent moral facts, but there are other possibilities. For instance, constructivists and ideal observer theorists might hold that moral facts are the result of certain real or hypothetical procedures, such as considering what we'd endorse if we were fully informed and ideally rational, or a process of practical deliberation. Such moral standards may not be stance-independently true, but they do represent moral facts distinct from our beliefs, what we were raised to believe, and what we want to believe.

| **FMO #2** *No truth* | All ideas about what is morally right and morally wrong are products of individuals, cultures, and communities and nothing more |
|---|---|

Agreement with this item is intended to reflect rejection of stance-independent moral facts. Yet agreement or disagreement with this item may not tell us whether people believe there are such facts.

This is because the item focuses on whether *ideas* about what is morally right or wrong are constructed (rather than discovered). This is consistent with both realism and antirealism. This is because the item does not ask about the putative moral facts or what makes them true. Such facts would either exist or not exist regardless of people's ideas about those facts. Thus, a moral realist could believe that people's *ideas* are the product of people, even if the moral facts themselves are not. For comparison, suppose you were asked:

> *All ideas about science are the products of individuals, scientific institutions, and communities and nothing more.*

This seems true enough. All *ideas* about science *are* the product of people, institutions, and communities. But that is not directly related to whether scientific *facts* just are or are reducible to the ideas of individuals, institutions, or communities. Scientific *facts* are conceptually distinct from scientific *ideas*, which may or may not successfully refer to the facts. Note that even saying that moral facts are the "product" of people and groups wouldn't escape this problem, since to say that these facts are the products of people and groups could be understood to reflect an epistemic relation rather than a causal one. That is, to the extent that moral facts are "the products" of people and groups, this could simply mean that people discover the facts, rather than literally make them true in virtue of their mental states and activities, e.g., their beliefs, attitudes, or desires.

So far, this indicates that there are two ways in which this item is poorly phrased: the use of "ideas," and the use of "products." But there is also reason to worry about the use of "nothing more." Nothing more than what? This could mean the ideas aren't the product of anything other than these things. Or it could mean that, with respect to morality, people have various ideas about what is morally right or wrong, but there is nothing beyond these ideas. That is, there are ideas, but no additional facts about what is stance-independently right or wrong apart from those ideas. The former would have nothing to do with moral realism, so presumably the item is intended to reflect the latter interpretation.

Yet it does a poor job of this. It's just not clear what this nothing more could be referring to, so there is little reason to be confident that if one agreed with this, that they agreed that there were no stance-independent moral facts. At the same time, to disagree with this would suggest that there was something more than just people's ideas about morality. This is even worse, because nothing about the item specifies what there might be in addition to these ideas being the product of individuals, cultures, and communities. This *could* mean that in addition to these, that these ideas refer to stance-independent moral facts. Or it could mean any number of other things. For comparison, imagine asking people to express how much they agree with this statement:

> *All ideas about what food is good or bad are the product of individuals, cultures, and communities and nothing more*

I disagree with this statement, but I don't believe there are objective gastronomic facts. Rather, I disagree because I think ideas about what food is good or bad are not merely the product of individuals, cultures, and communities. There are species-typical evolutionary and physiological factors relevant to human food preferences. The fact that many of us enjoy eating bacon, French fries, and chocolate cake, but we do not enjoy eating sand, garbage, and buckets of rusty nails, isn't merely the product of our subjective preferences or our cultural background. While these may provide proximal explanations of our food preferences, they don't provide an ultimate explanation (Scott-Phillips, Dickins, & West, 2011).

Transposing this same concern to the original question about morality, we might likewise disagree with the notion that ideas about morally right or wrong likewise aren't merely the product of individual belief and the influence of culture and community: ideas about moral standards could be influenced by natural selection, environmental conditions (e.g., living on an island with limited resources might make coercive methods of population control more appealing), and a rational

responsiveness to practical considerations (e.g., rules against violence and stealing may serve one's self-interest). Simply put, the something more that one might have in mind when disagreeing with this item could be something other than stance-independent moral facts.

One could even disagree for epistemic reasons, because one might simply not want to commit to the idea that there is nothing more to ideas about what is morally right or wrong than what is item references, one could be unsure whether this is true without having a substantive belief about what else there might be. In short, this item appears to express a stance on the etiology of moral ideas. It is unclear whether or not, and to what extent, agreement or disagreement with this item reflects one's stance towards moral realism.

| **FMO #3** *No truth* | What people believe to be morally right and wrong are merely social conventions that could have been different |
|---|---|

Presumably, this item is intended to be interpreted as the claim that, while people may believe their moral standards are true, their beliefs are not true because they are just social conventions that could have been different (which hints that social conventions are in some relevant way arbitrary). Agreement with this item is intended to reflect the belief that there are no moral truths, while disagreement is intended to reflect the belief that there are moral truths. Unfortunately, like many other scale items, both agreement and disagreement are consistent with both belief and disbelief in moral truths.

First, one could agree with this item and still believe there are moral truths. This is because it does not follow that if beliefs are social conventions that those beliefs aren't true. It's just that their truth may be relative or reflect a socially constructed set of institutional facts. For instance, it could turn out that people's moral beliefs are true relative to their respective cultures. If so, one might believe that, in a certain respect, one's moral beliefs are "merely social conventions that could have been

different," but that they nevertheless reflect facts about what is true relative to the standards of different people or culture. Likewise, the notion that people's moral beliefs are social conventions is consistent with various forms of constructivism. For example, it's true that it's illegal to drive on the left side of the road in the United States, even if the law is an arbitrary social convention that could have been different. Social conventions can still serve as socially constructed institutional facts. It's true that the rules of chess could have been different, but that doesn't mean there aren't facts about what the rules of chess are. This is a problem for this item, because agreement with this item does not entail that the participant denies that there are moral truths. At best, it would only entail that they deny that there are stance-independent moral truths. But that's not what this category, the "no truth" category, is intended to convey. It's intended to convey the rejection of *any* moral truth, including relativistic and constructed moral truths. Otherwise, it is unclear how it could be distinct from the relativism subscale.

Responses to this question could also be orthogonal to whether one believes there are moral truths, since it could also be interpreted as a descriptive claim. Even if you think there are moral facts, you could still believe that what people *believe* to be morally right and wrong is generally the product of social conventions, and that those beliefs could have been different. This descriptive interpretation could be exacerbated by the inclusion of the notion that people's moral beliefs "could have been different." Participants may be inclined to agree that this is the case in a way that pushes them towards expressing greater agreement than in its absence, simply because it does seem true that people's moral beliefs could have been different. For comparison, suppose participants were asked to express their level of agreement with the claim:

*What gods people believe in are merely the result of social conventions that could have been different*

A religious person could believe this is generally true. They could think about the vast majority of the world's population, and recognize that most people believe what they are brought up to believe. Yet

the religious person could still believe *some of these people are correct*. This might be a bit of a stretch, since those who are correct presumably aren't *merely* believing on the basis of social convention, but something more: they could believe that some people believe for reasons, or a result of divine revelation, and so on. Yet for the intended interpretation to work, this would require "merely" to be doing an enormous and perhaps implausible amount of work. This is because the inclusion of the word *merely* would have to indirectly entail that there are no moral truths. That is, the participant would have to recognize that the item means something like:

Moral beliefs are nothing more than social conventions, and *therefore there are no moral truths*

In other words, participants would have to interpret the item to reflect the view that, *because* moral beliefs are "merely" social conventions, *there are no moral truths*. However, this requires the participant to recognize a subtle and indirect meaning that is only implied by the item, rather than stated explicitly. Part of the reason for this comes from what it would mean to disagree with this item. Would disagreeing with this item entail that you believe there *are* moral truths? No. You could think that there are no moral truths, but you could also believe that moral truths aren't merely social conventions. Since disagreement with this item does not entail that there are moral truths, it may not be clear why agreeing with this item should entail that there aren't. Instead, it might seem less like this is a statement intended to reflect a stance on whether there are moral truths, and more like a statement intended to reflect a descriptive claim about what causes people to hold their moral beliefs (i.e., their surrounding culture). If so, one could accept or reject such a claim regardless of their metaethical standards. Note, also, that the intended interpretation requires participants to interpret a subtle, indirect implication of a statement in a precise way. Given the interpretative difficulties participants have with clear and straightforward questions, this may be a tall order.

Another problem with this item is that, while it is intended to reflect the belief that there are no moral truths, it does so by requiring participants to agree with a statement that, if true, would be

inconsistent with the most common philosophical positions that deny there are moral truths: error theory and noncognitivism. An error theorist believes when people make moral claims, they are implicitly committed to false presuppositions. As a result, their moral claims are false. An error theorist would not, therefore, think that moral claims are "merely social conventions." Rather, they think that moral claims are truth-apt, and involve one or more mistaken presuppositions about what the world is like. For instance, a common form of moral error theory holds that when people make moral claims, they intend to make propositional claims about stance-independent moral facts. But since there are no stance-independent moral facts, such claims are false. Such an error theorist would not think that moral beliefs are merely social conventions; they are substantive attempts to describe what the world is like, but fail.

Noncognitivists, on the other hand, believe moral claims have no propositional content but instead express some nonpropositional claim, e.g., a descriptive claim, or an emotional state. A noncognitivist would likewise deny that moral beliefs are "merely social conventions": social conventions are institutional norms or rules that regulate behavior within a given community. Whatever a noncognitivist thinks about social conventions, they regard moral judgments as expressions of some nonpropositional state. As such, whatever they are, moral beliefs aren't *merely* social conventions, they are also expressions of nonpropositional attitudes. In fact, strictly speaking, noncognitivists might deny that people have moral "beliefs" in the first place, since a moral *belief* could be understood to reflect a truth-apt claim about what is true. Given that agreement with this item is intended to reflect the belief that there are no moral truths, but agreement is inconsistent with the two most prominent metaethical positions that deny there are no moral truths, this poses a serious challenge to this item: if we're to imagine that there are error theorists and noncognitivists among participants, these participants have no way to properly express their positions. They may both deny there are moral truths, *and* deny that moral truths are "merely social conventions."

In fact, this problem is not limited to the primary metaethical positions that deny moral truth. Both realists and antirealists could deny that moral beliefs are merely social conventions. One could believe that moral beliefs are shaped by our evolutionary history, such that we have an evolved predisposition to adopt some moral beliefs. Or you might believe that moral beliefs are the result of personal reflection, and aren't merely a matter of social convention. After all, if moral beliefs were *merely* social convention, how would we explain the existence of people who reject the moral standards of the societies they are in? Part of the problem with this item is that it, at best, *embeds* a specific set of descriptive non-metaethical assumptions about what moral beliefs are into a claim intended to convey a general metaethical position. Someone who does not agree with those descriptive claims has no way to both simultaneously agree with the metaethical implication of the statement (i.e., that there are no moral truths) *and* disagree that moral statements are "merely social conventions." This is a bit like a covert double-barreled question that effectively operates like the conjunction of the claim "there are no moral truths," and the claim that "moral truths are merely social conventions." One has no way to express different levels of agreement with each of these claims.

That a specific descriptive account is embedded in this item also points to why disagreement with this item does not entail that the participant believes there are moral truths. This is because rejecting a particular account of *why* there are no moral truths does not entail that you believe there are moral truths. It only entails that you believe there are no moral truths *for the reasons specified by that item*. One way to put this is that, while agreeing that moral beliefs are social conventions tells us what the participant thinks they are, to disagree that they are social conventions only tells us what the participant thinks they aren't. It doesn't tell us what they think they *are*. It doesn't tell us, for instance, that among the things moral beliefs are, they are claims about what is true or false, and that some of those claims are true. Yet in order for disagreement to reflect belief in moral truth, participants who disagree with this item would have to disagree specifically because they think that moral norms aren't

merely social conventions *because* they are also claims about what is true or false, and that some of these claims are true. It's not clear that this is why most participants who disagree with this item do so. We would have to specifically assess *why* they disagree to know whether this item is valid.

Many items exhibit this structural flaw: they treat agreement and disagreement as a single spectrum, according to which agreement entails the belief that there are no moral truths, while disagreement entails that there are moral truths. This is a mistake and illustrates a more general problem of researchers not thinking carefully about how to interpret both ends of a Likert scale. Researchers will often treat each end of a scale item as mutually exhaustive ends of a single spectrum of some phenomenon, X. Yet their items sometimes fail to properly reflect X. Instead, they mistakenly take some subset or instance of X, or a reason for believing X as a proxy for X. Then, when they ask whether people agree with the statement in question, they treat disagreement as an indication that one disagrees with X, rather than disagreement with the specific instance or subset of X, or reason for believing X. For example, suppose X = theism. To agree that X is to agree that God exists, while to disagree with X is to disagree that God exists. Now, imagine a researcher wants to measure whether participants are theists (that is, whether they believe X). Suppose they ask participants to express their level of agreement with the following statement:

*"The God of the Bible is real."*

Assuming the item is interpreted as intended, agreeing with this item necessarily entails that the participant is a theist. Yet disagreement *does not necessarily entail that the participant is an atheist*. A participant could believe God exists, but not the God described in the Bible. The same applies to indicating some reason *why* someone would believe in God:

*"The majesty of creation reveals God's existence."*

Once again, to agree entails that the participant is a theist. Yet to disagree does not entail that the participant is an atheist. A theist could believe God exists, but that the majesty of creation doesn't

reveal God's existence. The problem occurs whenever researchers present an item in such a way that, to agree with an item necessarily entails X, but mistakenly assume that to disagree with the item necessarily entails not-X.[133] Someone who believes there are no moral truths could just disagree that moral beliefs are merely social conventions. In fact, this is exactly what I think. I don't believe there are moral truths, but I also don't believe moral beliefs are merely social conventions. I think there is more to them than that. Yet there is no way for me to express this using the FMO.

There are other minor issues with this item. Philosophers and psychologists may understand a "social convention" in a specific, technical way that doesn't reflect how ordinary people understand this term. Is there some non-technical, shared understanding of "social convention"? I'm not sure. Similarly, I'm not sure how participants would interpret "could have been different." Such modal language could mean a wide variety of things.[134] Another problem worth reiterating is that this item is supposed to be conceptually distinct from moral relativism, yet it is not clear that this is the case. The notion that moral beliefs are merely social conventions seems fairly close to what a cultural relativist would think about moral claims. This is a problem for this item, because cultural relativists believe there are moral truths. But it is also a problem because it indicates that items intended to reflect a distinct category may not be conceptually distinct.

---

[133] This can also work in reverse, where disagreement does correctly entail not-X, but agreement does not necessarily entail X.

[134] e.g., it could mean under identical conditions, with the presumption that the laws of physics don't entail that the same events must necessarily occur, or it could mean that even if determinism is true, that things could have been different under some counterfactual conditions in which circumstances were slightly different. Indeed, it's unclear whether philosophers mean the same things by phrases such as "could have been different" or "could have done otherwise" as ordinary people, whether ordinary people mean anything in particular, or that if they do, that there are substantive philosophical commitments implicit in their views, or that philosophers themselves have adequately reflected on what this phrase means, all points Dennett (1984) stresses in challenging dogmatic appeals to the notion of "could have done otherwise." In short, there is little reason to presume that ordinary people would interpret such phrases in any particular way, in the same way as one another, or in the way researchers may be inclined to suppose.

| **FMO #4**<br>*No truth* | It is an illusion to think that anything is really morally true or false |
|---|---|

This is one of the stranger items on the FMO. It's not clear that beliefs can be illusions, rather than that illusions cause us to hold false beliefs. Also, people can have false beliefs without those beliefs being illusions. So this seems to characterize beliefs in a strange way, and to ask participants whether they reject moral truths for a specific reason, rather than for any reason at all. But I also want to emphasize just how *weird* the item is. It strikes me as very unconventional way of phrasing things to state that "it is an illusion to think that…" What does that mean exactly? Typically, an illusion is something that seems one way, but is actually some other way. While it may seem to people that some things are really morally true or false, but they are not, why would we say that the thought something is morally true or false is *itself* the illusion, rather than the illusion causing people to think things are true or false? If I'm in the desert, and there is a mirage, the natural way to describe this as an illusion would be to say that there was the illusion of an oasis. The mirage is the illusion; not the *belief* that there is an oasis. So we wouldn't say "It's an illusion to think there is really an oasis," since this treats the thoughts *about* the illusion (in this case, a mirage of an oasis) as the illusion itself. That's a strange and unconventional way to phrase things. And there are much more natural and simple ways to reflect the claim that there are no moral truths. Why not say "it's a mistake to think that anything is really morally true or false"? Or why not just say "Nothing is morally true or false"? That directly conveys what this subscale is supposed to reflect.

Also, this is supposed to be the "no truth" category. Yet this item also indicates that it's an illusion to think anything is morally *false*. This subscale is only supposed to involve the belief that there are no moral truths, not that there are no moral falsehoods, either. If you're an error theorist you think that moral claims are uniformly *false*. You just don't think it's *true* that anything is *morally incorrect* or

*morally wrong*. Zijlstra seems to have conflated a metaethical position about the truth status of first-order moral claims with endorsing or rejecting first-order moral claims. It's not clear that ordinary people would understand the difference, but, strictly speaking, agreeing that it's an illusion to think anything is morally false would rule out error theory. Since error theory holds that there are no moral truths, this item would appear to mistakenly exclude a metaethical position that should be included. Yet the more serious problem is conflation, and resulting ambiguity, between thinking things are morally "true" or "false," and thinking that things are morally "right" or "wrong." Think about how much more natural it would be to state that "it is a *mistake* to think that anything is really morally *right* or *wrong*." Yet by using the terms "true" and "false," this item is open both to this interpretation, where true/false stand in for first-order normative concepts of right/wrong, or in some other way, where true/false reflect non-normative indicators of the truth status of first-order normative claims.

Another problem with this item is that, like FMO #3, one could believe that there are no moral truths, but not believe that "it's an illusion" to think that there are. The notion that it's an *illusion* to think there are moral truths could be interpreted to imply some substantive notion of the psychology of those who mistakenly think anything is morally true or false. We wouldn't necessarily say that if someone is incorrect about something that this is because they are subject to some kind of illusion. Not all of our beliefs are the result of how things seem, phenomenologically.

Finally, the use of "really" may be somewhat of a problem. What does *that* mean? Does it mean stance-independently wrong? If so, that would be a mistake, since one could believe there are moral truths, but they are stance-dependent. If it doesn't mean this, then I'm not sure what it is supposed to mean.

| **FMO #5** **Relativism** | When two people have opposing beliefs about a moral issue, it is not necessarily the case that either or both are wrong |
|---|---|

This item is an abstract version of the disagreement paradigm. It therefore inherits many of the problems associated with the disagreement paradigm. This is enough to sink any confidence in its validity. Yet this item has a few unique problems. One of the problems with this item is its complexity: it is not *necessarily* the case that *either* or *both* are wrong? That's *tricky*! Participants are given a conditional statement with a negation of the modal operator and a disjunction, so we get something like:

$$\forall x\, \forall y\ (B(x,\ p) \land D(y,p)) \rightarrow \neg\Box (W(x,p) \lor W(y,p))$$

Even if we wanted to hyper-simplify this, participants would at the very least have to understand a negation of a modal claim regarding necessity with reference to a disjunctive claim, e.g., $\neg\Box(P \lor Q)$. It is not reasonable to expect ordinary people to be able to readily understand this. They are especially unlikely to do so given the context in which this item is presented. They are responding to an individual scale item that appears alongside other scale items. There is little incentive to take the time to figure out precisely what a single sentence means. "Either or both" is also hard to parse. It means A, B, or (A $\land$ B). This minimizes the risk of modal operator scope ambiguity by eliminating ambiguity between inclusive and exclusive readings of "or," but at the cost of an unwieldy and clunky phrasing that is technically accurate but hard to interpret.

The interpretative difficulties with this item are further compounded by the use of multiple negative qualifiers. It's *not* the case that people are *wrong*? Such language can reduce the quality of measures (Cassady & Finch, 2014; Hughes, 2009; Johnson, Bristow, & Schneider, 2004; Suárez Álvarez et al., 2018). That such language is used alongside modal language and disjuncts only compounds the difficulty of parsing this sentence.

To make matters worse, it's doubtful ordinary people will interpret "necessarily" in the same way as philosophers or one another, or even interpret it literally at all. In fact, "necessarily" may not do the work researchers want it to do at all. Suppose you're an ordinary person, and you're told that two people disagree about some moral issue. You are then asked whether you think that it's *necessarily* the case that at least one of them must be mistaken. You could interpret this in the intended way: accept that both people have conflicting moral positions, then consider whether it's necessarily the case that conflicting moral positions could both be correct (presumably by entertaining some form of relativism). Yet you could also interpret the question to be asking whether it's necessarily the case their opposing views are genuinely the result of conflicting moral standards. You could instead think that, e.g., both beliefs capture an element of a broader truth, that they could be conceiving of different situations, and each is correct about the respective situation that they are referring to, or they could both have positions that reflect equally valid ways of conforming to the same abstract moral rule, or that this is an epistemic question about whether it's possible that either (but not both) could be correct (this would require a performance error since this would be a modal operator scope error, but given the complexity of the question this isn't obviously implausible), and so on. When given the opportunity to explain why they thought two people disagreed, some of these possibilities are just the sorts of things people would say (Bush & Moss, 2020). And when given the explicit option to select options reflective of some of these views when presented with a disagreement, participants frequently selected such responses rather than exclusively favoring the strictly intended interpretations. In other words, *not necessarily* could serve as a queue that prompts participants to consider the many ways two people could both be correct that don't require moral relativism. In short, this item not only suffers from the problems associated with the disagreement paradigm, it is an especially confusing and complicated version of it. For instance, the statement "If two people have conflicting moral beliefs,

it's possible that they can both be correct" means roughly the same thing but eliminates some of the more confusing elements of FMO #5.

Assessment of open response data that involved asking participants to explain why they agreed or disagreed with this item, and to explain what they thought this item meant support the conclusion that most participants did reliably and clearly interpret it in metaethical terms.

| **FMO #6** *Relativism* | There is not one but many different answers to the question of what is morally right and wrong and these can be equally correct |
|---|---|

This item is worded in a way that makes it especially susceptible to the left hand conflation. A realist who endorses the existence of general moral principles can believe there are different ways to conform to that principle. The rule "show respect for the dead" may be stance-independently true. However, the precise way in which one demonstrates their respect for the dead may depend on local customs. In some cultures, this may involve cremation, in others burial, and still others endocannibalism. A realist may also believe that individuals are permitted an individual prerogative, even if that prerogative is circumscribed by moral constraints. For instance, someone may believe that "it is up to each person and a matter of personal decision whether they get an abortion." This is consistent with realism: one could believe it is stance-independently wrong to coerce people into having abortions or not having abortions, wrong to prohibit the liberty to choose whether to have an abortion, and so on. And they may believe that while people may choose to have an abortion or not, that it would be morally wrong for someone to choose to have an abortion for trivial or malicious reasons.

The problem with items like this is that they are intended to reflect relativism, but fail to disambiguate the notion that moral claims can be true or false relative to different moral standards from the notion that there can be multiple means of complying with a normative moral standard that

are equally correct *within the same moral framework*. There is nothing remotely implausible or unfamiliar about the latter: there are multiple correct ways to cut vegetables, solve math problems, build bridges, and so on. Such considerations may even be more salient to participants than considering relativism. As such, this is not a good item for assessing whether people are relativists.

Assessment of open response data that involved asking participants to explain why they agreed or disagreed with this item, and to explain what they thought this item meant support the conclusion that most participants did reliably and clearly interpret it in metaethical terms (see **Chapter 4**).

| FMO #7 *Relativism* | What is ultimately morally right and wrong is different for people with different moral views and from different cultures and societies |
|---|---|

This item could readily be interpreted as a descriptive claim about what is morally right and wrong *according to people* from different cultures. The central problem is that when this item states that what is morally wrong "is different" for people from different cultures, that the notion that it "is different" could be interpreted as the claim that different moral standards actually apply to different people, or it could be interpreted as the claim that *what people believe* is morally right or wrong is different. Consider a similar remark that simplifies and highlights the ambiguity:

*What is true or false is different for people with different perspectives and cultures*

While this could be interpreted as a claim about truth relativism, it seems to me at least as plausible that this statement means that what people believe to be true or false differs in accordance with their perspective and culture. In fact, this is overwhelmingly how people interpreted this item when asked to explain why they agreed or disagreed with it, and to explain what it means (see **Chapter 4**).

Another shortcoming with this item is that even if it were interpreted as intended, it would reflect agent relativism, but not appraiser relativism. An appraiser relativist would have to disagree with this item, despite endorsing the more common form of relativism among philosophers, and ex

hypothesi the most likely form of relativism among ordinary people as well. This shortcoming is compounded by collapsing individual subjectivism and cultural relativism into a single item and presenting it as a conjunct. A subjectivist may believe moral truth depends on one's moral view, but not their culture or society, while a cultural relativist might think the opposite. Collapsing the two presumes that this distinction won't matter to people. This may simply not be true. At worst, it lowers the precision of the item, since it cannot distinguish different forms of relativism from one another.

Finally, there is the phrase "ultimately." What does that mean? I study metaethics, and I don't know what that means. What do ordinary people think it means? Do they think it means the same thing as one another? Does the term influence how they interpret the item? And does that interpretation encourage or discourage an intended interpretation? I don't know. There's something questionable about scale items dealing with subtle questions like the nature of moral truth tossing in terms like "ultimately," without regard for the role that term plays in influencing participant interpretation.

| FMO #8 *Relativism* | What is morally right and wrong is relative to the moral beliefs of an individual, culture, or society |
|---|---|

This item is the best of the lot. Unfortunately, it's not obvious that ordinary people understand "relative" as intended, i.e., as a metaethical position about the indexicality of moral claims. There is little reason to be confident they would do so. Ordinary notions of relativism appear to entangle normative, descriptive, and metaethical considerations (Bush, 2016). As a result, agreement with this item may be understood to convey tolerance for people with different moral standards (a normative claim), or to convey the recognition that different people have different moral beliefs (a descriptive claim). Ordinary people also conflate relativism with contextualism (whether a general moral rule applies depends on situational factors), the etiology of moral beliefs (people often focus on *how we*

*acquire* moral beliefs, a related but distinct notion from descriptive relativism), and various other conflations. It may seem simple and straightforward to present participants with a direct and explicit question about whether morality is "relative." Unfortunately, people reliably fail to interpret this as intended. Once again, evaluation of open response data reveals that these conflations comprise the majority of participant responses when asked to explain why they answered this question the way they did and to explain what they think it means.

| FMO #9 *Universalism* | What is ultimately morally right or wrong is the same for all people at all times and places |
|---|---|

Like other items on the FMO, it is not clear what "ultimately" means. This may be a minor concern, yet there may still be reason to worry that including terms and phrases without carefully considering how they impact participant interpretation may be a mistake. The more serious problem with this item is that participants may conflate universalism with absolutism. By stating that what is right or wrong is "the same" for everyone in every time and every place could imply that there are moral rules that are rigid and insensitive to context, even though that is not what this item is supposed to convey. Universalism is also distinct from stance-independence, yet this item could be interpreted to reflect either or both.

| FMO #10 *Universalism* | Although people or cultures sometimes ignore moral concerns, moral norms apply anywhere and everywhere |
|---|---|

This item is double-barreled. What if you agree that people sometimes ignore moral concerns, but you do not agree that the same moral norms apply anywhere and everywhere? You have no way to express this. It is also trivially true that people and cultures sometimes ignore moral concerns. Who could

disagree with that? By frontloading this item with an obviously true remark, this could bias participants towards agreeing with it merely in virtue of how obviously true the first part of the statement is. It also entangles one's metaethical position with a non-metaethical descriptive claim: that people and cultures sometimes ignore moral concerns. To agree with the metaethical element of this remark *requires* expressing agreement with a descriptive claim, independent of whether or not one endorses that claim. Someone who, for whatever reason, did not think people sometimes ignore moral concerns, but still thought that moral norms were universal would have no way to express this. If the objection to this is that it's not plausible that anyone would think this, then why include the first part of this item? Why not just use the statement "Moral norms apply anywhere and everywhere"?

There is an even more serious problem with this item, however. Once we set aside the first part of the item, which isn't relevant to metaethics, consider the second part: "moral norms apply anywhere and everywhere." This item does not state that the *same* moral norms apply anywhere and everywhere, yet this is essential to the item representing universalism. Without conveying that the moral standards that apply anywhere and everywhere must be the same moral standards, this *does not express a universalist stance on morality at all*.

This problem is sufficient to undermine the face validity of this item on its own. Yet there is another problem. What does it mean for moral norms to "apply"? This could be read in the intended, normative way. That is, it could mean that it's a fact that people are subject to some universal set of moral standards. Yet it could also be understood in descriptive terms. For instance, it's illegal to drive on the left side of the road in the United States. A natural way to express this is to say that this law "applies everywhere in the United States." The respect in which this law is in effect is that there is intersubjective agreement regarding the boundaries of the United States and the jurisdiction of the relevant law. The respect in which it applies to us is a *descriptive* fact, not a normative one. It is simply true that people in the US are subject to US law, regardless of whether we think they should or

shouldn't be. Likewise, when we say that moral norms apply to people, does this mean that there is some normative fact about what moral standards people are subject to, whether this is the case relative to some moral framework, or in virtue of there being stance-independent moral facts that determine what those people should do? Or does it instead refer to the mundane descriptive fact that anywhere and everywhere people are held morally accountable by the people around them, are morally judged, and so on? Or does it mean something else entirely? Speaking for myself, I don't know what it means to say that a moral norm applies to someone. The term "apply," when used in this context, is obscure, and could mean a variety of different things, and perhaps in its obscurity it would turn out on reflection to not mean anything in particular at all.

| FMO #11 *Universalism* | What is morally right and wrong for me here and now is also morally right and wrong for people elsewhere, even for people living in different countries and part of different cultures |
|---|---|

Like other universalism items, this could be conflated with absolutism, stance-independence, or a descriptive claim.

| FMO #12 *Universalism* | Despite the diversity of moral views between individuals, cultures, and societies, there are moral norms that should apply universally |
|---|---|

This is a normative claim about what should be the case, and not what *is* the case. It is not an appropriate measure of belief in universalism. Even if you don't think there are norms that apply to everyone, you could still think that there ought to be, or that it'd be good if there were, or that it would be in everyone's interests to subscribe to the same universal standards. The *should* in use here isn't explicitly moral, and one could have nonmoral reasons for thinking the same moral standard should apply universally, e.g. someone might think that it would have positive practical consequences for people to adopt some moral norms, such as norms against murder or stealing. In fact, I do think this.

In spite of being an antirealist that does not believe there are any universal moral facts, if you were to ask me if I think there are any moral norms that should apply universally, I would have a long list of norms to suggest.

A second, less serious problem with this item is that it is double-barreled and entangles the metaethical component (universalism) with a descriptive component (descriptive relativism). Universalism does not require the belief that there is a diversity of moral beliefs. In fact, some philosophers argue for universalism by denying that there is significant moral diversity. Some do so on *a priori* grounds (Cooper, 1978; Davidson, 1973; Foot, 1978a; 1978b Myers, 2004; cf. Lillehammer, 2007), while others argue that careful examination of the empirical evidence would reveal that claims that there is widespread moral diversity are overstated (Gowans, 2021).[135] This is typically achieved by presenting evidence of widespread moral agreement regarding at least some minimal set of moral standards and attempting to show that apparent disagreements about fundamental moral values are primarily due to disagreements about the nonmoral facts (Gowans, 2021). Since an outright rejection of moral diversity is one of the central arguments universalist philosophers make as part of their objection to relativism, it seems strange to presume this rationale isn't available to ordinary people, and that to endorse universalism requires doing so in spite of widespread moral diversity.

| **FMO #13** **Absolutism** | Although people disagree about what is morally right and wrong, I believe in the existence of specific moral principles that can settle any moral disagreement |
|---|---|

The absolutism subscale is supposed to include items that entail both *belief in general moral principles* and a commitment to those moral rules having *no exceptions*. For instance, an absolutist might endorse a

---

[135] These references are provided by Gowans (2021), who summarizes *a priori* arguments and empirical arguments against descriptive relativism.

general moral rule, such as "do not lie," and believe that there are no exceptions to this rule, so people should never lie, regardless of the circumstances.

Unfortunately, this item does not specify that the moral principles are absolute. That a general set of moral rules "can settle any moral disagreement," does not mean that the rules are inflexible or do not have exceptions. A rule that is flexible, commensurable with other moral considerations, and permits exceptions can still do so in a principled way, and could therefore reliably solve any (relevant) moral disagreement. It just does not follow that if one's rules are sufficient to solve moral problems that those rules must be absolute. Think about the alternative: a set of moral principles that cannot settle any moral disagreement. Would such a set of principles fail to be "absolute"? No. They would just be *incomplete* or *inadequate*. If you had a set of moral rules that could address nine out of ten situations, but could not settle the tenth, the rules you have could still be absolute, they could just fail to apply to the tenth issue. Since this item simply does not make it clear that the moral principles in question are absolute, it is not a face valid measure of absolutism.

Another problem with this item is that it refers to *specific* moral principles. Yet a specific moral principle is not necessarily a general one. A particularist (i.e., someone who denies generalism about moral rules) could believe in specific moral principles. It would even make sense to use the term specific *for the exact purpose of distinguishing one's position from generalism*. That is, "specific" could be used to mean that the principle in question is *not* general. Thus, not only does "specific" not help in conveying belief in general moral principles, but may actively work against the intended interpretation. This item may therefore fail not only to convey that the moral principles in question are absolute, but that they are general, as well.

There are still further problems with this item. The emphasis on the ability to settle moral disagreements is ambiguous and potentially misleading. In principle, you could believe there are absolute general moral principles, even if those principles cannot settle any moral disagreement. The

problem is ambiguity between the ability to settle disputes *in principle* or *in practice*. Suppose your moral principles are sufficiently comprehensive that they can determine what people should do in all possible situations. If so, then there would be a fact of the matter about what side of a moral disagreement was correct. Yet it does not follow that we would know what this fact was. Epistemic limitations could bar access. If so, then while the moral disagreement could be settled in principle, it could not be settled in practice. Yet another worry about practical resolutions to moral disagreements looms even larger. Suppose you know what the moral facts are, and they conform to some general and absolute set of moral principles. This does not mean that you could successfully convince people with contrary moral beliefs. People may be stubborn, or have defective epistemic practices, or for some other reason remain recalcitrant. The notion of "settling a moral disagreement," could be understood to refer to some abstract, idealized disagreement in which people are rational and interested in mutually arriving at the truth. But real world moral disagreements are not like this. They carry all the psychological idiosyncrasies and limitations of actual people.

For comparison, imagine that one's mathematical principles allow one to correctly solve any mathematical problem. Does this mean that you could settle any *actual* mathematical disagreement? No. People can be confused, or ignorant, or incompetent, or refuse to consider your position, or be committed to some alternative account of mathematics. Yet your inability to convince people would not mean that your position isn't correct. The ability to settle disagreements *in practice* may be a more natural interpretation of this item than the ability to do so in principle. Ordinary people are plausibly more oriented towards considering actual events in the real world, rather than hypothetical or abstract considerations. Yet to interpret this item to reflect the view that we should be able to resolve moral disagreements in practice would be to interpret it in an unintended way. A belief in absolute general moral principles would at best only entail that they could solve moral disagreements in principle, not necessarily in practice, for the same reason that an absolute and general set of mathematical axioms

may be able to solve any math problem, even if you could not settle any actual mathematical disagreement.

Yet this objection ignores a deeper worry about the notion that one's moral principles ought to be able to settle any dispute: such a characteristic doesn't even follow from a belief in absolute general moral rules. You could believe that there are absolute and general moral principles, but also believe that they may be insufficient to resolve every moral disagreement even in principle. The ability to settle any moral disagreement in principle would require something like "moral comprehensiveness": the belief that the moral principles that exist are sufficient to determinately settle all moral problems. Such a quality does not require absolutism or generalism, nor does generalism or absolutism logically entail comprehensiveness. These are simply different properties of a set of moral principles.

Like other items on the FMO, this item is also double-barreled. Once again, the participant must express agreement both with the claim that "people disagree about what is morally right and wrong" *and* the claim that there are "a specific moral principles that can settle any moral disagreement." In principle, someone could accept one of these claims but reject the other, yet they are linked in such a way that one must agree with both or neither. This, by itself, is not a problem. It seems implausible any significant number of ordinary people would deny that there are moral disagreements. The problem is that, like other items on the FMO, it begins with a claim that is completely unobjectionable, then follows this with a more controversial claim. By first presenting people with a claim almost everyone would agree with, participants who would otherwise claim that they do not "believe in the existence of specific moral principles that can settle any moral disagreement," are faced with the cognitive burden of agreeing with *part* of the statement, but *not the whole thing.*

The claim that people disagree about what is morally right or wrong is also a descriptive claim, so agreeing with this statement requires agreeing with a descriptive claim. This is not appropriate if one's goal is to isolate and measure people's metaethical standards. Finally, although this item is intended to measure belief in absolute general moral principles, it could be mistakenly interpreted as a descriptive claim, as universalism, or as the claim that there are stance-independent moral facts. Note that the latter part of the claim states that, "I believe in the existence of specific moral principles that can settle any moral disagreement." Specific moral principles that could resolve moral disagreements can exist *even if you do not endorse those principles*. For instance, Christianity offers a set of moral guidelines. Perhaps those rules could settle any moral disagreement. The use of "specific" could also allude to the possibility that there is only *one* set of moral standards, which could imply universalism. Finally, the notion that one believes in "the existence" of moral principles could imply the reification of those principles, i.e., that those principles are not merely subjective standards individuals hold, but that theory exists *independently of the people who endorse them*. If so, this could imply moral realism, even though this isn't the intent of the item. This could be amplified by the notion that these principles could settle moral disagreements, since one of the reasons why a moral principle could settle all moral disagreements *is because they are correct*. In fact, this seems more plausible than the notion that they can solve any moral disagreement because they are general and absolute.

| FMO #14 *Absolutism* | Certain actions are morally wrong and they remain morally wrong even in the rare case that no one believes so |
|---|---|

The notion that some actions are wrong even if nobody believes that they are would indicate moral realism, not absolutism or generalism. Nothing about agreeing or disagreeing with this item would clearly indicate a belief or disbelief in the notion that there are general exceptionless moral rules, such as "never lie." This is not a face valid measure of absolutism, and has been miscategorized. This item

is also double-barreled since it refers to actions that are morally wrong even in the "rare case" that no one believes this. It's unclear why the actions in question would have to be ones for which it is rare that no one thinks they are wrong.

| | |
|---|---|
| **FMO #15** *Absolutism* | There are absolute moral rules that apply to all people, including those who do not acknowledge these principles |

This item entangles absolutism with universalism. Absolutism does not require or entail universalism, and by combining the two, one must either endorse or reject absolutism and universalism at the same time. The inclusion of "including those who do not acknowledge these principles" exacerbates this problem further, since it alludes to stance-independence, which could give the unintended impression that this item reflects moral realism.

It is also unclear how people understand the term "absolute." Do they reliably interpret it to mean "rules that have no exceptions"? That's an empirical question, and it is not safe to assume that they will interpret it in the way researchers intend. This item also fails to refer to the notion that there are *general* moral rules that determine whether lower-level actions are correct or incorrect, so it fails to serve as a face valid measure of one of the criteria for the construct. Finally, like other items, this one makes use of the term "apply," despite its obscurity. This could prompt participants to interpret the item in descriptive terms. After all, local laws apply to me in the sense that I am subject to them and they will be enforced if I break them, but it does not follow that these laws are "true" or that they are absolute.

| FMO #16 *Absolutism* | There is, in all circumstances, one correct answer about what is the morally right thing to do |
|---|---|

This item does not indicate that the correct answers are a product of general moral principles. As such, this item fails to serve as a face valid measure of one of the criteria for the construct. However, it may be the best representation of absolutism of the items on the FMO.

| FMO #17 *DCT* | The correct answer to any moral issue can be found in a sacred book or text (for example, the Bible, the Qur'an, the Torah, or another) |
|---|---|

This is not a valid measure of DCT. According to Zijlstra, DCT is "the view that whether an action is morally right or wrong depends on the commands of a divine being" (p. 2). This item does not state that the answers that can be found in religious texts depend on, or are constituted by, God's commands. A sacred book could include claims about what is morally right or wrong, but this does not require that those moral facts are true in virtue of being commanded by God.

Furthermore, note that However, Zijlstra also states that "Those who support this theory regard religious texts and/or authorities as sources of moral knowledge" (p. 2). While it is true that they often do so, this is not a necessary feature of DCT itself. DCT simply holds that moral facts depend on God's commands. These commands do not have to be revealed through religious texts or mediated by religious authorities, even if they could be, or even if some are in practice. Yet this item states that the correct answer to *any* moral issue can be found in a sacred book or text. Even proponents of DCT do not have to think this is the case. This would require that, for example, a Christian would have to believe that *every answer to every moral question* can be resolved by consulting the Bible. This is not an implication of DCT, and it would be plausible for Christians believe that the Bible is not an exhaustive guide to addressing every possible problem in ethics.

This item would be more appropriately characterized as a claim about moral epistemology, and a rather implausible and strong claim at that: that all moral knowledge may be acquired exclusively by consulting religious texts. This *simply isn't a feature or entailment of DCT*.

| **FMO #18** **DCT** | The only actions that are ultimately morally right or wrong are those actions that God prescribes |
|---|---|

This looks like a face valid measure of DCT. See, it is possible! Note, however, that technically this item only states right or wrong actions happen to be the ones God prescribes, but it does not say that they are wrong *because* God prescribes them. Someone could, in principle simply think, regardless of *why* anything is right or wrong, God knows and prescribes what is wrong even if it isn't God's prescribing it that makes it the case that it's wrong. So, even this item may not really capture DCT.

| **FMO #19** **DCT** | God is the only true source of knowledge about what is morally right or wrong |
|---|---|

DCT does not require that God is the only source of moral knowledge. It's also not clear what it means for God to be the "true" source of moral knowledge. A source could be a final, or terminal, but there could also be intermediary sources of knowledge. For instance, if God issues moral decrees to priests, and priests inform people of God's decrees, it would be reasonable to say that the priests are a "source of knowledge," and, since they are correctly informing us of God's actual commands, they would in an important respect be a *true* source of knowledge, since they are certainly not a *false* source of knowledge. What they wouldn't be is the ultimate, or final source of knowledge. But that's not what this item states.

| **FMO #20** **DCT** | Without the existence of God, nothing is truly morally right or wrong |
|---|---|

It is logically possible to believe that God is necessary for there to be moral truth without believing that moral facts specifically reflect or depend on God's commands. There's a lot I'm willing to quibble about, and maybe someone more into theistic morality would take greater issue with this item, but it strikes me as adequate.

### S3.3.7 Objectivity of Morality Scale (JRT5)

While the MRS and FMO represent comprehensive efforts to devise valid measures of folk metaethical belief that I believe fall short, due in part to the inherently difficulties of specifying metaethical distinctions in a way participants will understand, some studies present measures that have not gone through this more rigorous process. My colleagues and I have done this ourselves. Sometimes there is no satisfactory validated scale or measure available, so you make your best effort. Even so, the measures one comes up with can sometimes turn out, on reflection, not to be ideal. This next scale is a brief, recent attempt to devise a short metaethics scale. Johnson, Rodrigues, and Tuckett (2020; hereafter JRT) devised a 5-item objectivism scale (JRT5) purportedly "derived from previous research," i.e., Goodwin and Darley (2008; 2012). Since G&D's conception of objectivism conforms to my use of realism, the JRT5 is presumably intended to be a scale for measuring realism. JRT state that an EFA indicated two factors, a *normativity* factor (3 of the items) and a *subjectivity* factor (2 of the items). The items on the JRT5 appear in **Table S3.3** along with a summary of the reasons why these items are not face valid.

**Table S3.3**

*The JRT5 Moral Objectivism Scale (Johnson, Rodrigues, & Tuckett, 2020)*

| Items | | Conflations |
|---|---|---|
| **Normativity** | | |
| JRT #1 | *Every good person on earth, regardless of culture, holds these beliefs.* | Normative conflation<br>Descriptive conflation<br>Universalism conflation<br>*Not about realism\** |
| JRT #2 | *The truth of these beliefs is self-evident.* | Epistemic conflation<br>*Not about realism\** |
| JRT #3 | *A society could not survive without its citizens holding these beliefs* | Practical conflation<br>Descriptive conflation<br>*Not about realism\** |
| **Subjectivity** | | |
| JRT #4 | *If someone strongly disagreed with you about one of these beliefs, it is possible that neither you nor the other person are mistaken [R]* | Disagreement paradigm<br>Overcomplicated<br>"Strongly" is unnecessary |
| JRT #5 | *There are no clearly true or false answers to these questions. [R]* | Epistemic conflation<br>*Not about realism\** |

*Note.* Items flagged as *not about realism\** denote items for which no plausible interpretation would reflect the construct these items are intended to measure. Thus, the items don't conflate realism with other considerations, but fail to represent realism at all.

The inclusion of the "normativity" factor is puzzling, since normativity isn't an element of realism, and therefore should not be included in a scale intended to measure realism. However, the label does not do a good job of capturing the content of the items within it. Only one item includes a substantive normative element: "Every good person on earth, regardless of culture, holds these beliefs." JRT #2 is an epistemic claim with no normative moral content, and JRT #3 is a descriptive claim about the practical consequences of holding beliefs rather than a normative claim.

Out of the two items in the *subjectivism* factor, JRT #4 is a version of the disagreement paradigm. This version of the disagreement paradigm cannot distinguish subjectivism from other forms of relativism or noncognitivism, so this item could not distinctively serve as a measure of subjectivism, but could instead only serve at best as a measure of realism versus antirealism. JRT #5 is an epistemic claim, so it cannot serve as a measure of realism. It also has nothing to do with subjectivism. This factor is simply mislabeled: neither item has anything to do with subjectivism. With four of the five items lacking any reasonable claim to face validity, the JRT5 is not a valid measure of moral realism. Let us now consider each item individually:

| JRT #1 **normativity** | Every good person on earth, regardless of culture, holds these beliefs. |
|---|---|

The claim *could* be interpreted as a claim that morality requires that all people share the same moral standards. One problem with this is that a relativist or antirealist could believe that all good people share the same moral standards, namely, *their* moral standards. I endorse moral antirealism. I do not believe there are stance-independent moral facts, but I also do not believe there are any universal moral facts. However, I do have a position on which moral standards are good or bad. I am just as capable as a realist or universalist of having a stance on which moral standards are good, and I can, consistent with antirealism, maintain that every good person holds some shared set of moral beliefs. There may be a mistaken presumption implicit in this item that if you're an antirealist or specifically a relativist that you cannot maintain that all good people would share some set of moral beliefs in common. This simply isn't true. The likely culprit for this misunderstanding is a failure to disambiguate agent and appraiser relativism, and to presume that if you're a relativist that you're an agent relativist, and therefore think that I am good if I adhere to my standards, and that other people are good if they adhere to theirs. But a relativist need not think this. They could think that they are good and others

who don't share their standards are bad *relative to their standards*, and that they are bad and others are good *relative to those people's standards*. There is simply no conflict between being a relativist and thinking anyone who does not share your moral beliefs is bad.

However, even if this were the case, this would entail universalism, *not* realism. Since this scale was purportedly modeled on G&D's scale, it is worth noting that they explicitly distinguish their notion of "objective" from "universal," and maintain that their goal is to measure the former and not the latter. As a result, this item fails in every way to reflect the intended construct of interest.

| **JRT #2** *normativity* | The truth of these beliefs is self-evident. |
|---|---|

The notion that the truth of particular moral beliefs is self-evident is a specific and very strong epistemic claim. It has nothing to do with normativity, so it is mislabeled. But the means by which one thinks we can acquire moral knowledge is distinct from whether or not one believes that moral facts are stance-independently true. Moral realism is a *metaphysical* claim concerning the nature of moral truth; it is not a claim about how we acquire knowledge of these moral truths. Many moral realists, perhaps the majority, would not endorse the notion that their moral beliefs are not only true, but self-evidently true. Thus, disagreement with this claim is consistent with being a realist. Since disagreement cannot distinguish antirealists from realists who do not think their moral beliefs are self-evident, it is not a legitimate method for distinguishing realism from antirealism.

Note, as well, that any response to this item presumes cognitivism. To agree with it is to agree that there are moral truths, and that those truths are self-evident, while to disagree is to agree that there are moral truths, but to deny that they are self-evident. Notably, insofar as a belief in moral truth is presumed to be a belief in stance-independent moral truth, both agreement and disagreement with this item would presume the truth of realism. But insofar as truth is not presumed to be stance-

independent, then no response to this item would indicate one's stance towards realism at all. One would have to think that self-evidence necessarily entails realism, while denying self-evidence doesn't, yet this is a fairly strong assumption to bake into a question about realism. At the very least, any possible response presumes that one thinks there is *some* kind of moral truth in principle and implies that there are moral truths in practice as well, which means that noncognitivists and error theorists have no way to respond. Since Beebe (2015), Davis (2021), and Pölzler and Wright (2020a; 2020b) have evidence that significant subsets of their participants endorse these views, it is not legitimate to use a forced choice paradigm that precludes a substantial portion of participants from responding appropriately.

Finally, note that this item states that the truth of *these beliefs* is self-evident. It's all or nothing. Participants who might regard some moral claims as self-evident but others as not being self-evident have no way to express this. This item, like many others on metaethics scales, presumes not only the determinacy of folk metaethical views, but the uniformity of those as well, despite ample prior evidence of variation across moral issues.

| **JRT #3** *normativity* | A society could not survive without its citizens holding these beliefs |
|---|---|

This item is not about metaethics, nor is it even about normativity. Instead, it is a claim about the practical necessity of particular beliefs. A society could not survive if people didn't believe they needed to eat or drink water. It does not follow that it is a stance-independent moral fact that we should eat and drink. Just the same, that some moral belief may be practically necessary does not entail that it is a stance-independent normative fact. I'm an antirealist. I don't think society could survive if people did not have rules prohibiting violence, theft, and wanton destruction. That doesn't mean I think we have a stance-independent moral duty to enforce these rules.

| JRT #4 *subjectivity* | If someone strongly disagreed with you about one of these beliefs, it is possible that neither you nor the other person are mistaken [R] |
| --- | --- |

This item inherits all the problems associated with the disagreement paradigm. It has a few other odd features, however. This item asks whether if a person disagreed with the participant about *one* of multiple beliefs, whether it is possible neither they nor the other person is mistaken. Given that people give different answers to different moral issues, for many participants *this would make no sense*. This is like giving someone a list of foods they like and dislike, then asking "if someone served one of these foods for you, would you like it?" There's no way to answer this without knowing *which* food you are referring to. The only way around this is to presume that it is irrelevant which belief was selected, because one applies the same metanormative standard to all of them. Given that few participants do so, this is not an appropriate assumption to make. Like many other items that employ abstract measures, the structure of the item presumes metaethical uniformity.

| JRT #5 *subjectivity* | There are no clearly true or false answers to these questions. [R] |
| --- | --- |

By including the term "clearly," this item is an epistemic question about the degree to which some of the questions presented have clear answers. As such, it is not a valid measure of moral realism, because it is concerned with how *obvious* the answers to these questions are, not whether there is a stance-independent fact of the matter about them. Yet even if "clearly" were removed, then the item would be expressing a claim that most closely resembled noncognitivism, not antirealism in general. Error theorists think there are answers to these questions (they are all false) as do relativists (they are true or false relative to different moral standards) and constructivists (they are true or false according to some

constructive process e.g., the conclusions rational agents do or would agree to using the appropriate procedure; Bagnoli, 2021). Lastly, this item asks only if there are true or false questions to the questions given in the study. It does not follow that if participants believe there are no "clearly true or false" answers to *those* questions that they think there are no "clearly true or false answers" to *any* moral questions. Gauging a person's metaethical standards by extrapolating from their response to a subset of moral considerations may not be an ideal way of assessing their metaethical standards, since it is more prone to error than methods that eschew such extrapolation.

**S3.4 Nichols & Folds-Bennett's (2003) response-dependence paradigm**

Nichols and Folds-Bennett (2003; hereafter N&F) introduced an innovative and surprisingly underutilized[136] paradigm for evaluating whether ordinary people endorse moral realism[137] by assessing whether they believe moral claims are *response dependent*. As N&F define it, "the basic idea is that a property is response-dependent just in case that property is constituted by the responses it elicits in a population; so, the same object or event might have different response-dependent properties in different populations" (p. B25). They offer "icky" as an example of a prototypical response dependent property. Since ickiness depends on the subjective attitudes of particular individuals, ickiness can only be judged relative to the potentially varying responses of different people. If the property "moral wrongness" were response-dependent in the same way, then there would be no stance-independent fact about whether a prototypical moral violation, such as hitting someone, were wrong or not. Instead, the moral wrongness of hitting a person could only be judged relative to the responses of different individuals.

---

[136] To my knowledge, no attempts have been made to replicate this study or use a similar paradigm.
[137] They use the term "moral objectivism." However, the examples and descriptions that they provide indicate that they are referring to moral realism.

This observation led N&F to design a straightforward test: if children regard properties like "icky" as response dependent, but do not regard moral norms as response dependent, this would provide at least some indication that they treat moral norms as stance-independent.[138] This is what they found. Children did treat properties like "icky," "yummy," and "fun," as response-dependent, but not moral properties like "good" or "bad." Unfortunately, there are several shortcomings with the paradigm they used. Children were first asked whether something had a particular property, e.g. whether grapes are yummy. If they said "yes" they were then asked a question that was designed to reveal whether they treated the property as response-dependent:

> You know, I think grapes are yummy too. Some people don't like grapes. They don't think grapes are yummy. Would you say that grapes are yummy *for some people* or that they're yummy *for real*?[139] (p. B27)

The judgment that grapes are yummy "for some people" was interpreted as response dependent, while judging that they're yummy "for real" indicated that "yumminess" is not response-dependent. An analogous procedure was used for the moral properties "good" and "bad." One problem with this question is that participants were given a forced choice between "for some people" and "for real," but these responses are neither mutually exclusive nor exhaustive of possible responses. After all, if grapes are yummy for some people does this mean they're not yummy "for real" for those people, or not really yummy at all? And what does it mean to say they're yummy "for real"? A natural alternative to this option would be that they're "not yummy for real," but this option isn't available. Instead, the implication is that if grapes are yummy "for some people" that somehow this isn't "real"

---

[138] Alternatively, if they treat them the same, this would suggest that children may not be objectivists after all. And if children don't appear to treat any properties as response-dependent, they may simply fail to draw a distinction between objectivism and nonobjectivism (B. 25). Gill (2009) refers to this possibility as metaethical indeterminism, and argues that it may apply to folk metaethics in general. According to Gill, people may fail to draw various metaethical distinctions when engaging in everyday moral thought. If so, then there would be no fact of the matter about whether everyday use of terms and concepts like "moral wrongness" are treated as objective of nonobjective.

[139] If they said "no", the experimenter would select alternatives until they found one that the child would say "yes" to, then adjust the script accordingly.

yumminess, as though response-dependent properties aren't non-real properties. But response-dependent properties are real properties, and something being yummy to someone can still be thought of as yummy "for real." Think about your own food preferences. Do you think when you eat something and find it tasty, that it isn't tasty "for real"? I don't. I think what it means for something to be tasty "for real" *just is* for it to be tasty relative to whoever it is tasty to. Subjective tastiness isn't *fake* tastiness.

Conversely, what would it mean to reject this choice and say that grapes are yummy "for real"? This is contrasted with "for some people." That they *aren't* yummy "for some people"? Technically, the logical contrast to "for some people" would be "for no one," but since "for some people" pragmatically implies a non-negligible quantity, but probably not a majority or everyone, this remark is ambiguous, and the natural contrast to it could be "everyone," "most people," "no people," or some combination of these. In other words, the pair of response options are treated as though they are mutually exhaustive and conflicting positions, but they are neither mutually exhaustive nor do they clearly conflict with one another, yet this is further exacerbated by each side of the putative dichotomy having an ambiguous negation, yet among the possible options for these negations, none seem to fit well with the response options participants are actually given.

Since selecting one of these options could plausibly be interpreted as rejection of the other, yet neither choice is unambiguously inconsistent with the other, it is unclear what participants take their responses to indicate when they choose one option over another. In short, it is unclear what either option means when they are presented as two presumably incompatible choices or how the presentation of either as an alternative to the other may have influenced how participants interpreted them. It'd be a bit like asking someone if they think that "some people are good at math," or if "math is real." Since these don't appear to conflict with one another, it's unclear how people would interpret this such that their responses would indicate anything in particular. But the failure to

provide a proper dichotomy could force people to infer or fill in the blank by spontaneously theorizing about what one or both of these remarks could mean, and such considerations could result in idiosyncratic interpretations that go undocumented and heighten the risk of interpretative variation and of unintended interpretations.

It's also not obvious that judging that grapes are yummy "for real" indicates response independence. A forced choice question between the *quantified* "for *some* people" and a second alternative that isn't explicitly quantified could be interpreted as a *descriptive* question about the *proportion* of people who think grapes are yummy, since one potential interpretation pragmatically implied by the term "for real," in the context of a direct contrast with "some," is *many*, *most*, or *all*. As a result, children who believe that people usually think grapes are yummy may interpret "for some people" and "for real" as two ends of a continuum about the overall proportion of people who think grapes are yummy, with the former representing comparatively fewer and the latter representing comparatively more. If so, their choices may simply reflect attitudes about how many people think grapes are yummy, rather than attitudes about whether yumminess is response dependent.

"For real" could also be taken to express a stronger attitude; indeed, "for real" is sometimes used as an exclamation or to convey seriousness, so it could even be interpreted as a question about how strongly the participant felt that grapes were yummy. "For real" could even imply strong consensus. Children may already be aware that food preferences vary. They are less likely to be told, or to encounter people, who think that helping others is not good, or that hitting people or pulling hair aren't bad. If so, children may be more likely to choose "for some people" for preference items (e.g. about what's icky or boring) than about moral items, because they recognize people often have different food or play preferences but rarely have different basic moral attitudes about helping and hitting. If so, we would expect children to display less apparent response-dependence for extremely

high consensus taste or play preferences, but more apparent response-dependence if more concrete, ambiguous, or controversial moral items were used (though the latter may be more appropriate for older participants). If so, this might explain why children continued to favor "for some people" for preference-related items (like whether grapes are yummy) in the second study, but treated disgusting acts and violations of social conventions like they treated moral acts, since disgusting acts and social conventions likely enjoy higher consensus than taste preferences. Finally, children could be more motivated to choose "for real" for moral items because doing otherwise could signal that they aren't "good" kids that take morality seriously. If so, the result could be due to demand effects (Orne, 1962).

More generally, these problems illustrate the sheer opacity of the question. It's not clear that children aged 4 to 6 years old can be expected to understand the question as intended and respond appropriately. And even if children did treat the moral items used in this study as response-independent, *all* of the moral items concerned harm/care. It is unclear whether the same response pattern would generalize to other moral transgressions. It may not even be the case that the same pattern would generalize to other moral issues related to harm or care. After all, research on adults using an expanded range of moral violations has consistently found that most people regard some moral transgressions in realist terms and others in antirealist terms (e.g. Beebe, 2015; Beebe et al., 2015; Goodwin & Darley, 2008; 2012; Wright et al., 2013) with the most recent studies finding a majority favored antirealist responses for moral items (Davis, 2021; Pölzler & Wright, 2020a; 2020b). At best, N&F's findings only show that children might be realists about a narrow range of moral issues, but this does not justify generalizations about the moral domain as a whole.

Given these concerns, it's unclear whether N&F found much evidence that children are moral realists. Even so, paradigms designed to assess response dependence remain a viable option for evaluating folk objectivism. Note, as well, that these studies had very low sample sizes ($n = 19$

and $n = 13$). Future attempts could evaluate comprehension, include a broader range of moral and preference items, conduct higher powered studies, develop and test appropriate analogues in adult populations, and develop alternative wordings that circumvent the potential difficulties with the wording N&F used.

**S3.5 Fisher et al.'s (2017) direct question paradigm**

*Direct question* paradigms are any methods for assessing folk metaethical belief that involve explicitly asking people whether they endorse a particular metaethical position using the technical terms that characterize that position, or using technical terms or phrases intended to reflect that position. In truth, the degree to which a paradigm is direct or indirect is a matter of degree, with some studies presenting comparatively more indirect approaches, such as Zijlstra's (2021) implicit moral objectivism paradigms, through scales that include a range of items, some of which are fairly direct, e.g., the FMO's inclusion of the item "What is morally right and wrong is *relative* to the moral beliefs of an individual, culture, or society," (emphasis mine), which explicitly employs language used to reflect relativism. There is no sharp dividing line, so this category is best used to reflect items at the extreme end of the continuum for explicitness and directness. One criterion I'd propose for distinguishing "direct question" paradigms is that they *only* rely on direct questions, rather than including such questions among other measures. The FMO does have some direct questions, but these are scattered amongst less direct questions. What distinguishes a pure form of a direct question is the presumption that explicitly asking people whether they endorse realism, antirealism, objectivism, relativism, or some other metaethical position is, by itself, a valid way to measure that person's metaethical standards.

At present, Fisher et al. (2017; hereafter FKSK) is only one study that employs a direct paradigm. They ask participants:

*"'Should [action][140] be allowed?' Please tell us whether you think there is an objectively true answer to this question."*

*1 (Definitely no objective truth)—7 (Definitely objective truth)*

I previously critiqued this measure in Bush and Moss (2020). One observation I did not make in that critique was that this paradigm, like many others, relies on forced choice: *all* responses are treated as a determinate stance about whether there are objective moral truths. This obscures any possibility of indeterminacy and necessarily generates a pattern of results that, no matter what they are, could be taken as an indication that people have a position on the status of moral realism. Direct questions contribute to the potential for the appearance of determinate folk metaethical views, which the appearance of folk determinacy could be an artifact of study design.

A more distinct and serious problem for direct paradigms is that they are only valid if participants interpret stimuli as intended. FKSK explicitly state that the goal of the direct paradigm is to eliminate "ambiguity about whether we really are asking about whether there is an objective truth about the topic" (p. 1127). I do not believe they were successful. It would only remove ambiguity if participants reliably interpreted "objective" in the same way as the FKSK. Simply using the word "objective" does not ensure that the question isn't ambiguous, since people could interpret the term "objectively" as it appears in the question in a variety of different ways, many of which could differ from researcher intent. In other words, FKSK mistakenly presume that using a word ("objective") that corresponds to a concept you label using that word (*objective*, understood to refer to stance-independence), that participants will have no trouble understanding the word "objective" to refer to the concept *objective*. This is not plausible because words themselves can be ambiguous, and "objective" is no exception. "Objective" could mean *unbiased*, it could mean *measurable according to some shared,*

---

[140] Moral issues were randomly selected from: same-sex marriage, marijuana legalization, teaching evolution in school, abortion, and violence in videogames" (Fisher et al., 2017, p. 1126).

*standardized measure* (e.g., thermometers provide an objective measure of the temperature), *capable of being tested* or *subject to a discrete, fixed subset of possibilities that exclude subjective considerations* ("These exams provide an objective test of employee suitability"), *discernible to others* (e.g., some medical diagnoses are distinguished by whether others can perceive their symptoms), and so on. "Objective," like many other words, is polysemous, and can even be used to refer to notions that make no sense in the context of the question, such as a *goal* or to particular elements of grammar. It is not reasonable to assume that simply asking people whether morality is "objective" would eliminate any ambiguity, without any evidence that people would reliably interpret it in the same way in the context of the question.

There can also be little doubt that FKSK intended for their question to reflect realism, since they describe objectively true claims as ones that are "established by facts independent of any particular person's judgment," which indicates that by "objective" they mean "stance-independent" (p. 3). The only way that their question would serve as a valid measure of realism would be if participants reliably interpreted "objectively true" to mean stance-independently true. Yet when asked "In your own words, what does it mean to say that moral truth is objective?" I found that only 12.3% of participants clearly interpreted "objective" to mean "stance-independent," with a majority (71.9%) instead interpreting the notion that moral truth is objective in some unintended way, such as the notion that moral judgments are unbiased, or the notion that moral claims are "black and white" such that we can make clear and definitive distinctions between right and wrong. The paradigm did not employ the precise language used by FKSK. Perhaps if it had the rates of intended interpretation would be higher. But given the overwhelming evidence that participants struggle to interpret questions about metaethics as intended, and that they do not reliably interpret questions that refer to morality as "objective" or "relative," in unambiguously metaethical terms, the onus is on those who employ direct paradigms to demonstrate that participants interpret them as intended.

Another problem with FKSK's study is not the validity of the measure itself, but how they use the results of their direct question paradigm to bolster the validity of their other findings. FKSK employee a version of the disagreement paradigm in their first three studies that is similar to the one employed by Sarkissian et al. (2011):

> *Earlier studies show that people take opposite positions on the issue of [issue]. Given that people have opposite views, at least one side must be wrong.*

[1 = Strongly Disagree, 7 = Strongly Agree]

FKSK devised the direct paradigm because they were concerned that participants may not understand that FKSK are "asking a question about the metaphysical issue as to whether there is an objective truth about the question under discussion" (p. 1127). FKSK claim the following that their results replicate previous findings, in that they find the means of the two conditions are significantly different in the predicted direction. Since they find a pattern of results that were similar to their previous study (using the disagreement paradigm), they conclude that "This result replicates our main finding from the previous experiments and suggests that participants were correctly interpreting the original measure" (p. 10). This does not follow. You cannot conclude that because two studies found a significant difference in the same direction that both of those studies were measuring the same thing, much less that they were measuring what you were intending to measure. Suppose the initial studies they conducted generated replicable results using invalid measures. you cannot demonstrate that this faulty method is valid by finding a similar result using a new method. The new method may also be invalid. Given that neither method was developed independently, they could even be valid for similar reasons, e.g., that the questions are ambiguous or confusing in a way that directionally favors one pattern of response over another. After all, the same findings are consistent with the possibility that participants interpreted *both* questions in the same unintended way, or interpreted them in different unintended ways that incidentally yielded a similar response pattern.

To illustrate why, suppose researchers wanted to know if people are happier after exercising than watching TV. After an hour of either exercising or watching TV, they give participants a survey, but accidentally give them the wrong one, and they are instead asked how much they agree with the statement "I am *thirsty.*" Unsurprisingly, participants in the exercise condition report greater agreement. After discovering this error the researchers redo the study. But once again, they mix up their survey with another one, only this time participants are asked how much they agree with the statement "I am *tired.*" Once again, participants in the exercise condition report greater agreement.

Since both studies found higher agreement in the exercise condition, does this demonstrate that both studies accurately measured *happiness*? Of course not. Both questions were measuring an entirely different phenomena that simply happened to differ in the same direction as each other: participants who exercise are both thirstier and more tired. It does not follow that these questions are measuring the same thing as one another. More importantly, this illustrates how both questions could be consistent but fail to measure what the researchers intended to measure (i.e. *happiness*) in the first place. Of course, if two studies appear to measure the same thing, and yield similar results, this could still be taken as *some* evidence for their validity, but mutual corroboration alone is insufficient to establish validity.

## S3.6 Behavioral studies

There have only been two studies examining the behavioral consequences of manipulating people's belief in realism or antirealism. Rai and Holyoak (2013) found that participants presented with an argument for relativism were more likely to cheat and steal than participants presented with an argument for realism.[141] Young and Durwin (2013) report that participants primed with a statement

---

[141] Rai and Holyoak (2013) refer to it as "absolutism." However, they are referring to *realism* as the term is used here since they define absolutism as "The philosophical position [...] that some moral beliefs are objectively true, and reflects facts that are independent of any social group's specific preferences" (p. 995, emphasis ours).

in favor of realism were twice as likely to donate to a charity than participants primed with relativism. Unfortunately, neither of these studies used valid manipulations

### *S3.6.1 Rai & Holyoak (2013)*

Rai and Holyoak (2013) devised a test to examine whether people primed with relativism were more likely to cheat and steal than people primed with realism. Why suspect relativism may compromise moral behavior? They claim that realists fear that embracing realism could render people incapable of judging people with different moral standards. This could result in tolerating other people's actions, even when those actions are harmful. Such tolerance could then become internalized, causing relativists to engage in immoral behavior. Rai and Holyoak are quick to note that such unsavory consequences are not part of the conceptual content of relativism:

> Note that there is no intrinsic reason why a relativistic conception of morality need adopt all of these positions. In philosophy, meta-ethical relativism accepts that our moral beliefs are ultimately subjective, but does not hold the normative position that this subjectivity forces us to tolerate behaviors that we find morally disagreeable, nor that our own behavior should necessarily be impaired. (p. 996)

Since endorsing relativism does not *necessarily* entail any particular normative implications, the relationship between relativism and immoral behavior could only be a contingent one. In other words, if one could only believe in relativism *if* they engaged in immoral actions, then there would be no need to test whether relativism caused immoral behavior: it would be a feature of belief in relativism that it did so. Since it is not, any connection between the two must be established empirically.

Rai and Holyoak attempt to demonstrate that belief in relativism causes immoral behavior by showing that, when participants are presented with an argument for relativism, they engage in comparatively worse moral behavior than participants presented with an argument for realism. However, in order for these manipulations to demonstrate that it is increased belief in relativism that is causing the increase in moral behavior, the arguments that they gave to participants would have to *only* directly manipulate belief in relativism. If they include arguments for anything else, we would not

be able to tell whether increased immoral behavior was caused by whatever psychological changes were induced by these other arguments. In other words, they must avoid introducing confounds. Unfortunately, this is exactly what Rai and Holyoak did. Rather than presenting arguments for realism and relativism, they instead presented participants with statements that simply assert that realism and relativism are true, and then argue that *because* realism and relativism are true, that we have a moral obligation to be to either impose our values on others (in the case of realism) or that we must refrain from imposing our values on others (in the case of relativism). Here is their description of their manipulations:

**Argument for realism**

> In the moral absolutism condition, participants were told that some moral values are objectively right or wrong and it is our duty to impose our values on other groups of people regardless of what they believe because female genital mutilation causes irreparable harm and is an intrinsic form of violence ("This is not a situation where we should exercise tolerance for other cultures' practices, because ultimately, they are morally wrong…. Our moral beliefs are based in intrinsic facts about what is right and wrong in the world…. We should realize that our feelings are telling us something important and be willing to act on them."). Killing newborn infant girls in countries that favor boys was used as a source analogy for explaining the objectivity of our moral values ("We know that the practice is wrong because the wrongness of murdering newborns based on their gender is not a matter of opinion—it is simply evil"). (pp. 996-997)

**Argument for relativism**

> In the moral relativism condition, participants were told that our moral values are subjective opinions and we cannot impose them on another group of people because they see female genital mutilation as a necessary, purifying act ("…it is not our place to judge and it would be wrong for us to impose our values on other people…. If we grew up in a culture where female genital mutilation is practiced, we would think it was the morally right thing to do…. We have to step back from our immediate gut reactions and realize that our own moral beliefs are simply a product of our cultural upbringing rather than any objective set of criteria"). Male circumcision was used as a source analogy for explaining the subjectivity of our moral values ("…male circumcision is also painful, can have risks…. Yet, it is seen as normal, and perhaps even necessary by many people in the United States"). (pp. 997).

The explicit purpose of their study was to examine "whether exposure to arguments for moral relativism and moral absolutism could impact moral behavior" (p. 996). Neither of these manipulations present an argument for realism or relativism. Rather, they consist of mere assertions that morality is realist or relative, followed by the claim that these metaethical positions have strong normative implications.[142] The bulk of the manipulation appears to consist of exhortations to act in certain ways. In the case of realism, participants are told that because realism is true, we must be intolerant of others, impose our moral standards on them, oppose violence, and act on our feelings. In the case of relativism, participants are told that, because relativism is true, they must not impose their values on others, they must not judge others, and they must not trust their gut feelings. With respect to the specific moral issue in question, they are also given a rationale for why the practice should be tolerated since they are told it is seen as "necessary," and "purifying," which introduces substantive information about *why* the people who endorse these views believe what they do, which could serve as mitigating or exculpatory considerations that alter the moral status of the actions in question.

Note that the manipulations that they use appear to provide false information to participants as well. By their own lights, Rai and Holyoak acknowledge that relativism does *not* entail that we must tolerate others. Yet they present participants with an argument that *because* relativism is true, *therefore* we must tolerate others and abstain from imposing our standards on them. This is, strictly speaking, *false*.

---

[142] Since they do not provide the full text of the manipulations, it is possible they included arguments for realism and relativism. If so, it is strange that they did not include this in their description. However, even if they did, this would barely help. Absent such arguments, they both failed to provide an argument for realism/relativism, *and* inappropriately included arguments for something else entirely. If they did include such arguments, this would simply mean that they had arguments both for realism/relativism, *and* arguments for something else entirely. The latter would still represent a confound, and would still render their study invalid.

Yet the more serious problem with these manipulations is that they have almost nothing to do with arguing for realism and relativism, and almost everything to do with normative arguments that require us to act in specific ways. This is bizarre, given that Rai and Holyoak explicitly acknowledge that metaethical positions have no necessary normative implications. If people's moral behavior did change as a result of these manipulations, we have no way of knowing whether this is due to a change in their belief in realism or relativism, or if it is instead due to a change in their beliefs or attitudes as a result of the rest of the substantive content of the manipulations.

Recall the fear Rai and Holyoak attribute to realists. The fear is that if people embrace relativism, this will cause them to be tolerant of others, and if they are tolerant of immoral behavior, this will cause them to be tolerant of their own immoral behavior. If this account were correct, what we'd want to show is that belief in relativism causes people to be more tolerant of others, and we'd want to show that tolerating others leads to immoral behavior.

The proper way to test for this would be to provide evidence of this causal pathway. This would involve demonstrating that belief in relativism *caused* increased tolerance, and that increased tolerance *caused* immoral behavior. The proper way to test for the first of these causal claims would be to manipulate relativism, then look to see if it caused an increase in tolerance. In other words, suppose you wanted to demonstrate the following causal connection: A→C. You suspect that this is because A→B→C. Yet it could turn out that ∼(A→B), or that ∼(B→C). If you wanted to show that A→B→C, you would have to show that A→B *and* that B→C. You cannot simply cause people to believe B by telling them "A→B" then testing whether C is the case, then, if it is, conclude that A→C. It could be that B→C but not A→B! You don't have to think that if A→C that this is because A→B→C. Yet to show that A→C, you would have to in some way manipulate A, then see if this resulted in C. If you *also* manipulate B, or manipulate B *instead*, then you cannot claim that A→C. All you have evidence for is that B→C, or at best that A&B→C. In the case of relativism causing immoral behavior, this

would mean that if A = relativism, and C = immoral behavior, this would require manipulating relativism, then seeing if this compromised moral behavior.

Yet this is not what Rai and Holyoak did. Instead, they *directly told participants that realism and relativism entailed certain normative implications*, e.g. intolerance and tolerance. In other words, they presented participants with an argument that served to directly induce the causal association that they should have been testing for. This is the equivalent of *causing* people to think that A→B and that B, then testing for C. Even if you show that C, this does not demonstrate that A→C. Strangely, their description does not even include a direct argument for realism or relativism, but instead seems to presume they are true, then describe the normative implications of their truth (which, to reiterate, is not entailed by these positions *according to Rai and Holyoak!*). Yet even if they did include arguments for realism and relativism, this would just mean that they manipulated belief in A, B, and A→B, then found C. This would still not show that A→B or that B→C.

To illustrate why this is inappropriate, suppose researchers wanted to test the hypothesis that belief in God causes increased charitable giving. How should they test this? Presumably, they would manipulate belief in God, then see if increased belief in God caused increased charitable giving. Yet suppose that, instead of presenting people with arguments for and against the existence of God, they instead presented participants with the following arguments:

(1) Belief in God condition: Because God does exist, we should be more selfless and more empathic towards others

(2) Disbelief in God condition: Because God does not exist, we should more selfish and less empathic towards others

Now suppose they found that participants in condition (2) gave less to charity. Should we conclude that decreased belief in God causes people to engage in less charitable giving? No. We have no way of knowing whether the claims that God does or does not exist are driving the change in charitable giving, or whether the rest of the manipulation is doing the work. These wouldn't even be properly

described as manipulations of belief and disbelief in God. At best, they would also have to be described as manipulations of selflessness/selfishness and empathy. Unfortunately, the structure of these manipulations closely parallels the structure of the manipulations employed by Rai and Holyoak. Simply put, they did not cleanly manipulate realism and relativism, so their results do not provide clear evidence that increased belief in relativism leads to increased immoral behavior.

The second of their studies uses a different manipulation, but it is not much better. Drawing on prior research, participants in the realism condition are told that:

**Realism**
*Morality is defined by things that are just morally right or wrong, good or bad. There are absolutely clear guidelines, that always apply to everyone, whatever the circumstances.*

**Relativism**
*Morality is defined by values that are shaped by our culture and upbringing. There can never be absolutely clear guidelines and what is right or wrong depends entirely upon the circumstances.*

Neither of these adequately or exclusively manipulates the intended metaethical position. The realism condition does not manipulate realism at all. To do would require telling participants that there are stance-independent moral facts. No part of this statement reflects this view. Instead, it expresses several metaethical and non-metaethical views. The realism condition expresses moral absolutism ("just" morally right or wrong, "whatever the circumstances"), moral universalism ("always apply to everyone"), and an epistemic claim that these moral guidelines are "clear." This is simply not an expression of realism.

The relativism manipulation is a little better. The first part states that morality "is defined by values that are shaped by our culture and upbringing." This is a bit vague. What does it mean to say that they are *defined*? Is this a descriptive claim about how people acquire moral beliefs? If so, then it isn't, strictly speaking, a metaethical stance, but an etiological stance about how moral beliefs emerge in a population. And what is meant by "shaped"? Does this mean these sources have *some* causal

influence on the content of our moral values? Or does it mean that they are strictly determined by them? This statement is *vague* and does not clearly indicate that moral claims are *true or false* relative to the moral standards of different people and groups. It is not an adequate expression of relativism. This is followed by similar content to the realism condition. Once again, this content is not about the proper metaethical concept. The notion that there "can never be absolutely clear guidelines" expresses either an epistemic claim, a claim about particularism or a general sensitivity to contextual considerations, or both, while the claim that "what is right or wrong depends entirely upon the circumstances" seems to express particularism. Particularism is not the same thing as relativism. You can be a realist and a particularist, or a relativist and an absolutist. This item simply expresses the wrong concepts, and is therefore not an appropriate expression of relativism.

This may be a bit pedantic, but neither of these statements express *arguments*, either. Both merely consist of assertions. Perhaps that is a strength of the study. If such minimal manipulations can produce strong behavioral effects, this would support their conclusions. On the other hand, perhaps it should raise our suspicions about whether changes in metaethical beliefs are really behind these results. There can be no doubt this is how Rai and Holyoak interpret their results:

> Taken together, the present findings indicate that meta-ethical worldviews related to moral relativism and moral absolutism *can have a causal impact on people's moral judgments and behaviors*. Specifically, increased moral relativist and decreased moral absolutist perspectives may *lead to* relaxed moral standards and willingness to engage in immoral behaviors. (p. 999, emphasis mine)

Their findings do not provide good evidence for these claims. There are also several other reasons to worry about these findings.

One concern is that they provide no manipulation checks, so there is no direct evidence that they successfully manipulated people's metaethical standards. Second, there are serious concerns about the external validity of their findings. Manipulating people's degree of belief in relativism and realism does not necessarily demonstrate that people's metaethical worldviews have a causal impact

on their behavior in the real world. This is because the manipulations used in studies like this typically produce *temporary* changes in people's beliefs or attitudes, test the *immediate* behavioral changes only, and often do so by *increasing or decreasing confidence* in the belief, *not causing a discrete shift from belief to disbelief* (or vice versa). There is a considerable difference between what it is like to have a particular worldview, and what it is like to have one's current worldview challenged by a rival worldview. Studies that challenge a participant's current worldview by espousing some alternative worldview may induce psychological states that differ from the psychological states present in the mind of someone who has come to adopt that worldview, and had time to internalize and come to grips with it. Consider how a group of devout Christians might feel if they are given an antitheist passage that forcefully argues against the existence of God and the afterlife. Someone whose worldview is challenged by this passage may experience confusion, anxiety, fear, sadness, anger, doubt, a reduced sense of purpose and meaning, and so on. If you test their behavior immediately after being exposed to this passage, you may very well find differences compared to participants who did not read the passage. Perhaps they would be more aggressive, more likely to lie or steal, or express less concern for their family members. Yet this would not be good evidence that *becoming an atheist* would cause these changes. An atheist is someone who doesn't believe in God. It is not someone who does believe in God, but whose views have just been challenged. A Christian in a state of doubt is not the same thing as an atheist, and we wouldn't (and shouldn't) expect insights about the behavior of one to generalize to the other. Likewise, a participant exposed to an expression of realism or relativism is not identical to a person who embraces realism or relativism with respect to morality. There are several important considerations to keep in mind:

(1) If a manipulation successfully induces a temporary change in belief, differences in measured outcomes may be temporary. There is little reason to think a temporary change in belief would cause permanent psychological changes in the absence of evidence or a compelling theoretical rationale

(2) Even if a manipulation successfully induced a permanent change in belief, differences in measured outcomes may be temporary

(3) Studies typically do not induce permanent changes in belief. Even if they did, if a permanent change were necessary for a permanent concomitant change in measured outcomes, appropriate evidence (e.g. longitudinal) would be necessary to establish the permanent change in both the belief and the measured outcomes

(4) Studies that fail to provide evidence that they successfully manipulated the belief of interest provide weaker evidence than those that do, since we cannot be confident that the belief in question was actually manipulated

(5) A change in one's confidence in their beliefs is qualitatively different from a discrete change from one belief to another. Studies which show that reduced confidence in a belief without showing a discrete change in belief should not be used as evidence of the latter. We cannot assume that a theist with reduced confidence in the existence of God is in the same state of mind as someone who went from being a theist to being an atheist

It is unlikely that the manipulations used by Rai and Holyoak successfully convinced anyone to be a realist or a relativist. At best, participants were likely exposed to a particular perspective about morality that they didn't already have, or that wasn't salient, and were somewhat more inclined towards it than prior to their exposure to it. The inference that relativists in the wild are more likely to engage in immoral behavior is a bit of a leap. Think again of Christians exposed to an argument against their beliefs. It is understandable that if a Christian is convinced of atheism that, in the immediate wake of their loss of faith, they may behave differently. They might find themselves groping for a new worldview, or they may grapple with feelings of confusion, loss, betrayal, and other negative psychological states. It can take time to come to terms with these changes and to develop a new worldview. And even after one adopts a new worldview, it may take time to internalize that view in a way that fully manifests in one's behavior. That interim period may be characterized by behavior that isn't representative of what it is like to disbelieve in God, but instead reflects the behavioral changes that characterize a transitional state from one worldview to another. Likewise, people who transition from one metaethical stance to another may or may not show any long-term behavioral changes. Any

changes that we observe may instead reflect the fallout of the transition period that occurs when one moves from one worldview to another.

Another problem with their findings is that Rai and Holyoak did not test participants in advance to determine what proportion were already relativists. This is a problem, because some recent studies found that a majority of participants consistently favored relativist responses to questions about metaethical questions across multiple paradigms and with respect to multiple moral issues (Pölzler & Wright, 2020a; 2020b). How can we claim that convincing people that relativism is true would cause decreased moral behavior if a substantial proportion of those participants were already relativists to begin with? Whatever effect the manipulation had on them, it could not have been the result of causing a non-relativist to *become* a relativist, *if* we think they were already relativists to begin with.

They also presume, plausibly enough, that most participants would agree with the *normative* content of the concrete moral examples they used (e.g., female genital mutilation and arranged marriages). This may be relevant, since a participant's metaethical stance may vary depending on the moral issue in question, and they recognize that future studies should look at the behavioral consequences of responses to different moral issues. But note that they generalize from the behavioral consequences of statements ostensibly about metaethics that are embedded in a broader set of stimuli centered on specific normative issues, such as female genital mutilation. Yet by entangling metaethical claims with normative claims, we cannot be sure that the impact that exposure to the normative claims, or the interaction of exposure to metaethical and normative claims, is having on participants. And we cannot assume that if participants are convinced to adopt a relativistic stance towards a particular moral issue that they have or would adopt a relativistic stance towards other moral issues, or towards morality in general. Thus, while Rai and Holyoak want to draw inferences about the behavioral consequences of a generally relativistic worldview towards the moral domain as a whole, their findings

do not permit such inferences. The majority of prior research on folk metaethics overwhelmingly suggests that participants can and often will express different metaethical standards towards different moral issues. Maybe they would do so in a way that is stable and consistent under particular circumstances, but at present we do not know how and why people would develop such a consistent worldview, and we cannot infer that statements supporting realism or relativism and then applying these standards to a particular moral issue successfully prompt a consistent realist or relativistic stance towards the moral domain. Indeed, we don't even know if ordinary people share a clear conception of the moral domain, and there is some reason to doubt that they do (Wright, Grandjean, & McWhite, 2013), or that every population does so (Berniūnas, 2014; 2020, Machery, 2018; Stich, 2018).

Lastly, if I am correct that most people have no determinate metaethical standards, these findings may have few if any broad behavioral implications. If most ordinary people don't have a determinate metaethical view in the first place, then even if a commitment to relativism would compromise moral behavior, this would be irrelevant: if nobody actually were a relativist, then these behavioral consequences would not manifest. Even if people did have a determinate stance, it does not follow that this would have any significant impact on their behavior. Studies that expose people to a particular philosophical position might cause an immediate change in behavior, but that change might be present only when the philosophical position is made salient. A recent reminder that one is a relativist may prompt lax moral behavior, but under most ordinary circumstances the fact that one is a relativist may fade into the background, and have little or no impact on one's behavior or thought processes.

Given these many considerations, there is little reason to believe that a commitment to relativism would cause people to behave immorally. As an addendum, note as well that if it did do so via the route alluded to here, this compromised moral behavior would be predicated on a mistake: relativism *does not entail that we should tolerate other people or that we should abstain from imposing our values on*

*them.* This is a normative position, not a metaethical one. As such, even if relativism caused a decline in moral behavior, it would do so only through its contingent association with an inferential mistake about its normative implications. If participants were exposed to a statement regarding realism or relativism, along with an accompanying statement that this had no normative implications and therefore had nothing to do with whether there should be more or less tolerant, this may attenuate or eliminate any negative behavioral consequences. Note that relativism simply maintains that moral standards are true or false relative to different standards, such as the standards of individuals or groups. If you are a relativist, you could still think that, relative to your standards, a particular action is wrong for everyone, everywhere, and you could still impose your moral standards on others. Moral relativism does *not* require tolerance for other people's moral standards. If participants are informed of this, such mitigation via correction about the normative implications of metaethical positions could be mirrored by similar awareness among members of a population committed to relativism. If so, then it could turn out that it is not relativism that causes immoral behavior, but mistaken beliefs about the implications of relativism.

### S3.6.2 Young & Durwin (2013)

Young and Durwin (2013; hereafter Y&D) report the results of two studies which found that participants primed with a statement in favor of realism were more than twice as likely (50% compared to 22%) to donate to charity than participants exposed to a statement in favor of antirealism in a field experiment (study 1) and expressed greater willingness to donate to charity under lab conditions (study 2).

These are remarkable results, and there is considerable reason to worry about the strength of these effects. Y&D report that the data in the field study (study 1) were gathered by one of the experimenters themselves, A. J. Durwin. As such, the canvasser was not blind to their hypothesis, and this could have had a significant impact on their results. This is especially plausible given that the

canvasser employed a script that included numerous points at which the canvasser's body language, tone, perceived enthusiasm, assertiveness, and other characteristics relevant to success in receiving donations could have influenced donation rates. Here are some examples from their description of the study:

> (1) "Engagement of the passerby began with a smile and asking the passerby whether he/she had ever heard of the charitable organization"
>
> (2) "If the passerby continued walking slowly but did not stop, the canvasser asked the passerby to stop for just a minute to help him practice his presentation."
>
> (3) "If the participant initially refused to donate, the canvasser attempted to persuade the participant to donate, focusing on the relatively low cost to the donor and the relatively high gain for the people in need."

Note that these are just a few of the many steps at which the degree to which an unblinded researcher could have unconsciously influenced the outcome, by exhibiting more effort, enthusiasm, persuasiveness, and other characteristics in the realism condition compared to the antirealism condition. The whole point of blinding researchers to hypotheses is, of course, to minimize precisely these kinds of biases. The problem is that ensuring the canvasser was blind to the hypothesis was *especially* crucial in this type of study, given that the success rates were heavily dependent on canvasser performance. Perhaps there is little reason to worry. After reporting their results, they make the following claim:

> We replicated this basic pattern in hypothesis-blind canvassers (see Supplementary Results). In sum, priming participants to consider moral realism doubled donation rates.

Unfortunately, the supplemental data does not provide strong support for the results of study 1. In the supplemental data, they report that the same primes (realism vs. antirealism) were used by seven street canvassers blind to hypothesis. However, whereas in study 1, donation rates were compared across conditions by the same canvasser, donation rates in the supplemental section compared donation rates per hour collected on the day they used the prime to their "lifetime average," which

included the day they used the prime. Their rationale for doing so is that the charitable organization

they worked with already collected data. The results are not very encouraging. Here is what they report:

> Consistent with Experiment 1, the number of monthly donations per hour trended towards being higher on the day the moral realism prime was used compared to lifetime averages (realism: 4.77, lifetime: 2.52; $t(6)=1.02$, $p=.051$; Supplementary Figure 1). The two canvassers who used the antirealism prime did not receive any donations (antirealism: 0, lifetime: 2.14). A mixed-effects ANOVA yielded an interaction between prime (realism vs. antirealism) and test (prime vs. lifetime average) ($F(1,7)=5.71$, $p=0.048$, partial $\eta^2=0.45$). (Young & Durwin, 2013b)

There are several reasons to worry about these results. First, given their own data, they *failed* to replicate

the results reported in the main study, since it was not the case that $p < 0.05$. While the reported $p$-

value is "marginally" significant, it is still misleading to claim to have "replicated" the "basic pattern"

of one's results in a supplemental section when the results were not statistically significant, and to

report the results of an interaction effect in an ANOVA that barely skirt by on statistical significance.

Second, they did not compare the results of the realism versus antirealism and control conditions for

the same canvasser, but instead compared results of realism prime to the canvasser's lifetime average,

so they were not comparing results from the same measures, nor did they use the same type of

analyses. As such, it is also misleading to describe the comparison as a "replication" in the first place.

Setting aside concerns about the data, I want to focus on the content of the statements

themselves. The statements are purportedly intended to reflect realism and antirealism. Unfortunately,

like the manipulations used by Rai and Holyoak, both are poor representatives of the metaethical

positions they are intended to represent:

**Realism**
*Do you agree that some things are just morally right or wrong, good or bad, wherever you happen to be from in the world?*

**Antirealism**
*Do you agree that our morals and values are shaped by our culture and upbringing, so there are no absolute right answers to any moral questions?*

The realism condition does not clearly convey moral realism. It does not state or allude to the notion that there are stance-independent moral facts. Instead, it asks whether some things are "just" morally right or wrong, "wherever you happen to be from in the world." This more closely approximates universalism, *not* realism. This is troubling because Y&D cite Goodwin and Darley (2008; 2012) in their description of realism, despite the fact that Goodwin and Darley go out of their way to explain that realism is not the same thing as universalism. The realist condition could also be interpreted as an expression of moral absolutism. The notion that some things are "just" right or wrong may imply a rigid, "black and white" set of moral standards that are insensitive to context. Finally, it could be interpreted as a descriptive claim that some things are regarded as morally right or wrong everywhere. Given these possible interpretations, and empirical evidence that people frequently interpret expressions intended to convey realism in these and other unintended ways, it may be that few if any participants interpreted the question to be a question about whether moral realism is true.

The antirealist condition likewise fails to adequately convey antirealism. First, it's worth noting that Y&D conflate antirealism with subjectivism and cultural relativism:

> Importantly, moral antirealists do not deny the existence and importance of moral values; antirealists simply assert that moral values reflect the beliefs of a person or a culture, rather than immutable facts that exist independent of human psychology. In other words, like subjective preferences, (e.g., chocolate tastes better than vanilla), rather than objective facts, moral values may depend on the psychology of an individual or community. (p. 302)

Only subjectivists and cultural relativists think that moral values reflect the beliefs of people or cultures, *noncognitivists* do not think moral values reflect beliefs at all, but instead reflect non-propositional states. Their description also appears to exclude error theory, since error theorists do *not* think that moral values are like subjective preferences. Instead, they think that moral values reflect beliefs *and* that those beliefs concern matters of immutable facts that exist independent of human psychology; they just think there are no such facts.

Furthermore, the notion that "our moral values are shaped by our culture and upbringing" is a descriptive, psychological claim about how we acquire moral beliefs. It is also uncontroversially true. This question is, like others, frontloaded with a descriptive claim that does not entail any particular metaethical position. Rather than conveying a simple propositional claim, it includes the more complicated claim that *because* something is true, *therefore* something else is true. This is similar to a double-barreled or complex question. It also makes no sense. Relativists don't think that relativism is true *because* our moral beliefs are shaped by our culture and upbringing, yet this statement suggests that this is the reason why one would endorse relativism and reject realism. I would disagree with this question, and I'm not a realist!

Lastly, even the implication of the fact that our moral values are shaped by our culture and upbringing does not clearly convey relativism: it simply states that because this is the case, that "there are no absolute right answers to any moral questions." This does not clearly convey the claim that there are no nonrelative, or stance-independent moral facts, so it cannot serve as an expression of antirealism, unless participants interpreted "absolute" to mean "stance-independent," and there is little reason to think the majority of participants would do so. Second, participants could easily conflate this question to be a question about the rejection of universalism, or the rejection of absolutism, *not* the rejection of realism. In short, this question does clearly reflect a question about whether antirealism or relativism is true. It is ambiguous, and could be easily interpreted in a variety of unintended ways. Since the authors do not provide a clear description of what constructs they have in mind by "realism," and "antirealism" it is not even clear what question would serve as an appropriate operationalization of the intended interpretation.

Given these considerations, there is good reason to doubt that Y&D's studies accurately and reliably manipulated belief in realism and antirealism and that they did so without introducing confounds. If researchers wish to transform philosophical concepts into testable claims about human

psychology, it is incumbent on researchers to adequately understand the philosophical concepts that they operationalize, and it is incumbent on editors to identify reviewers familiar with the relevant philosophical distinctions.

**S3.7 Theriault et al. (2017, 2020)**

Theriault et al. (2017; 2020) provide an interesting and unique measure of folk realism and antirealism. All participants were presented with a set of moral and nonmoral statements. For each statement, they were asked a set of three questions:

(1) To what degree is this statement about **facts**?
(2) To what degree is this statement about **morality**?
(3) To what degree is this statement about **preferences**?

They describe their rationale for this approach as follows:

> We wanted to test participants' intuitions about metaethics without unnecessarily constraining their responses. Prior work has typically imposed a zero-sum relationship between judgments of morals as objective or subjective [...] and although this may reflect the philosophical distinction, it also constrains how participants are allowed to express their intuitions. (p. 1588)

Fair enough! This is as good of a reason as any I could think of. Many studies *do* force participants to categorically favor one or another of different metaethical accounts. Allowing them the option to endorse more than one metaethical account is an excellent idea. Theriault et al. add that "It is possible that participants see morals as both fact-like and preference-like to some extent, and a categorical (or one-dimensional) approach rules out this outcome before testing this" (p. 1588). Their set of three questions are intended to solve this problem.

Unfortunately, their attempt completely fails to provide a valid measure of folk realism and antirealism. First, and most importantly, agreeing that a given moral statement is a "statement about facts" does not provide a valid measure of moral realism. As David Moss and I point out in a critique of another study that makes the same mistake, "both objectivists and subjectivists believe that

statements express matters of fact" (Moss & Bush, 2021, p. 2; Joyce 2015). In other words, whether a moral claim conveys a *fact* is not exclusive to realism; subjectivists (and other realists, e.g., cultural relativists) *also* think moral claims express facts, they just don't think those facts are stance-independently true. Indeed, as the quote above illustrates, Theriault et al. characterize their measures as one intended to capture *subjectivism in particular*. Yet subjectivism *entails* that moral sentences are "about facts." Since both realists and subjectivists think moral claims are "about facts," this item cannot be used as a measure of realism as opposed to subjectivism, since both realists and subjectivists should completely agree.

This problem alone completely invalidates their measures. Yet it isn't the only problem. Consider a statement they provide in the text: "It is irresponsible for airlines to risk the safety of their passengers." Now, consider the question of whether this statement is "about facts." This question is *unclear*. What does it mean for this sentence to be "about" facts? "About" in what respect? There are a variety of distinct facts this statement could be "about," some of which wouldn't be normative facts, they'd just be references to the mundane descriptive content of the passage. For instance, it would be hard to make sense of the question unless one assumed certain facts about airlines and passengers and the relationship between them, not least of which is the fact that airlines and passengers *exist*. The question even asks if the statement is about "facts," rather than "a fact." So is it intended to ask whether the statement is about *multiple* facts? That's strange. If the goal of the question is to determine whether the participant thinks there's a stance-independent moral fact of the form, "it is a stance-independent moral fact that it is irresponsible for airlines to risk the safety of their passengers," would that be just one fact? The question also makes no explicit reference to whether the facts the question is asking about are *moral* or *normative* facts, so we have no way to know (without asking) whether the facts participants have in mind are the relevant kind of fact. Note, as well, that there are descriptive readings of the claim, e.g., one can believe it's a descriptive fact that airlines have taken on the

responsibility to care for passengers, and that as a matter of descriptive fact they are held responsible according to various laws, social norms, and so on for treating passengers in particular ways.

It's also strange to ask whether the statement is *about* facts, because a moral realist would hold that a moral statement *asserts* a fact. If I say, "Julius Caesar was born in Rome," is this a statement *about* facts, or is it a statement *of* fact? If you wanted to know whether I thought this statement was stance-independently true, it would be strange to ask whether I thought it was "about" facts. It would make far more sense to ask me if the statement *was* a fact. Asking whether such statements are "about" facts is at best a strained and awkward way of asking the question that could prompt further confusion.

Finally, another problem with this item is that asking whether a statement is about "facts" conflates whether the statement is *propositional*, that is, whether it is an assertion that is capable of being true *or* false, and whether it is a *true proposition*, i.e., whether it not only attempts to assert a fact, but succeeds at doing so. Take, for instance, this statement:

*The earth is flat.*

On Theriault et al.'s intended interpretation of a statement being "about facts," you should *agree* that this statement is "about facts." This is because their measure is intended to assess whether the statement attempts to report a stance-independent fact, not whether it succeeds at doing so. Indeed, they *better* mean this. If they don't, then they are inappropriately conflating a measure of people's first-order moral standards with their metaethical standards. Thus, their only option is to intend for the statement to reflect the notion of *propositionality*, not a statement about whether a propositional claim is, in fact, true. Incidentally, then, their measure of whether the statement is "about facts," is at best only a measure of *cognitivism*, not a measure of realism. Cognitivism is consistent with subjectivism, so this item could not tell us whether participants are realists or subjectivists. Yet it is also consistent with error theory. A participant could consistently judge that all of the moral statements they were given are "about facts," but are *false*. Yet all this relies on the assumption that participants interpret the

question to be whether the statements are statements capable of being true *or false*. Asking whether they are statements "about facts" is ambiguous and misleading, and ordinary people could readily interpret this to be a question about whether the moral statements in question *are true*. Thus, their question conflates a metaethical question (are the statements propositional) with a normative question (is the moral assertion true or false?), and even if it were interpreted as intended, it couldn't possibly serve as a measure of realism, since agreeing with the item would only convey cognitivism, which is consistent with a variety of antirealist positions.

With this many methodological shortcomings, their measure of realism is not simply invalid, its invalidity is *overdetermined* by multiple, fatal problems, each independently rendering the measure invalid. Their realism question isn't their only measure, however. Could their study be rescued by appealing to one of the other measures that they used? Unfortunately, it cannot. Their question about whether the issue in question is "about morality," is irrelevant for our purposes, since it isn't intended as a measure of people's metaethical stances or commitments. But their final measure, the degree to which the statement is "about preferences," is. Regrettably, this question is just as flawed as their statement about facts. First, the use of "about" remains very strange. Consider, again, their example statement:

*It is irresponsible for airlines to risk the safety of their passengers*

What would it mean for this statement to be "about preferences"? Someone who asserted this statement could be asserting their preference. In which case, the statement would *be* a preference. But asking whether the statement is "about" a preference is ambiguous. Is this a question about whether the statement *is an expression of a preference*, or whether the statement *makes references to preferences* in the sentence *itself*? Consider, for instance, the following statement:

*Alex loves pineapple on her pizza.*

This is a statement *about* a preference. Yet it is not a subjective statement or a statement *of* preferences, which would be something more like "I like pineapple on pizza." It is a descriptive claim that is *stance-independently* true or false (unless you're an antirealist about descriptive claims of this kind). If a statement refers to or describes a preference, the statement would be "about preferences," without this in any way indicating that someone who agrees endorses subjectivism. Thus, there are two ways in which a sentence could be "about preferences." The sentence could refer to or describe preferences or it could express a preference. It's not clear, given their wording, which (if either) of these interpretations they intend.

Yet there is a far more serious problem with this item: moral realism does not entail that moral claims don't express preferences in addition to expressing facts. Suppose a moral realist were to assert:

*It is morally wrong to torture babies for fun.*

The realist will believe this is a statement about a fact, i.e., that it's a stance-independent fact that it's wrong to torture babies for fun. Yet such sentences are *also* typically used to pragmatically express the speaker's personal moral attitudes and preferences, i.e., that they prefer that people don't torture babies for fun, that they find it objectionable, disgusting, repugnant, etc. As such, agreeing that a moral statement is "about preferences" in no way suggests subjectivism or any other form of antirealism. Indeed, one serious problem with treating "preferences" and "facts" as distinct is that expressions of preferences can also be expressions of facts: facts *about* the preferences. As observed in Moss and Bush (2021), "statements about preferences can also express facts (e.g., 'I prefer classical to country music')" (p. 2). Even the reverse is true: "even straightforward factual statements, such as 'I know *King Lear* by heart' or 'I have never seen *Star Wars*' can reflect the speaker's tastes, opinions, or preferences" (p. 2).[143] This issue is made worse once we consider some of the other moral items Theriault et al. use. Consider these statements:

---

[143] I cite Moss and Bush (2021), but David Moss made these points and provided these excellent examples.

*Driving after drinking heavily is a stupid and selfish way to behave.*

*The deplorable conditions of Chinese electronics workers should not be ignored.*

*Sport fishing to kill and eat fish is barbaric and evil.*

Even if you are a moral realist, it would be absurd not to think these sentences *merely* convey a sober statement about what is morally right or wrong. They are most plausibly interpreted as *also* pragmatically conveying the speaker's subjective attitudes and preferences. This is achieved through the use of intense, emotionally charged evaluative language, e.g., "stupid," "deplorable," and "barbaric." Would *you* have any doubt that a person who made the last of these three claims *preferred that people don't engage in sport fishing*? Of course anyone who said this is expressing the preference that people don't engage in sport fishing. Theriault et al. use a variety of examples that pack in emotionally charged language that seems optimally designed to simultaneously convey both facts about the speaker's moral beliefs *and* the speaker's preferences. Other statements exemplify other flaws with their design. Consider this item:

*It is wrong to harm cockroaches just because humans find them disgusting.*

By stating that some people find roaches disgusting, this item explicitly describes both a fact and a preference simultaneously. This illustrates one of the ambiguities I referenced earlier: a statement could be "about preferences" because it expresses the speaker's preference, or because it describes preferences. By referring to their disgust at cockroaches, this item describes other people's preferences, and thus it is a statement "about preferences" even though this has nothing to do with subjectivism, or indeed about metaethics at all; it's just a descriptive claim. Finally, consider this item:

*Harry Potter should be banned from school libraries for idolizing witchcraft.*

Although this item does include a moral assertion ("Harry Potter should be banned from school libraries"), it *also* includes the nonmoral descriptive claim that "Harry Potter idolizes witchcraft." Regardless of your moral position on whether Harry Potter should be banned, this item *also* includes

a "statement about facts," that is, it asserts at least one fact: that Harry Potter idolizes witchcraft. It is, of course, irrelevant that this is *false* (see above). The problem is that the inclusion of this remark makes this a statement "about facts," but the fact in question is a nonmoral descriptive claim. Thus, even an antirealist could agree that this is a statement about facts. Consider, for instance, the following statement:

*We should throw idiots who think the earth is flat in jail because the earth is obviously round.*

Even if you're a moral antirealist who denies there are stance-independent moral facts, this is still a statement "about facts" because one of the facts it is about is the fact that the earth is round. It is not appropriate for an item asking people whether a sentence is "about facts" to be used as a measure of moral realism when the item in question includes both a moral claim *and* a nonmoral descriptive claim.

Like Rabb et al. (2020), Theriault et al. do not appreciate the role pragmatics play in our assertions: statements of preferences can express (or be "about") facts, and statements of fact can pragmatically convey preferences. In practice, any competent realist or subjectivist would recognize this, and regard claims like those used by Theriault et al. as expressions of both facts and preferences. Thus, unfortunately, level of agreement with these questions is simply not a valid measure of whether a person endorses realism or subjectivism.

**S3.8 Davis's (2021) flowchart method**

First, this makes assessment of metaethical stance towards individual items far more onerous. Each moral issue requires multiple potential questions. This increases the length and complexity of the task, which limits one's ability to present participants with a variety of moral issues without paying the price: such studies cost more, incentivizing researchers to include fewer participants and reducing power, and such studies risk tiring or boring participants. These aren't insurmountable difficulties, though. One could always run more and larger studies with adequate resources.

A second, more serious problem is that the central methodological problem with research on folk metaethics is that participants struggle to interpret questions as intended. The inclusion of so many additional subcategorization tasks introduces numerous additional ways that participants could interpret stimuli in unintended ways. Again, this isn't insurmountable, but it provides many additional paths to high levels of interpretative variation (participants interpreting stimuli in ways that differ from one another) and low rates of intended interpretations.

A fifth problem with this approach is that participants still face a number of forced choices. Such studies *already* rely on forcing participants to respond to a set of categories and distinctions that concern academic philosophers. A more fine-grained approach risks foisting even more forced choices onto participants that don't necessarily reflect their positions. For instance, even if someone rejects realism, that doesn't mean they have a determinate antirealist stance. One could have determinate stances or commitments at a general level, but no determinate stance or commitment at a lower level of specificity. For instance, many people may believe in God, but have no determinate stance about whether God's omniscience is incompatible with free will. Thus, even if participants have a determinate metaethical stance or commitment with respect to realism or antirealism, it's unclear whether they'd also have a determinate stance with respect to the subcategories on offer. Is it plausible that ordinary people will readily understand the distinction between naturalism and non-naturalism, and that they have a determinate stance on the matter, or speak in a way that commitments them to a naturalist or non-naturalist metaphysics? I'm already arguing for skepticism that ordinary people are realists *at all*; it's even less plausible that they'd have a determinate stance or commitment on more fine-grained subcategories. For comparison, even if most people preferred red or white wine over the

other[144], it would not follow that they generally have a well-developed and determinate view on the merits of a Pinot Noir over a Malbec, and it would be unsurprising if they couldn't distinguish them.

A fourth problem with this approach is that using auxiliary or follow-up questions can introduce novel methodological problems. For instance, Davis first asks participants whether they believe in God. However, previous studies have found that religious priming increases realism (Yilmaz & Bahçekapili, 2015a). Asking such a question could have inappropriately increased realist responses. Researchers may be able to mitigate this by varying the order of questions and testing for order effects, but this may be difficult or impossible if the content of some questions is contingent on the content of previous questions.

A fifth and final problem with these studies is that they rely on the specific categorization schemes proposed by researchers. Such categorization schemes may be controversial even among professionals in the field. Having participants work through a kind of flowchart with follow-up questions will require each researcher to make decisions about the way in which metaethical subcategories are nested within broader categories. That researchers may disagree is not idle speculation. Davis includes Divine Command Theory (DCT) among the realist response options, and treats it as distinct from naturalism and non-naturalism, yet I don't think it is distinct from non-naturalism and, more importantly, yet *I don't consider DCT a form of realism*. There may be *versions* of DCT that are a legitimate form of realism, there may also be versions that aren't. Insofar as a proponent of DCT regards moral facts as stance-dependent, and stance-*in*dependence is *the* defining characteristic of a realist metaethical position, such versions of DCT simply aren't forms of realism. So is DCT a form of realism? I'm not sure. I suspect that it can be construed in ways that are consistent

---

144 Ballester et al. (2009) found that ordinary people are capable of distinguishing red and white wines, but both experts and novices struggled with rosés. I can't think of any obvious implications for metaethics, but at the risk of offending sommeliers, I suspect distinctions in metaethics are at least as tricky as the difference between a rosé and a red or white wine. If trained experts can't even distinguish basic categories of wine, perhaps in an odd way this should bolster our suspicion about the ability of non-experts to distinguish obscure metaethical positions.

with both realism and antirealism. I'm also not sure that "supernaturalism" isn't just a type of non-naturalism. More generally, a variety of metaethical positions cross-cut various other distinctions, with subsets of particular positions appearing in multiple places on any supposed "flowchart." Indeed, the very notion of a flowchart is questionable. Where do we put constructivists? Are *all* relativists antirealists? Is stance-independence even the appropriate dividing line? Some metaethicists insist there are forms of objectivism that don't entail full-blown realism. Others want to add epistemic conditions to their accounts of realism. It's typical to include semantic theses in one's description of a metaethical position, even though others deny that semantic theses are necessary (Kahane, 2013). There is no consensus on how to characterize grounding, whether grounding is necessary, which forms are, if so, or how many forms there are, or how best to characterize various types of grounding. One of the most revered moral realists, Parfit, even denied realism requires any metaphysical claims, as does Scanlon, another prominent realist. Where do we put *them* on our flowcharts? Are their views of realism illegitimate on analytic grounds? Can we presume that ordinary people couldn't possibly endorse realism without a metaphysical element?

I don't expect answers to any of these questions. A dozen or more dissertations couldn't resolve the matter. A dozen *careers* couldn't resolve these questions adequately. My point is simply that the idea of a genuinely accurate, uncontroversial, theoretically neutral flowchart is impossible. Flowcharts are, at best, a convenient oversimplification for the sake of introducing undergraduates to the topic, *not* a way of actually characterizing the metaethical landscape. It would take a much more sophisticated nonlinear set of questions to even begin to clarify what a particular person's metaethical positions are. It cannot be done with a handful of simple questions.

**S3.9 Zijlstra's (2021) implicit measures**

Zijlstra (2021) offers a novel and interesting way of assessing whether ordinary people are moral realists. Zijlstra claims that there are two ways to be a moral realist: explicitly, and implicitly. These

ways of being a realist incidentally correspond (if imperfectly) to my distinction between stances and commitments, respectively. Zijlstra appeals to a number of researchers who claim that even if ordinary people did not explicitly endorse moral realism, that they may speak or think in ways that commit them realism anyway, and that *this is in fact what is most relevant*. For instance, Björnsson (2012) states that "the primary task of metaethical theories is to account for this *engaged* behavior, rather than for what is in effect lay people's theoretical interpretations of it" (p. 9, as quoted in Zijlstra, p. 4). Likewise, Enoch (2005) claims that

> "[W]hat is relevant is not the explicit metanormative beliefs – much less the explicit metanormative statements – of participants in normative discourse. What is relevant, rather, are the deep metanormative commitments embedded (perhaps implicitly) in normative discourse and practice themselves. (p. 773, footnote 31, as quoted in Zijlstra, 2021, p. 4)

Yet Zijlstra's primary inspiration seems to come from another article from Enoch (2014). Enoch devised a series of three intuition pumps that he is confident most people would respond to in a way that suggests they are implicitly committed to realism. Zijlstra transformed these intuition pumps into a set of three studies, each designed to assess whether ordinary people are implicitly committed to realism.

One immediate cause for worry is that of the 150 participants recruited for the study, 53 either failed to complete the study or failed one of two comprehension checks were excluded from analysis (though we're not told the proportion excluded for failing the comprehension checks in particular). At 35.3% this is an extraordinary number of people to fail, far exceeding the 10% cutoff for exclusion due to inattention suggested by Bergenholtz, Busch, and Praëm (2021) or the 25% cutoff for failure to comprehend suggested by van 't Veer and Giner-Sorolla (2016). Before we've even looked at the results, we're already dealing with a potentially self-selected sample of participants who don't reflect the population they were drawn from, threatening the internal validity of Zijlstra's findings, though

Zijlstra maintains that results were no differences between participants excluded from analysis. Setting this concern aside, the three tests are:

1. The joke test
2. The phenomenology of disagreement test
3. The counterfactual test

### 3.9.1 The joke test

The first test is based on the idea that jokes can reveal that we regard some normative domains in realist terms, but others as merely subjective. Enoch presents a joke:

> A child hates spinach. He then responds that he's glad he hates Spinach. To the question "Why?" he responds: "Because if I liked it, I would have eaten it; and it's yucky!" (Enoch, 2014, p. 193)

This is allegedly funny because the child fails to realize that taste preferences are subjective. By mistakenly thinking of taste preferences as stance-independent, the child has said something amusing. Yet this wouldn't work for a normative domain that we do regard in realist terms, because the juxtaposition between speaking of something we're implicitly antirealists about in explicitly realist terms would no longer be present, we'd just be speaking about something we're implicitly realists about in explicitly realist terms. This first study assesses whether people treat moral claims like the joke above.

Participants were presented with three Enoch-style jokes: a taste joke, a factual joke, and a moral joke, each adapted from Enoch (2014).

**Taste condition**
A child hates spinach. He then responds that he's glad he hates Spinach. To the question "Why?" he responds: "Because if I liked it, I would have eaten it; and it's yucky!"

**Factual condition**
Consider, for instance, someone who grew up in the twentieth century West, and who believes that the earth revolves around the sun. Also, she reports to be happy that she wasn't born in

the Middle Ages, "because had I grown up in the Middle Ages, I would have believed that the earth is in the center of the universe, and that belief is false!"

**Moral condition**
Suppose someone grew up in the US in the late twentieth century, and rejects any form of racism as morally wrong. He then reports that he's happy that that's when and where he grew up, because "had I grown up in the 18th century, I would have accepted slavery and racism. And these things are wrong!"[145]

For each condition, participants were asked:

Can the above story be regarded as a joke? [Yes/No]

To what extent do you think the above story is funny? [0-100]

Zijlstra found that the proportion who judged each condition could be regarded as a joke was significantly different from the other conditions: 63% of participants judged that the taste scenario could be regarded as a joke, while 32% judged that the factual condition could be regarded as a joke, and 8% judged that the moral condition could be regarded as a joke. The median score for how funny each joke was 20 for the taste condition, 9 for the factual condition, and 1 for the moral condition.

Zijlstra interprets these findings as support for the notion that ordinary people are implicit realists. Yet there are numerous problems with this conclusion. *None* of these scenarios were especially funny; the spinach joke was far below the midpoint, while the factual and moral conditions were near or at the floor of the measure, respectively. Strictly speaking, it's unlikely *any* of these scenarios, jokes or otherwise, are actually funny. This takes a bit of the wind out of the sails for these conditions, but Zijlstra or Enoch could still insist that the taste condition at least did a little better, and that counts for something. They could also point out that the proportion who thought the taste condition could be regarded as a joke was so much higher than the moral condition that this, at least, lends weight to Enoch's claim.

---

[145] Zijlstra does not explicitly state whether the examples he presents from Enoch were the actual stimuli that were used, but these appear to be the conditions given to participants.

Unfortunately, it's far from clear that it does. The central issue with these conditions is that what matters is *why* more people think the taste condition is a better candidate for a joke than the other conditions. It *could* be due to a sensitivity to mistakenly treating subjective values as objective, but without direct evidence that this is *why* participants judged this condition to be more of a joke than the other conditions, we cannot know whether an implicit commitment to moral realism can explain the difference across conditions. This points to a more general issue with these conditions: there are multiple features of each condition that could contribute to the degree to which the conditions are regarded as jokes *that have nothing to do with realism/subjectivism*. One scenario involves a child's reaction, while the others don't mention children. The factual and moral condition involves a counterfactual that places a person in the distant past. They are significantly longer and more complicated, as well. Yet the more critical issue is that the substantive content of the different scenarios could contribute to differences in the degree to which people regard them as potential jokes, and we simply don't know how varying the content (from taste, to factual, to moral claims) may have influenced the degree to which the scenarios were regarded as potential jokes, and how funny they might be. Note, as well, that participants are simply asked to judge whether these scenarios could be regarded as a joke. But ordinary people may judge that or not something could be a joke isn't merely due to its structure, but to whether or not it's actually funny. In other words, whether a scenario is a joke or not *may not be independent from how funny it is*. Children disliking green vegetables is a common trope, if a rather trite one. Yet being wrong about the earth being the center of the earth may simply not be very funny. Likewise, it seems plausible many people don't think we should make jokes about slavery or racism. People may think e.g., "it's not appropriate to make jokes about that…" where they have in mind atrocities or taboo topics. If so, they may have regarded the scenario about racism and slavery as less funny and as less plausible a candidate for a joke simply because a lighthearted quip about slavery and racism isn't funny. In other words, people's normative and evaluative attitudes about the content of the scenarios could

influence the degree to which they consider the scenarios funny, which could in turn influence the degree to which they regarded the scenarios as jokes.

In addition, note that participants are given these scenarios stripped of any social context that would normally be relevant to judging whether these scenarios are intended as jokes. If all three scenarios were delivered by a stand-up comedian on stage, there may be little dispute about whether they "could be regarded as jokes." So what exactly does it mean to ask whether they *could* be regarded as jokes? Whether something is a joke or not is, presumably, contingent on the intentions of the person presenting the scenario (i.e., whether they are intending to deliver a joke). Without such context, ordinary people engage in some unknown processing of these scenarios before rendering a judgment. Perhaps they imagine how plausible it would be, in ordinary circumstances, that each of these scenarios would be presented with the intention of conveying a joke. This would be a sensible way to respond. Yet if so, the taste condition may simply be a more plausible candidate for a joke than the second two. Take the moral scenario: people's moral values, especially opposition to slavery and racism, are highly central to their identities. We also know that our moral standards are highly historically contingent, and that had we lived in the past, we could have had terrible moral values. There are plenty of non-humorous circumstances in which a person would make a point of expressing their opposition to the standards and attitudes of people in the past. There are no similar reputational benefits to signaling disgust with spinach.

The percentages Zijlstra reports are also underwhelming. 63% is hardly a resounding proportion in favor of the taste condition being a joke: many people didn't see it as a joke. Strangely, 32% also regarded the factual condition as a joke. *Why?* Enoch's account does not explain why factual conditions would be intermediate between taste and moral conditions. One might instead expect a far more discrete division, with a strong majority regarding the taste scenario as a joke, and almost nobody regarding the fact and moral conditions as jokes. Yet this isn't what I found. Results are consistent

with nearly half of participants regarding the taste condition as not being a joke, which leaves ample room for many participants to exhibit an implicit commitment to realism.

We may also recognize a certain *arbitrariness* to taste preferences that isn't available for factual or moral issues. That is, we may not closely identify with our taste preferences, recognizing that they could have been different, and that this would make little or no practical difference. If we actually did like spinach, then it wouldn't be "yucky" to us, and *this wouldn't matter*. We'd be the same person, just with different preferences. And it may be especially appealing to imagine alternate versions of ourselves with preferences we find unappealing. For instance, versions of ourselves who unironically enjoy Nickelback's music. Conversely, there may be nothing amusing about imaging versions of ourselves that torture people or commit atrocities. Such a person isn't someone we'd want to be, and in some ways may not even be us.

Another concern with these findings is the possibility that *ordinary people are bad at counterfactual reasoning*. It may be fairly easy to imagine versions of ourselves with different food preferences. But how easy is it to adopt the point of view of someone with factually mistaken beliefs about basic scientific facts, or who has moral views we consider completely repugnant? There's no harm in imagining versions of ourselves who like spinach, but it may be difficult to imagine a version of ourselves who thinks slavery or racism is morally acceptable.

Finally, participants may have simply judged that the moral scenario wasn't funny due to concerns about social desirability (Grimm, 2010).[146] People may have wanted to avoid judging a scenario describing slavery and racism as a "joke," or regarding it as funny, even if they would, in private, find jokes about slavery or racism to be funny, and would have regarded a scenario involving

---

[146] Zijlstra (2021) acknowledges this, stating that:

"It is possible, however, that the significant difference in responses between the factual and moral stories is explained by the fact that people tend to provide a socially desirable response. That is, perhaps some people found that moral issues are not a laughing matter and that therefore they judged it as even less funny than the factual story." (p. 9)

racism or slavery as a joke (since, as I've argued, whether something is funny could influence whether it's regarded as a joke). Racism is one of the least desirable traits, while slavery is almost universally regarded as one of the greatest human rights abuses conceivable. Judging such a scenario to be a joke and to regard it as funny could signal an undesirable lack of opposition to racism and slavery.

Finally, note that Zijlstra only appeals to a single moral issue. We don't know how well these scenarios would perform if the substantive content of the taste, factual, and moral conditions were changed. Would we get more or less the same results? Without knowing, it's hard to know what factors may have contributed to people's judgments. And that brings us back to the central problem with this study: *we don't know why people expressed the judgments that they did*. Without independent evidence or good reasons to think that Enoch's claim about why the taste condition is a joke and the moral condition isn't, we are far from having substantive evidence of an implicit commitment to realism.

### 3.9.2 The phenomenology of disagreement test

The second test was based on the phenomenology of moral disagreement. This test concerns whether, when we think about what it feels like to experience a moral disagreement, whether it feels more like a disagreement over stance-independent factual matters, or more like our preferences. To test for this, participants were asked to either "think about a moral disagreement about abortion or a disagreement about a different moral issue they felt strongly about" (Zijlstra, 2021, p. 7). Then they were asked whether this disagreement feels more like a dispute about whether dark chocolate or milk chocolate tastes better, or whether it feels more like a dispute about whether human actions contribute to global warming. This was the wording Zijlstra provides in Appendix 2:

> In this part of the study, we will consider what it feels like for you to engage in a disagreement.
> Now, think of some serious moral disagreement. For example, about the moral status of abortion. Suppose that you are engaged in such a disagreement. Imagine this, as it were, from the inside. You are in this disagreement yourself. Perhaps you think that there is nothing wrong with abortion, and you are arguing with someone who thinks that abortion is morally wrong. Or, perhaps you think that abortion is morally wrong and you are arguing with someone who thinks that there is nothing wrong with it.

Please explain how it feels for you to engage in this kind of disagreement. Please note that there is no correct answer to this question: We would simply like to know how it feels for you to engage in moral disagreements. In particular, please tell us whether it feels more like disagreeing over which chocolate is better, or like disagreeing over objective facts like whether human actions contribute to global warming or not?

[1] It feels more like disagreeing over which chocolate tastes better.

[2] It feels more like disagreeing over whether human actions contribute to global warming.

Other.

Zijlstra found that 77.5% of participants regarded moral disagreements as more like factual disagreements, while 22.5% judged that they felt more like matters of taste (4 people chose "other", but were not included in the final percentage).

22.5% isn't nothing. That's nearly a quarter of the participants. Once again, Zijlstra did not find an overwhelming indication of an implicit realism, just a majority. Yet Zijlstra states that "the second test suggests that the feel of moral disagreements is more about getting an objective fact right than about stating one's own preferences" (p. 9). This is unclear. This *could* mean that a higher proportion of people treat moral claims as stance-independently true, rather than as a matter of preference, but it's not clear that this is how people would interpret such a claim. But it could also mean that moral disagreements are generally about both stance-independent facts and about preferences, but they're just more about the former. It's odd when researchers draw generalized conclusions like this that are hard to interpret in a straightforward way.

Yet this study also has methodological problems. Like the joke test, we don't know *why* most people judge that moral disagreements are more like disagreements over whether people contribute to global warming than disagreements over which kind of chocolate tastes better. One reason *could* be that they regard both moral claims and scientific claims as stance-independent, but claims about chocolate as stance-dependent. Yet why presume that this is the only, or primary factor accounting for the response pattern we observe? There's no direct evidence that it is! What matters is not what

proportion of people respond in any particular way, but *why they do so*. And absent evidence that implicit realism is the best explanation for these results, all we're left with is a pattern that *could* be explained in this way, but no evidence that it actually is. Of course, it may still be the best explanation, but is it?

I don't think so. First, participants are asked to compare *abortion* and *climate change*. Both are highly politically charged disagreements, and one might regard them as more similar for that reason alone. By comparison, disagreements about dark and milk chocolate are not, to my knowledge, politically charged.

Second, and more importantly, climate change, abortion, and other serious moral issues have *enormous practical significance*, whereas chocolate preferences don't. In some ways, my opposition to torture feels more like my opposition to policies that would cause environmental catastrophes than it feels like my personal preference for peach tea, but this does not indicate an implicit commitment to realism, but because opposing torture and opposing bad social policies are both *really important to me*. My taste preferences aren't. Why did Zijlstra use climate change as an example, given how our attitudes towards climate change are deeply bound up not only in our political concerns, but our practical concerns with the environment and the future of humanity?

When we think about how we feel about abortion, climate change, and our chocolate preferences, there is more involved in these feelings than whether they feel stance-independently true or not. They can feel *important*, or *emotionally charged*, or *central to our identity*, or *practically significant*, and so on. And these comparisons can link disagreements in two domains for reasons other than a shared, implicit regard for both as stance-independent facts. More generally, without additional information, we don't know what it is people are feeling when they respond to these questions, so why should we presume that involves feeling like moral facts are stance-independent?

In addition, there are elements of moral disagreements that are similar to factual disagreements, but dissimilar to disagreements over taste, which don't entail or require a commitment

to realism. Recall that participants are asked to imagine that they are *engaging* in a disagreement. That is, they aren't merely aware that someone else has a different view about morals, facts, or taste, but they are actually *arguing* with someone who disagrees. Think about what a factual dispute looks like: people may point to data, appeal to various alleged facts and findings, and so on, in an effort to persuade the other person. What does a dispute about chocolate look like? I'm not sure. People probably don't engage in many disputes about which is better. At least far fewer than they engage in disputes about matters of scientific fact. Yet when such disputes do occur, the substantive features of the exchange are likely to be quite different from disputes about scientific matters. For instance, there are probably few non-normative facts under such circumstances. It's not likely, for instance, that a proponent of dark chocolate would point out that "dark chocolate activates 61.2% more taste receptors," or that it "releases more dopamine." They *could*, but a prototypical taste dispute is unlikely to involve the same kinds of appeals as factual disputes

Now consider moral disagreements. Such disagreements often *do* appeal to nonmoral factual considerations, such as the consequences of outlawing abortion, or the risk to a mother's life, when life begins, the details of abortion procedures, the psychological consequences of abortions, the economic costs of restricting abortion access, and so on. And while people do make appeals to normative moral considerations, such as a right to bodily autonomy, a right to life, and so on, these are interspersed among a wide variety of non-normative considerations as well. Such considerations are relevant *even for antirealists, including moral subjectivists.*

This brings me to what is the most serious flaw with this study: an antirealist can regard moral disagreements as more similar to factual disagreements than disagreements about taste. Unless participants want to choose "other," and few people seem motivated to go for such options, participants have to choose between a response option interpreted as realism, or a response option that suggests a kind of crude subjectivism. Yet this isn't the only option available to antirealists.

Speaking for myself, moral arguments don't feel like disputes about science or preferences. They feel like negotiation or diplomacy: I recognize others have goals that conflict with my own, and I seek to navigate interactions so as to form compromises, alliances, and agreements that optimize for achieving my goals. While this is similar in some ways to a dispute about preferences, what I'm *not* doing is arguing about which of our preferences is "correct." It's unclear, in the case of a disagreement about chocolate preferences, whether any practical solution is to come of such a dispute. If I prefer dark chocolate and someone else prefers milk chocolate, what are we trying to achieve by arguing? It's not plausible I'd be really invested in convincing someone else to eat more dark chocolate. What would be the point? Yet despite rejecting moral realism, I am very much invested in other people acting in accordance with my moral standards. And this is one of the key differences that undermines the appropriateness of Enoch's (and by extension, Zijlstra's), comparison: taste preferences are preferences about self-regarding actions. Even if a moral antirealist regards their moral standards as "preferences," in one respect—namely, they don't think of their moral standards as stance-independently true—that does *not* mean they regard them as merely *self-regarding* preferences.

One of the key features of our taste preferences is that, for the most part, they are only about us. I prefer dark chocolate over milk chocolate. But I don't care what chocolate preferences other people have. I don't support laws that ban milk chocolate or punish people who eat it. Yet take my preference that people don't torture babies. It's not simply that I prefer not to torture babies. It's that I *very strongly* prefer that *nobody* torture babies. I do support laws against baby torture, and want people who torture babies to be punished. My preferences are thus *other-regarding*. Thus, while I'm a moral antirealist, my moral disagreements aren't anything like disagreements about chocolate. Such disagreements are frivolous, trivial, and have no practical significance. Disagreements about moral matters, on the other hand, are not only practically significant, *they are the most practically significant concerns of all*, almost by definition! As such, they're not at all like disagreements about chocolate.

This illustrates a key weakness in Zijlstra's study: participants are given a choice between a response interpreted as realism and an extremely unattractive antirealist response that doesn't even reflect how antirealists are necessarily inclined to think about moral disagreements. Someone with an implicit commitment to antirealism may judge moral disagreements to be more like factual disputes than taste disputes not because they're implicitly committed to realism, but simply because it's the least bad of the two options. That is, it's not that they accept realism, but that they reject what amounts to an especially crude form of antirealism on offer. In short, participants are given an option between responses that amount to realism or *a very narrow and specific conception of antirealism*.

Not only do I personally reject the notion that moral disagreements are like taste disputes, cultural relativists would as well, since they don't think moral standards are reducible to matters of individual taste. Likewise, constructivists are antirealists, yet they think that moral theories are devised in accordance with particular decision procedures. They might think that e.g., our moral standards are those that an ideally rational and fully informed version of ourselves would endorse, or they endorse that set of moral standards we'd collectively agree to under conditions in which we were unaware of our place in society (Bagnoli, 2021; Milo, 1995). Regardless of what procedure they favor, constructivists may regard moral disagreements in a way that is utterly unlike disagreements over matters of taste, and instead bears many of the hallmarks of a dispute about facts, without this entailing stance-independence about moral values. Finally, error theorists regard moral disagreements as disputes about factual issues, they just think everyone engaged in such disputes is mistaken. What all these examples illustrate is that there are many forms of antirealism that are consistent with rejecting the notion that moral disagreements are like trivial disagreements over taste preferences. It's unclear what position these people should favor, yet in many ways judging moral disagreements to be more similar to factual disputes is a sensible approach, even though they're not realists.

Note that antirealists can and do engage in moral argument for persuasive purposes, as well. It is possible for a moral antirealist to convince someone that an action is wrong, or change their mind about a moral issue, by pointing to nonmoral facts, drawing attention to the consistency between some policy, principle, or action and the values of the person they're arguing with, or causing that person to reflect on their own values and change them. It's unclear how feasible this is for taste preferences, nor is it clear why anyone would be motivated to persuade someone to change their chocolate preferences.

Tellingly, Zijlstra also constructed a "low stakes" condition, where participants were asked to judge whether a moral disagreement feels more like a taste disagreement (again, over dark or milk chocolate) or a disagreement about flight times between NY and LA. Zijlstra found that 71.9% of people judged that moral disagreements felt more like taste disagreements. This appears inconsistent with what Enoch would predict, and suggests irrelevant factors drove people to select the "realist" response in the main version of the study.

Zijlstra devised two more conditions. One condition swapped out disagreement over chocolate with a disagreement over the taste of organic food for the taste response option, but kept the original global warming option for the factual response option. This time, 54.3% judged that moral disagreements felt more like a matter of taste. Finally, Zijlstra swapped out both response options, asking people whether moral disagreements felt more like disagreements about the taste of organic food or about flight times. Yet again, 72.3% felt moral disagreements were more like disagreements about taste.

These results reveal deep inconsistencies that raise substantial worries about the legitimacy of this paradigm, and, at the very least, indicate that there is little reliable evidence that the phenomenology test favors implicit folk realism.

### 3.9.3 The counterfactual test

Lastly, Enoch appeals to the notion of counterfactuals, asking, "Had our beliefs and practices been very different, would it still have been true that so-and-so?" (Enoch, 2014, p. 197, as quoted in Zijlstra, 2021, p. 6). Zijlstra provides an example of smoking: suppose we believed smoking was harmless and didn't cause cancer, and due to this belief, we never prohibited cigarettes. Would it therefore follow that smoking is, in fact, harmless? Obviously not. Likewise, Enoch proposes that we would feel the same way about moral claims. Even if historical events had unfolded differently, such that slavery were considered morally permissible, would we think that under such circumstances slavery would, in fact, be permissible? Enoch suspects that most people would respond, when confronted with such cases, with a resounding no. If so, he takes this to be an indication of an implicit commitment to realism. The last measure involved reading a short story adapted from Enoch's smoking example, which appears in Appendix 3:

> As a result of years of scientific research we now know that smoking causes cancer.
> Now, had our relevant beliefs and practices regarding smoking been different—had we been ok with it, had we not banned it, had we thought smoking was actually quite harmless—would it still have been true that smoking causes cancer? It is probably uncontroversial that the answer is "Yes". The effects of smoking on our health do not depend on our beliefs and practices. Rather, it is an objective matter of fact.
>     The question that we therefore ask here is "Had our beliefs and practices been very different, would it still have been true that so-and-so?".
>     Let us apply this question to morality. For example, some people believe that gender-based discrimination is wrong. Maybe you also believe that it is morally wrong or maybe you do not. If you do not, imagine something else that you think is morally wrong. Would it still have been wrong had our relevant beliefs and practices been different?
>     [1] No, had our relevant beliefs and practices been different than it would not be wrong.
>     [2] Yes, had our relevant beliefs and practices been different than it would still be wrong.

70% chose the second response, judging that gender discrimination would still be immoral even if our beliefs and practices were different, while 30% judged that gender discrimination wouldn't be wrong.

This is not an impressive win for realism. 30% amounts to nearly a third of the participants, and this result was obtained in spite of using an item that is likely to spark extremely high levels of social desirability bias in favor of the "realist" response. Like previous items, note that it may be difficult or impossible for ordinary people to fully embrace the scenarios they're presented with, such that they can disentangle their current normative moral standards from their assessment of these situations. This can result in significant performance errors that reliably bias response options towards the "realist" response options.

Yet this scenario also suffers from unclear instructions. Participants are asked whether gender discrimination would still be wrong if "our relevant beliefs and practices been different." What does that mean? Relevant to *what*? Presumably, Zijlstra wants participants to think about whether gender discrimination would be morally wrong in a counterfactual scenario in which they thought gender discrimination was acceptable, and practiced gender discrimination. First, there's an unusual vacillation between what the participant thinks about the moral issue, and what people in general think. The participant is told to think about something they believe is morally wrong. Note the subtle bias here: the instructions presume cognitivism. But let's set that aside. The issue is that participants are asked about what *they* think is wrong, but are then asked whether it would still be wrong had *our* relevant beliefs and practices been different. Who is "our"? Does it include the participant or not?

Now consider how an antirealist might respond. I'm a moral antirealist. Do you think I would think something was morally acceptable simply because other people thought that it was acceptable, or acted as though it were? Suppose I am asked to imagine that history had gone very differently, and people in the United States engaged in ritual human sacrifice, and thought they were morally required to do so. Given my *current* moral standards, preferences, and so on, I would still think ritual human sacrifice was wrong, and were I transported to this society, I would continue to think human sacrifice is wrong. This is because my moral standards aren't based on what other people think under counterfactual

scenarios, nor even on whatever it is people in the actual society I am in currently think. My moral

standards are also not based on what some counterfactual version of myself would think. That is, I

don't think ritual human sacrifice would be morally good if some other version of me thought it was

morally good.

One serious problem with Zijlstra's question is participants are forced to choose between

realism and relativism, as though those were the only metaethical positions. This leaves anyone with

antirealist inclinations that isn't disposed towards relativism without a response option that would

reflect their views, including error theorists, noncognitivists, constructivists, individual subjectivists,

appraiser relativists, and quietists like myself, which could compromise the majority of potential

antirealist respondents.

However, another serious issue is that even if you endorsed relativism, the question would be

unanswerable, because it would be ambiguous. In order for a moral relativist to judge whether an

action is morally right or wrong, they need to know which moral standard is being indexed. Both of

these response options require participants to judge whether an action would be right or wrong under

counterfactual conditions in which "our relevant beliefs and practices" were different than they are.

The problem is: would they be right or wrong *for us* or right or wrong *for the people whose relevant beliefs*

*and practices were different*? If you're an appraiser relativist, then whether an action is morally right or

wrong depends on who the appraiser is supposed to be. Yet it's not clear from the question or

response options whether we're being asked whether *we* (the non-counterfactual version of ourselves

responding to the question) think that an action that we think is morally wrong would no longer be

wrong *relative to our current moral standards* if everyone else thought it was wrong, or whether the action

would be wrong *relative to the counterfactual version of ourselves in the world in which our beliefs and practices were*

*different*. In other words, each participant has their actual moral standard, which we can call *standards*$_{actual}$

and then there are the standards of the people in a counterfactual scenario in which people had

different beliefs and practices, which we can call *standards$_{counterfactual}$*. For an appraiser relativist, if they are asked whether an action is morally right or wrong, there is no fact of the matter *simpliciter*. The action could be wrong relative to their moral standards, *standards$_{actual}$* but not morally wrong relative to some hypothetical set of standards they reject, e.g., *standards$_{counterfactual}$*. The problem with Zijlstra's question is: *which of these standards is being indexed by the term "wrong"?* It's not clear.

Furthermore, this ambiguity could also prompt normative entanglement. If you are an appraiser relativist, you think that actions are only right or wrong relative to the standards of different appraisers. If you are asked whether an action would be wrong *if* people's standards were different, are you being asked whether the action would be morally wrong relative to the moral standards of those hypothetical people who had different beliefs and practices, or relative to your current moral standards? If the latter, then the answer would be "no," since whether an action is right or wrong relative to your current moral standards has nothing to do with other people's beliefs and practices. However, an appraiser relativist resolves the ambiguity in this way, then they would choose the "realist" response option, which would lead to their response being miscategorized as an endorsement of realism. If, on the other hand, the question is asking whether an action would be wrong relative to the standards of hypothetical people who considered the action in question wrong, then the answer would be a trivial yes, since for an appraiser relativist, whether an action is wrong relative to a particular moral standard *just is* to say whether it is wrong *according* to that standard. On the latter interpretation, the appraiser relativist would choose the antirealist option, and would therefore be correctly categorized as an antirealist. However, if you were disposed towards appraiser relativism, it may seem odd to be asked a question where the response options are trivially true or trivially false. In this case, the question would amount to asking something like "If our beliefs were different, would our beliefs be the same, or different?" While the answer would obviously be "different," that so obvious a

question is being asked may prompt people to consider whether they'd interpreted it as intended, which could prompt confusion or uncertainty.

I'll try to elaborate on this problem to drive home the point. Once we take the appraiser relativist's perspective, we can see why the question Zijlstra poses is incredibly bizarre. The appraiser relativist is asked to imagine an action *they* consider wrong. They are then asked whether, had "our" practices been different, such that "we" thought whatever the appraiser relativist currently think currently think is wrong relative to their *standards_actual* wasn't wrong relative to the standards of the people in this hypothetical situation, whether the action in question would be "wrong." To an appraiser relativist, if the question is about their own moral standards, then presenting them with a counterfactual makes no sense, since *standards_actual* don't change in accordance with counterfactual considerations. If the question is instead asking whether a counterfactual version of themselves would have different moral standards, the question is trivial, since it would amount to asking, "if your moral standards were different than they are, would they be different than they are?" to which the answer is, of course, "yes." The problem is that *if* the appraiser relativist resolved the ambiguity in the former sense, they'd give the "realist," response option, and be miscategorized by researchers as a "realist."

If, instead, they chose the antirealist response option, this could pragmatically imply that they are expressing that the action that is wrong according to their *standards_actual* wouldn't be wrong relative to their actual standards. That is, they'd be implying they hold the moral stance that *if* people think an action that is wrong relative to their current standards isn't wrong, then it isn't wrong *relative to their current standards*. Yet that would express *agent* relativism, *not* appraiser relativism. And an appraiser relativist may not wish to endorse a response that could be conflated with agent relativism, since agent relativism *does* result in the agent relativist judging that if someone else thinks that e.g., baby torture is morally permissible, then *it is permissible* for that person to torture babies. An appraiser relativist may not only disagree with this position, but find it highly objectionable, and not wish to have their own

stance be conflated with an agent relativist's stance. They may also not know whether their response option would be interpreted as an expression of appraiser or agent relativism. For instance, suppose a professional philosopher who understood the distinction between appraiser and agent relativism were responding to this question. How should they respond? *It's not clear.*

I don't see how to make much sense of the response options from an appraiser relativist perspective. However, it makes much more sense from an agent relativist perspective. To an agent relativist, *if* people think an action is morally permissible, then it is morally permissible *for those people* even if it would still be wrong for the agent relativist themselves to perform the action in question. An agent relativist may have little trouble with this question: if "our" beliefs and practices were different, such that we thought that a particular action the participant thinks is wrong isn't wrong, then it wouldn't be wrong (according to the agent relativist) *for those people* to perform the action. However, agent relativism is a highly specific form of relativism, and it's unclear whether or not participants would distinguish agent and appraiser relativism. It's also unclear what proportion of relativists endorse agent relativism, so even if the question and response options made more sense for an agent rather than appraiser relativist, this would mean that the question would make more sense only for a (possibly very small) subset of relativists. An agent relativist may have less trouble with the ambiguity about which moral standard is being indexed: it would make sense to suppose that they are being asked whether, if people had different moral beliefs and practices, whether it would be morally wrong *for them* to perform the actions in question. On the other hand, there is still some ambiguity here: even if the agent relativist thinks that if it would be morally permissible for members of a hypothetical society with different beliefs and practices to engage in actions that members of that society approve of, it doesn't follow that the agent relativist thinks that it's permissible for the agent relativist *themselves* to perform the action. Thus, the ambiguity remains: if some hypothetical society had different beliefs and practices, and considered something the agent relativist regards as wrong

(e.g., baby torture) to be morally permissible, it's still unclear what it means to ask whether the action in question would be "wrong." Wrong *for who*? For an agent relativist to whether an action is "wrong" or not depends on the standards of the agents performing the action. In short: even for an agent relativist, the question of whether an action would be wrong if our beliefs and practices were different simply doesn't make sense: to an appraiser relativist, it wouldn't be wrong *for those agents* to perform the action, but is that what this question is asking? Or is it asking whether it would be wrong for the agent relativist to perform the action? It's not clear. I suspect it's not clear because I suspect the question implicitly smuggles a realist preconception about actions: either they're "wrong" or they're not, *period*. Yet for both appraiser and agent relativists, actions aren't simply right or wrong, permissible or impermissible, and so on. Instead, whether an action is morally right or wrong varies relative to different moral standards. By asking participants whether a particular action would be "wrong," participants aren't given sufficient information about which standard is being indexed.

This leaves us with a serious dilemma: if you're non-relativist antirealist, neither response option would reflect your metaethical stance. If you're a relativist antirealist, the response options are ambiguous in a way that renders them effectively unanswerable. For relativists, moral claims *must* be indexed in order to make sense. Yet the response options do not respect this fact, instead presenting participants with a use of the term "wrong" without sufficient context for it to be clear what standard, if any, it is presumptively indexed to. This doesn't make any sense. When one regards claims as having an indexical component, there must be sufficient information to know how the claim is being indexed. To illustrate, imagine there were two people, Alex and Sam. Now suppose one of them says:

> *"I am Alex."*

Is this statement true or false? There is no way to answer, if we do not know *which* of the two people made the claim. Just the same, moral relativists cannot judge whether an action is "wrong" or not. They can only judge whether it is wrong *relative to a standard*. Zijlstra's response options don't make the

standard to which the putative wrongness of the action in question is relativized sufficiently clear. As a result, there is no way for antirealists to respond to this question that accurately reflects their views or does so in a way that isn't misleading. Furthermore, to choose the "antirealist" response option could imply that the participant is less committed to their moral standards than someone who chooses the realist response option, since it could convey that their current normative moral standards are contingent, unstable, or not applicable to people with different beliefs or practices, none of which is entailed by antirealism.

This scenario also relies on counterfactual thinking. It's not clear how capable ordinary people are of engaging in counterfactual thinking. At the very least, they lack the training and experience philosophers have with engaging in thought experiments. As such, there is also ample opportunity for confusion and performance error resulting from the task being confusing or too cognitively demanding.

I want to end with a general criticism of the way Zijlstra frames the results of these studies. In two of the three studies, Zijlstra found that a majority of participants favored the response option associated with implicit realism, while one of the three conditions yielded inconsistent results that didn't demonstrate that most people favored the response option interpreted as implicit realism. Results are, at best, mixed. I say at best because I am deeply skeptical of these findings for the many reasons outlined here. But let's set these aside, and take the study at face value. What would these findings show? They suggest that perhaps two thirds to three quarters of people are inclined towards implicit realism. Yet Zijlstra states that "Overall, these results provide support for Enoch's thesis that people are moral objectivists," (p. 9). What does Zijlstra mean, *people are moral objectivists*? Some people? All people? This statement is, at best, unclear, and at worst, misleading. One cannot conclude that *all* people are implicit realists because a tentative majority favor a particular position. Yet Zijlstra continues to make generalizations like this: "The results of the survey experiment reveal that people

do respond in ways that supports Enoch's conjecture that people must on some level be moral objectivists" (p. 9), and later, "Before we conclude that Enoch's tests show that lay people are moral objectivists [...]" (p. 10). Such remarks appear numerous times in the article. Unqualified remarks about what "people" are strike me as rather strange. *Which* people? All of them? Some of them? Do these findings generalize to the entirety of humanity, or only to specific populations? To illustrate how strange these remarks are. According to the United States Census Bureau (2021), 80.7% of people in the United States live in urban areas. That number is a larger proportion than the proportion who favored implicit realist responses in Zijlstra's studies. Imagine if we concluded from this result that: *People live in urban areas.* What would this mean? It could mean that among the places people live, this includes urban areas. It could mean that all people live in urban areas, that most do, that some do. It's simply unclear what this means.

Part of the goal of social scientific research is to provide data and statistics, and explain how these findings support or conflict with various hypotheses. While generalizations can be helpful in capturing the essence of a finding without overcomplicating the matter, researchers often lean into generalization and oversimplification to the point that they make statements that are confusing, unclear, or misleading. Unfortunately, this seems to have happened in this case as well. Even if 60-75% of people exhibited an implicit commitment to realism, it would not entitle realists to claim victory as though *everyone* were implicitly committed to realism. Incidentally, the latest poll of analytic philosophers in the Anglophone world found that 62% endorsed moral realism (Bourget & Chalmers, ms). This is amusingly close to the proportion of realists found in Zijlstra's studies, though I suspect the similarity is a coincidence. Part of why I find it amusing is that, in my frequent debates with moral realists, many have drawn on this survey result as a kind of trump card, leveraging the fact that a majority of philosophers are moral realists as some kind of compelling evidence for realism. It isn't. But there seems to be a strange tendency for people to regard a majority in favor of one view over

another as deeply impactful, as though crossing the threshold of 50% leads not merely to a proportionate edge over the opposition, but a discrete, qualitative edge that quickly approaches victory as one moves towards 100%, even if one falls short. There may be deep social roots in such sentiments. Philosophers often seem affronted or frustrated with me when I seem perturbed by the fact that I'm in the minority by endorsing moral antirealism. The expectation seems to be that I should place great esteem in the majority view, and that there is something impertinent, arrogant, and even foolish in persisting, with confidence, in a minority view. I suspect this tendency to think the majority has some kind of edge has bled into the way people interpret results in folk philosophical studies. Any instance in which a majority favors a particular view is often accompanied by vague or ambiguous remarks suggesting that people in general hold that view, or that everyone holds that view. If 70% of people endorse a particular view, why not just *say that*, even if it lacks the punch of saying "*people*" endorse the view?

**S3.10 Training paradigms**

Training paradigms present participants with instructions, training exercises, or detailed response options in an effort to instill adequate understanding of the relevant metaethical distinctions before measuring folk metaethical stances and commitments. At present, there are only two studies that have employed training paradigms, Wright (2018) and I address the Pölzler (2020a; 2020b). I address general problems with training paradigms in **Chapter 3**. Here, I discuss the distinctive methodological shortcomings of particular studies.

*S3.10.1. Wright's (2018) training paradigm*

*S3.10.1.1 Relativism vs. non-relativism*

Wright (2018) employs a training paradigm to evaluate support for relativism and to distinguish support for cognitivism vs. noncognitivism. To measure belief in metaethical relativism, Wright (2018)

asks participants to read a paragraph that describes the difference between relative and nonrelative terms before asking them to judge which of the two best reflects how they view moral issues:

> Consider the difference between the term "triangular" vs. the term "tall". The first of these terms is a non-relative term, meaning that the context in which it is uttered does not influence its truth value—e.g., the statement "That shape is triangular [i.e., it is a shape with three sides and three corners]" is either true or false of the shape being talked about no matter who says it, when it is said, or what frame of reference is being used. If it is true that the shape being referred to is triangular in one context, then (barring something happening to change the shape) it will always be true that it is triangular, regardless of the person making the statement and/or the time, place, situation in which it is uttered.

> On the other hand, "tall" is a relative term, and, therefore, the statement "Naomi is tall" could be true or false, depending on the context/the frame of reference under which it is uttered—e.g., whether we are comparing Naomi, who stands 5'6", to a group of women from a Black Hmong village in Vietnam (who, at their tallest, stand about 5") or to a group of NBA players (who, on average, stand about 6'7"). It would also be the case that we'd consider the statement "Naomi is tall" to be true if uttered by a Black Hmong woman, but not true if uttered by an NBA player. In other words, for relative terms, the person making the statement and/or the time, place, situation in which it is uttered makes a difference. Frame of reference is important for determining truth-values.

> Please keep this distinction between relative and non-relative terms in mind as you participate in the next exercise. (pp. 126-128)

These instructions are long and complicated and include a variety of sophisticated and unfamiliar technical concepts using terms that are likely to be unfamiliar to readers (e.g. "frame of reference", "truth value," "non-relative"), or use terms ordinarily familiar terms in narrow and specific ways (e.g., "relative," "context"), or present distinctions that could be readily interpreted in ways orthogonal to the relativism/non-relativism distinction, or For example, Wright uses "triangular" and "tall" to represent nonrelative and relative concepts, respectively. Yet these concepts more naturally represent the distinction between categorical and continuous differences *categorical* (e.g., yes/no, on/off) versus *continuous* (e.g., height, duration) variables.

This would not be a problem if this interpretation were not easy to confuse with the intended distinction. Unfortunately, it is unclear whether people wouldn't be disposed to confuse the two.

Evaluation of open response data in response to attempts to assess folk beliefs about realism and antirealism reveals that people frequently misconstrue the distinction between "objectivist" and "relativist" options as a distinction between the notion that morality is either "black and white" or that there are "gray areas". Although it is unclear what exactly this means, and meaning likely varies across participants, one possibility is that people could mistakenly interpret Wright's explanation of relativism and non-relativism to be a distinction between whether a given action or action-type is either categorically right or wrong or is right or wrong in some respects but not others. That is, people may conflate the distinction between relativism and nonrelativism with a distinction between absolutism and non-absolutism, generalism and particularism. For instance, when asked what it means to say that the truth of a moral claim is "objective", one respondent stated that:

> It means that whether an act is right or wrong can be determined precisely or in a binary manner based on the facts. And there is no continuous measure of morality or exceptions to the rules.

It is also possible for participants to interpret the instructions in a way consistent with conflating epistemic distinctions with metaethical ones, i.e., that the answer to some moral issues is clear but it is unclear in other cases. Many people who encounter moral issues do appear to be concerned with the distinction between moral issues that have obvious answers and ones that are difficult to morally evaluate, or are subject to a more nuanced evaluation, often *because* they incorporate situation-specific details, which could mean that ordinary people don't cleanly distinguish epistemic considerations from normative or metaethical distinctions.

In fact, Wright's instructions may actively cultivate these unintended interpretations. Since Wright employs terms like "context," and "situation," this exacerbates the possibility that participants will take relativism to refer to the notion that whether an action is right or wrong depends on the particular details of the situation, and that it is not the case that we can apply a rigid and absolute moral rule that is insensitive to the unique characteristics of each situation, that is, which could be captured

by the distinction between absolutism and contextualism, or the distinction between generalist and particularism.[147]

For instance, Wright states that, for a non-relative claim, "the context in which it is uttered does not influence its truth value [...]" (p. 126). Although some of the surrounding language implies that "context" the truth-value of the statement doesn't vary based on who makes it or when it is made, which could allude to relativism, it also includes the notion that its truth does not vary regardless of "the situation in which it is uttered." Yet relativism in metaethics does *not* refer to the claim that whether a moral claim is true or false depends directly on the context in which it is uttered. Rather, it depends on *who* is making the moral claim (agent relativism) or judging the claim (appraiser relativism). Consider four possible metaethical positions:

(1) *Agent subjectivism*: Moral claims are true or false relative to the standards of the person expressing the moral judgment

(2) *Appraiser subjectivism*: Moral claims are true or false relative to the standards of the person evaluating the moral judgment

(3) *Agent cultural relativism*: Moral claims are true or false relative to the standards of the culture of the person expressing the moral judgment

(4) *Appraiser cultural relativism*: Moral claims are true or false relative to the standards of the culture of the person evaluating the moral judgment

In the case of (1) and (2), the situation in which a moral claim is uttered is irrelevant. Whether the moral claim is true or false depends on the standards of the person making the claim, or the person judging the claim and those standards don't necessarily vary as a function of the situation.[148]

For instance, an agent relativist would say that if Alex believes abortion is wrong, and says "abortion is wrong," then this is true (relative to Alex's moral standards) regardless of the situation Alex is in. An appraiser relativist might think that if Sam hears Alex say this, but thinks that abortion

---

[147] These are subtle but different distinctions. Absolutism holds that there are moral rules
[148] Unless it is part of the person's standards that they do, but that is irrelevant because it isn't the appropriate kind of situational variation.

is not wrong, then Sam is correct relative to Sam's standards, since what matters is who is evaluating the claim, not who is making the claim (though, of course, Alex would presumably evaluate her own claims in a way consistent with her own moral standards). Once again, the situation in which it is uttered is irrelevant.

An agent cultural relativist would likewise not be concerned about the situation in which a claim is made. If Alex says that abortion is wrong, the agent cultural relativist would judge this to be true or false relative to the standards of Alex's culture. Alex could be visiting people in another culture, or even living on an alien planet, and this would be irrelevant. The situation does not matter. What matter is whether abortion is consistent with the standards of Alex's culture.

It's a bit of a stretch, but the only conventional form of relativism according to which one could reasonably regard the status of a moral claim as depending on the situation it is uttered in would be some form of appraiser cultural relativism. The reasoning might go something like this: whether an action is morally right or wrong depends on the cultural context in which the action is performed. For instance, suppose Alex believes it is morally wrong to steal, and it is also wrong according to the norms of Alex's culture. If Alex is in her community, and someone steals her wallet, Alex's judgment that this was wrong would be true. Yet if Alex was visiting another culture, and someone stole her wallet, but members of that culture did not regard stealing as wrong, then a relativist might think that if Alex judged that this act of theft was morally wrong, that Alex would be mistaken. This is because a relativist might think that whether an act is right or wrong depends on the culture in which the act is performed. If so, the judgment that "stealing is wrong," could be true or false depending on the *cultural context* in which it is uttered. Yet this is a narrow and specific form of relativism, and it would not be appropriate to use instructions that prompted people to think of relativism in this way as a method for explaining what relativism means. It is also an indirect, confusing, and strained way to convey relativism. Researchers should stop describing relativism as the view that whether an action is

right or wrong depends on the "situation" or the "context." There are far more natural interpretations of what this would mean that have nothing to do with relativism. For instance, someone might think that "it is morally wrong to get an abortion, except in the case of ectopic pregnancies," or they might claim that "it is usually wrong to kill people, but it is justified in some cases…" such a person might go on to list a wide and complicated variety of cases: self-defense, to save the lives of others, during just wars, to prevent a catastrophic event, and so on. It is natural to say, in such cases, that whether an abortion or an act of killing is morally permissible "depends on the context" or that it "depends on the situation." Note that this misleading characterization of relativism *is the very first distinct introduced*: participants are immediately told that the difference between "triangular" and "tall" is that the "context in which the first of these is uttered does not influence its truth value" (p. 126).

Open response data also supports the concern that participants often conflate the distinction between objectivism and relativism with the distinction between *absolutism* and *contextualism*, respectively. By *absolutism*, I mean the normative moral belief that a given type of action, e.g. theft or torture, is wrong in all circumstances. *Contextualism*, on the other hand, holds that whether actions of a given type, such as theft or torture, may or may not be morally permissible, depending on the circumstances, e.g. theft may be permissible in order to save a life, but not otherwise. For instance, when asked to explain what it means to say that the truth of a moral claim is relative, a respondent said:

> *[...] Most people would agree that stealing is morally wrong, that this is a moral truth. But someone who steals to survive is often excused because it's also a moral truth that we preserve ourselves and the survival of our species So the moral truth is relative, based on circumstances*

while another stated that:

> *It means that morality cannot be applied systemically and must take situational factors into account.*

As these examples illustrate, when directly asked what it means for a statement to be "relative," people often interpret this as a distinction.

Wright acknowledges another limitation of this study. After receiving instructions, participants were then told to consider a disagreement between two people about whether a third person who performed a particular action did something morally right or wrong, and were then asked to assess whether they think both people could be correct or whether one must be incorrect. This is a version of the disagreement paradigm, yet as Wright acknowledges, it specifically is asking for a contrast between appraiser relativism and the rejection of appraiser relativism. However, while Wright classifies those who reject appraiser relativism as expressing a "non-relativist" response, this does not mean that these people are non-relativists; it would at best only mean that they aren't appraiser relativists. It is still possible that they are agent relativists. It is also possible for someone to think that moral standards could be relativized to both agents and appraisers. In both cases, such relativists lack any appropriate way to respond to the question. Technically, agent relativists ought to favor the judgment that one of the two statements would be correct, but this requires a sophisticated capacity for recognizing that the distinct form of relativism one endorses is consistent with this claim, despite it *appearing* to convey something that more closely approximates realism. It is not reasonable to expect participants to exhibit this degree of sophistication, but to the extent that they did, this would mean miscategorizing some unspecified number of relativists as non-relativists.

Of course, since participants were given a version of the disagreement paradigm, the results of this study are saddled with all the attendant methodological problems of this approach. And the non-relativist option allows the participant to express that only one statement would be correct, but adds that this is so regardless "of the contexts in which it is being made," while the choice that both are correct may depend on "the context in which it is made," once again inviting a conflation between relativism and contextualism/particularism.

None of these issues represent the most serious problem with the instructions or response options. The most serious problem is that Wright presented participants with the wrong form of

relativism. Wright appeals to a distinction between whether something is "triangular" with whether someone is "tall." The former is nonrelative, presumably because whether something is a triangle or not does not depend on the standards of the speaker or the speaker's culture. Yet to contrast this notion with the proper form of relativism, Wright should have given an example where whether the claim was true or false *did* depend on the standards of the speaker or their culture. This is because metaethical relativism is typically characterized as a form of *stance-dependent relativism*, rather than *stance-independent relativism*. As Joyce (2015) notes, the distinction between stance-dependence and stance-independence is orthogonal to the distinction between whether the truth status of a claim is objective (i.e., non-relativized) or relative. Stance-dependent relativism would treat moral claims as true or false relative to the standards of people or groups. Stance-independent relativism would treat moral claims as true or false relative to some standard *other than* the stances of individuals or groups (or some other standard), while a stance-independent form of relativism would treat moral claims as true or false relative to some standard *other than* the stances of individuals or groups. It is difficult to imagine plausible examples, but in principle moral claims could be true or false relative to what time of day it is, or to one's geographic location. Such possibilities are bizarre, and are emblematic of how unusual such relativization would be. Metaethical relativism almost always presumes that moral claims are true or false relative to a standard of evaluation, and are thus stance-dependent. Yet Wright's notion of tallness is not a stance-dependent form of relativism, it is stance-independent. This is because whether someone is tall compared to others does not depend on the standards of individuals or groups. Consider Wright's example:

> "Naomi is tall" could be true or false, depending on the context/the frame of reference under which it is uttered—e.g., whether we are comparing Naomi, who stands 5'6", to a group of women from a Black Hmong village in Vietnam (who, at their tallest, stand about 5") or to a group of NBA players (who, on average, stand about 6'7"). It would also be the case that we'd consider the statement "Naomi is tall" to be true if uttered by a Black Hmong woman, but not true if uttered by an NBA player.

Taken literally, this last statement is false: whether "Naomi is tall" is true or false would not depend on who is making the claim, but who Naomi is being compared to. While it is plausible that a Black Hmong woman would compare Naomi's height to other Black Hmong women and the NBA player would compare Naomi's height to other NBA players, they aren't *required* to do so. If an NBA player visited a Hmong village, saw Naomi among the villagers, and said "Naomi is tall," they could mean "relative to the height of other women in the village," "relative to NBA players," "relative to the global average height of women," or even "because she is taller than me." In all of these contexts, it is not the beliefs or attitudes of the speaker that determine whether it is true that Naomi is tall, but the standard of comparison that the speaker has in mind, yet such standards are *not stance-dependent*. This is because it is a stance-independent fact whether Naomi's height is greater than the average of some group she is being compared to. This type of relativism is very different from the type of relativism discussed in metaethics. The typical (stance-dependent) relativism would treat whether "Naomi is tall" is true or false as dependent *on the subjective standards of the speaker or their culture*, or on the standards of the person or the culture of the person *judging* the speaker's claim. These are not the kinds of examples Wright provides in the instructions. Ironically, Joyce even uses tallness and pro basketball players to illustrate stance-independent relativism in order to contrast it with the stance-dependent form: "Consider: Tallness is a relative notion—John is a tall man but a short pro basketball player—but it is not the case that 'thinking makes it so.'"

In fact, the type of relativism presented in the instructions is inconsistent with the response options, since the response options given to participants presuppose a form of appraiser relativism that is most reasonably interpreted as a stance-dependent conception of relativism. This inconsistency further undermines the validity of these instructions, but incidentally also renders the response options invalid. Since participants were given the wrong instructions, and an inconsistent set of response options, the results of this study should be interpreted with considerable caution. Note also that no

mention is made of comprehension checks, and no training exercises were implemented. In their absence, we cannot be confident that participants interpreted instructions or response options as intended.

### *S3.10.1.2 Categorical imperatives*

Wright's (2018) next goal is to assess whether ordinary people think there are *categorical imperatives*, which she defines as moral claims that "make reference to objective values" that provide "people with a reason to do/not do the action *independently from* (and even *in spite of*) any actual desires, inclinations, beliefs (etc.) that they might have to do/not do it" (p. 129, emphasis original). This is roughly consistent with the kinds of claims captured by my characterization of moral realism: that there are stance-independent facts about what we should or shouldn't do, which could be chased out in terms of what we have "a reason" to do. To assess whether ordinary people believe there are categorical imperatives, Wright presented participants with the following question and response options:

> If there were people who did not believe that there was anything wrong with doing *x* (o *x*-ing)—and, indeed, they wanted to do it—would there be any reason for that person to nonetheless refrain from doing it?
>
> - *There would be no reason for them not to x.* They should feel free to x if they so desired. [NON-CATEGORICAL]
>
> - *There still might be a reason for them not to x.* People in their family/community might disapprove of x-ing or type of action that x-ing is [NON-CATEGORICAL]
>
> - *There still might be a reason for them not x.* It is against the law and they could get in trouble for x-ing or for engaging in the type of action that x-ing is. [NON-CATEGORICAL]
>
> - *There is still a strong reason for them not to x.* It would be bad for them to x, even if they don't think so and they wanted to do it (and even if no one else would disapprove or punish them for doing so). [CATEGORICAL]

There are serious problems with this question. One problem is that it does not explicitly ask whether there would be any *moral* reason to do or not do it. Wright is interested in whether people believe that there are categorical *moral* imperatives: facts about what we morally should or shouldn't do that give us reason to do something independent of their desires. While the context of a question could make

it obvious that the question is asking about morality in particular, this question goes out of its way to frame things in a way that could readily prompt nonmoral considerations. Consider the initial setup: we are told that there are people who don't think something is morally wrong. We are then asked whether there would be *any* reason to not do it anyway. Not only could this prompt considering whether there could be nonmoral reasons, this might even be *more* plausible than considering whether they have moral reasons even if they don't think that they do. This is exacerbated by the response options presenting nonmoral reasons why a person shouldn't do something. The whole setup of this question, from the instructions to the response options, treats "reasons" in a completely generic form. Yet the only way to know whether participants think we have categorical *moral* reasons to do things independent of our goals, standards, and values is to specifically rule out consideration of any nonmoral reasons. This question not only doesn't do that, but appears to make an active effort to include such considerations. As a result, it is not a valid measure of belief in categorical moral reasons.

This is because participants could readily imagine many nonmoral reasons why someone might "have a reason" to not do something, that have nothing to do with whether the action is immoral. This could include practical reasons, such as reputational and legal consequences. For some reason, *only* these options were included in the response options, even though there are many other reasons why a person might "have a reason" to not commit the actions described in the study, including practical and personal consequences, such as guilt, psychological trauma, or health risks. Yet the response option for categorical reasons only indicates that the person wants to do it and that others won't disapprove or punish them. This isn't sufficient for capturing categoricity, because it does not rule out non-categorical reasons why a person might "have a reason" not to perform the action in question. Thus, another problem with this question is that the response option distinguishing categorical from non-categorical responses does not sufficiently ensure that participants understand that the categorical response option holds that there are reasons why one should not do something

that have *nothing to do* with what would be in the interests of the person performing the action, even if that person were not aware of it. For instance, imagine there is a smoker living during a period of time where we did not know that smoking was bad for your health. This person wants to be healthy, but they also enjoy smoking. They want to smoke, and nobody will disapprove of them or punish them for doing so (in fact, many people will actively approve of them smoking!). Someone who denies there are categorical reasons could still believe this person has a non-categorical reason to not smoke: *it's bad for their health*. Since being healthy is among this person's desires, they could have a non-categorical reason to not smoke, without having any non-categorical reason to not smoke. Yet Wright provides no response options for situations where a person denies both categorical reasons and denies the non-categorical response options given. The response options given are not mutually exhaustive.

Furthermore, the response options are not even mutually exclusive. Suppose you believe in categorical reasons, and think that a person should not engage in cannibalism even if they want to and won't suffer punishment or disapproval. It is still *also* true that if they *actually did* engage in cannibalism, that it would be against the law and their family and community would likely disagree. Yet participants are asked to select the "best" answer. All three of these answers are true, so why would the categorical answer be "best"? What would that mean? It can't be "more true." One way in which it is best is that it allows the participant to signal their moral disapproval of the acts in question in a way the "non-categorical" options don't. If so, this would undermine the validity of the measures, since participants wouldn't be selecting this option because it was true and the others were not, but because it allowed the participant to convey socially desirable attitudes and avoid the pragmatic implicature of inadequate disapproval that would accompany selective alternative responses; in other words, it would avoid normative entanglement.

These response options are rendered even murkier by the second and third option asking there *might* be a reason for them to not perform the action, because their family or community *might*

disapprove or because they *could* get in trouble. This is very strange language to include in a set of response options. Interpreted literally, of course their family or community *might* disagree, and of course they *could* get in trouble, in principle. It can simultaneously be true that the second and third options might be true, and that the third option is true. Again, these responses aren't mutually exclusive, yet they are presented as if they were, and participants are forced to choose the "best" answer without any clear indication of what would make an answer "best."

There are almost too many other minor issues to list. Consider the first response option:

- *There would be no reason for them not to x.* They should feel free to x if they so desired

Like all response options to this question, this response option is double-barreled. It contains two claims: that there would be no reason for them to not x, and that they should "feel free to x if they so desired." What if you agree with one of these claims but not the other? You have no way to indicate this in your response. Yet there is a more serious problem with this response option. Suppose you deny that there are any categorical reasons. Does it follow that you believe they have no reason not to x? No. Part of the problem with this question is that it does not exclude the possibility that a person has non-categorical reasons not to x. Second, disbelief in the existence of categorical reasons does not entail that one believes that people should "feel free" to do anything they desire. This remark implies the participant's approval of the actions in question. Yet disbelief in categorical reasons does not entail approval of other people's actions. For example, suppose I am a moral antirealist. I believe murder is morally wrong, and I strongly disapprove of anyone committing murder. I just don't believe there are any categorical reasons to not murder. If I am asked whether a person who desires to murder has any reason not to, my first response would be: yes, there are all sorts of non-categorical reasons to not murder. As a result, I would avoid this response option, regardless of whatever my other response options are. Yet suppose there are no better options, so I return to this one. I am then confronted by the second part of this response: that this person should feel free to murder others. Well, I *don't* think

they should "feel free" to murder others. But this has *nothing to do* with whether I think there are categorical reasons or not: rather, it simply conveys *my personal non-categorical normative standards*. This is not a valid response option for measuring disbelief in categorical reasons, because the second part of the response option appears to be a question about the participant's normative stance towards the action in question, not its categoricity.

Another problem with this item is the double negation: that there would be *no* reason for them to *not x*. This is a cognitively demanding consideration that is then paired with another statement to form a conjunct. Such complex questions are not ideal for surveying nonphilosophers, since they increase risk of performance error.

Another problem is that the second and third response do not make it clear that disapproval and punishment would be reasons to not *x*. Participants are told that they might have a reason not to *x* paired with a descriptive claim. Do these descriptive claims *give* or *provide* non-categorical reasons? Do ordinary people understand that? And is it appropriate for this to be implicit in the response options rather than explicitly stated? Also, why are these *non-categorical* reasons? Someone could believe that we have a moral obligation to act in accordance with our society's standards, or to seek the approval of our family or community, independent of whether we want to do so. This may seem implausible to many readers, but note that people in more conservative or interdependent (or "collectivist") societies may very well regard striving for the approval of one's family and community as an important virtue or moral duty. People might also think we have a *prima facie* moral duty to conform to just laws, independent of whether doing so is consistent with our desires. Thus, belief in categorical moral norms is consistent with response options (2) and (3). Yet because these response options do not explicitly indicate that the categorical reasons one might have must be moral, one might also think that there are categorical reasons to act rationally or prudentially. For instance, one might think that we have reason to behave in ways that do not undermine our goals and life projects.

Someone who engages in cannibalism or other activities in the list of moral issues Wright includes could find themselves ostracized by friends and family, exiled by their community, or imprisoned. Recognizing this, participants could think people have categorical nonmoral reasons for selecting response options (2) and (3). Thus, these responses are not valid representations of non-categorical moral reasons since they don't adequately exclude the possibility of interpreting them as providing non-categorical nonmoral reasons or categorical moral reasons.

There are also several issues with the final response option, which is supposed to represent a categorical moral reason. One problem is that it is unclear what "strong" means. Are categorical reasons necessarily stronger than non-categorical reasons? It's not clear that they are, or if so, in what respect they are stronger. We might think that categorical reasons always override non-categorical reasons, but this is a substantive position all on its own, and one need not believe this in order to believe or disbelieve in categorical reasons. The fact that we may have reasons to do something that don't depend on our desires does not entail that those reasons are stronger than our reasons to act in accordance with our desires. As such, it is not appropriate to bake this presumption into one's notion of a categorical reason, since one will in effect be asking a double-barreled question that presupposes a particular conception of the relation between categorical and non-categorical reasons: namely, that categorical reasons are overriding.

Another problem with the use of "strong" is that it seems like an odd contrast to the previous response options. Why wouldn't the fact that something is against the law and would risk punishment not be a strong reason to avoid doing it? Why wouldn't severe reputational consequences not be a strong reason to avoid doing something? Note that these activities involve *cannibalism* and *selling children on the internet*. People who get caught engaging in these activities would have their lives ruined. They could spend decades or the rest of their lives in prison, and even if they avoided prison or eventually got out, they would be ostracized by friends and family, universally hated, lose their jobs, lose custody

of their children, face divorce, lawsuits, revulsion, an inability to earn a living, inability to maintain social media presence, they would have few or no friends, be unable to find romantic partners, and so on. In short, their lives would be absolutely ruined. How would this *not* be a strong reason to avoid these actions? Wright's response options seem to presume that only categorical reasons are strong reasons, but there is no plausible reason to think this is the case, or to think ordinary people would think this way. As such, many participants may favor the "categorical" response option not because they believe in categorical reasons, but because it is the only way to adequately convey the strength of the reasons to not eat people or sell children.

Another problem is that this response option doesn't adequately convey that the categorical reason is a *moral* reason. Note the wording: "It would be bad for them to *x*." I'm a moral antirealist. I don't think there are any categorical moral reasons. Yet I still think it would be for people to *x*, if *x* is selling children or eating people for completely nonmoral reasons. However, a much more serious problem with this item is that I *also* think it would be bad for them to *x* because eating people or selling children is bad *according to my subjective moral standards*. In other words, I *do* think it's immoral for people to do these things, and I think it would be bad for them to *x* even though I don't believe in categorical reasons. A fatal problem with this response option is that it entangles a question about the existence of categorical reasons with a question about the participant's own normative standards. In Wright's defense, one could argue that this response option does adequately set aside the participant's own normative standards. Note that it says that it would be bad for the person to x "even if no one else would disapprove or punish them for doing so. Yet there are several problems with this. First, it is not clear whether "no one else" includes the participant themselves, rather than *hypothetical* people in the *hypothetical* world in which the act is being performed. If it doesn't, then the conflation would still be present. If it does, then this is unclear and at the very least not obvious. It would also be very strange: it would effectively involve asking ordinary people to imagine a world where they didn't

disapprove of or have any inclination to punish someone for performing an action, independent of their actual moral standards, yet to consider whether the action would still be wrong even if they themselves didn't think that it was. This is a difficult counterfactual to entertain for a professional moral philosopher who has the time to think about it, and experience addressing these issues, yet we're expected to believe that ordinary people responding to a multiple choice question would carefully recognize and consider an extremely difficult counterfactual like this and respond appropriately? And that they do so after correctly navigating ambiguities in this response option, and the ambiguities and oddities associated with the whole set of response options as a whole, such as that they aren't mutually exclusive, and that categorical reasons and non-categorical reasons are conflated with strong and (by implication, not strong) reasons? Finally, on top of all this, the participant has to imagine a world where nobody disapproves or seeks to punish people for actions that it's hard to imagine *nobody* disapproving of these actions. After all, many of these actions have victims. Are we to imagine that the children being sold don't disapprove of this state of affairs, or that the companies people are stealing from don't mind? The reference to eating other people is underdescribed, so are we to imagine that someone consented to being eaten after a natural death, or are we to imagine that someone killed someone else for the purposes of eating them, but that this person didn't disapprove? Asking participants to suspend disbelief to imagine a world where nobody disapproves of acts we regard as violent and repugnant, including the victims of those actions is an incredibly tall order. People struggle to accept the stipulations of the trolley problem (Ryazanov et al., 2018). Are we to expect them to accept counterfactuals where neither they nor anyone else objects to child trafficking, or that people don't mind being robbed or eaten? This is exceedingly implausible, and represents an incredibly cognitively demanding task, where most of the work is happening in one of several response options in a multiple choice question, which people's attention may be diffused across the available response options.

Normative entanglement could also motivate some participants to favor the "categorical" response option even if they do not believe in categorical reasons. Any response that you select could be taken to convey information about your normative moral stance, or to convey information about your moral character or level of moral commitment about a given issue. When someone is asked to select from among the four response options available, one reason for favoring the "categorical" response option is that it is the only way to adequately convey one's personal opposition to the acts in question. This is because the first option could be taken to convey outright approval of the action, insofar as the participant would be agreeing that people should "feel free" to sell children. Yet the second and third options also do a poor job of conveying one's personal opposition to a particular action. Imagine you were in front of an audience, and you were asked:

*Do you think people have any reason to not sell children on the internet?*

Would you say, "yes, their family would disapprove"? Would you say "Yes, it's illegal"? Imagine how your audience would react: that's a *weird* response. And it's weird because there is an expectation that you will find selling children repugnant and evil, and failure to immediately express this attitude *pragmatically implies that you don't have this attitude*. Any person concerned about their reputation is far more likely to say something like "Yes, of course! That's disgusting and evil! We should lock up anyone who does that!" This illustrates that one strong motivation for selecting the "categorical" response option is to adequately convey one's moral opposition to the act in question, rather than to convey their belief in categorical moral reasons. The contrast between this response option providing a "strong" reason and the second and third response options not doing so could also amplify the degree to which normative entanglement could motivate participants to favor this response for unintended reasons.

It is also possible that participants will simply reject the features of the response options. While researchers may stipulate that someone committing the actions in question would not suffer

disapproval or be punished, participants are not robots, nor will they tend to be professional philosophers with the requisite training and understanding to accept stipulative conditions and properly represent them such that their responses reflect an adequate representation of the attitude or position they'd take if the stipulated conditions real. Previous research shows that participants reject the outcomes of trolley problems, even when those outcomes are explicitly stipulated as a feature of the hypothetical. Anyone who has engaged in philosophical discussions with ordinary people will also be familiar with people rejecting elements of a stipulated hypothetical situation: they may insist certain outcomes aren't possible, that the situation is unrealistic or "not real," or otherwise object to various features of the stipulation. Such concerns may generalize to other scenarios, and this may be especially likely in circumstances where people are asked to suspend their beliefs and attitudes about moral behavior. Try asking ordinary people to imagine that it is morally okay to throw babies into a woodchipper and see if they can readily do so without objections. Even people who do their best may still fail to do so adequately, or may have normative considerations unwittingly intrude in their responses, resulting in performance errors that undermine the validity of the measure.

Another serious problem with the response options that Wright provides is that response options (2) and (3), which reference disapproval and punishment, respectively, could be interpreted as claims about what *is* the case, while categorical option can only be reasonably interpreted as a hypothetical claim about what would be the case given some counterfactual consideration. This is especially true of (3), which states that "It is against the law," not that it "could" or "might be against the law. As a result, participants are effectively asked to choose between options that describe what *is in fact the case* and options about what *would* be the case under counterfactual conditions. This is bizarre, because *such considerations are consistent with one another.* Imagine being asked, for instance, if someone asked to choose which of the following options was true:

(1) It is a bad idea to drive through red lights because it is illegal

(2)  It would be a bad idea to drive through red lights even if it was not illegal

Which of these response options is "better"? What would that even mean? Most people would think both statements are true. Yet these statements are analogous to the third and fourth options Wright gave to participants. It's also cognitively demanding to provide response options that contain both hypothetical and non-hypothetical parts, and response options that are mostly non-hypothetical and others that are mostly hypothetical, and ask them to choose which one is "best." Best based on what criteria?

Finally, regardless of what response participants select, it is unclear what it would mean for there to "be a reason" for people to $x$ or not $x$. I don't know what this means. Will participants? I'm not sure, but I have my doubts. While philosophers may find the notion of internal and external reasons to be intuitive and clear, and to readily consider whether people "have" reasons of either kind, it is not at all clear that ordinary people draw this distinction, or conceive of reasons in the same way as philosophers. It is even less clear that they could be reliably induced to assess whether they think we have reasons of the relevant kind without training and without those concepts being salient and clearly conveyed by experimental stimuli. For instance, if participants select one of these responses, can we confidently infer that they endorse any particular account of reasons? Can we infer that participants who select (1)-(3) believe in some kind of internalist account of reasons, where reasons must necessarily relate to one's goals or desires? Can we infer that participants who select (4) don't think this, and instead think that there are external or categorical reasons that "apply" (whatever that might mean) to people independent of their goals or desires? It's not clear we can make any such inferences. Nothing about the question or response options adequately disambiguates different notions of reasons, nor reflects those accounts in the response options. Virtually any response option is consistent with any conception of reasons, and, since participants are forced to choose between categorical and non-categorical "reasons," it employs a forced choice paradigm that imposes either a

kind of top-down characterization of endorsement of either categorical or hypothetical reasons on all participants, as though participants must endorse one or the other. I'm a reasons quietist: I think both conceptions of "reasons" are conceptually confused. Such conceptions may be philosophical inventions. Even if they are intelligible and not a product of confusion, we cannot simply presume that ordinary people *either* endorse one or another of two competing philosophical positions. They may not endorse either. I suspect this is the case, and that ordinary people don't have any determinate notion of "reasons" that would conform to these categories. Yet this study, like many others, presupposes that philosophical concepts and distinctions are necessarily reflected in the way ordinary people think, resulting in a study that by its very design would give the illusion that this was the case even if it wasn't.

There is another serious confound with this paradigm: by asking participants whether they believe we have categorical reasons to not perform particular actions, Wright has conflated metaethics with normative considerations. Suppose you believe it is morally wrong to steal, but that it is not morally wrong to get an abortion. Even if you believe we have categorical moral reasons to act in accordance with our moral obligations, you would still judge that we have categorical reasons not to steal, but you would not believe we have categorical reasons not to get an abortion. Yet this is not because you are a pluralist about categoricity with respect to our moral requirements, but because *you don't think abortion is immoral in the first place.* Whether a participant believes that, in any particular case, whether a person who wanted to perform a particular action had categorical reasons is going to depend on their normative standards, and not merely their stance on categoricity. By conflating the two, we cannot infer that participants are pluralists about categoricity with respect to moral norms, since we cannot tell if their judgment that we don't have categorical reasons to not perform an action in any particular case is because it is immoral, but we have non-categorical moral reasons to not perform the action, or because it isn't immoral. In other words, suppose two participants gave consistently non-

categorical responses with respect to the moral issues in this study. It could be that one participant chose non-categorical responses because they endorse a stance-dependent conception of moral reasons. They believe all the actions in question are immoral, but believe we only have non-categorical reasons to abstain from those actions. Yet the other participant could believe that there are categorical moral reasons for not performing certain actions, but they did not believe any of the issues in this particular study were immoral. This may be implausible (few participants are going to think it is morally permissible to sell children on the internet), but the point isn't that the latter pattern of response is likely; the problem is that participants who provided "pluralist" responses could be pluralists about the categoricity of moral reasons *or* the moral status of any given action, and since we cannot infer which of these factors drove their response to any particular issue, we can infer almost nothing about the degree to which participants consistently judge moral issues to be categorical or non-categorical.

Even if this paradigm were interpreted as intended, the particular moral items selected are not representative or expansive enough to allow for generalizations about the moral domain as a whole. Thus, we would not be able to appeal to the proportion of participants who gave consistent categorical or non-categorical responses, or pluralist responses, to the moral domain as a whole. There is no reason to believe that the particular moral issues that were selected are a representative sampling of the moral domain (assuming there is a moral domain; see Stich, 2018). For instance, Wright reports that a majority of participants favored the categorical response for four moral issues, the non-categorical response for four, and were evenly split on the remaining two. Yet without knowing how well these 10 items reflect the moral domain as a whole, we would not be justified in drawing any inferences about people's general tendency to judge moral issues to be categorical or non-categorical.

For comparison, imagine presenting people with four extremely popular celebrities, and four extremely unpopular celebrities, and two obscure celebrities. If we asked participants how much they liked these celebrities, this might yield the unsurprising result that they tend to like the four popular

celebrities, dislike the unpopular four celebrities, and are indifferent to the two obscure celebrities. Would this allow us to estimate that people tend to like about 40% of celebrities? Of course not: the particular celebrities weren't randomly selected, so the proportion of participants provides little information about the proportion of celebrities they like in general. Researchers are aware of problems like these, i.e., studies that treat stimuli as random factors, and have proposed various solutions (Judd, Westfall, & Kenny, 2012). Unfortunately, these recommendations, and even awareness of this problem, have been largely ignored. As Baguley (2012) observes, such concerns may be of minimal concern if we don't wish to generalize from findings that employ a particular set of stimuli. Yet Wright does generalize to the moral domain as a whole, concluding that "In sum, this investigation revealed a high and consistent degree of pluralism in the way people think about moral issues and evaluate moral discourse" (p. 130). Second, if there is little variation in the domain you are studying, then any particular stimuli you select may adequately represent that domain. If so, there would be little problem with generalizing from any particular set of stimuli to that domain. Baguley offers precision-engineered equipment as an example, e.g., products produced in factories with highly precise machines may be fairly uniform. Any particular widget may be more or less the same as any other. But it's not plausible that this is the case for moral issues, and at the very least we're not entitled to presume this is the case.

Finally, note in **Table S3.4** how bizarre the results are. Am I supposed to believe 25% of people think that it's *not* the case that it would be bad to sell children on the internet, even if the person who wanted to do so thought it was okay, wanted to do it, and if nobody would disapprove or punish them for doing it? I would have liked to see written responses explaining why people made the choices that they did. I suspect many people would choose the second and third options, that there "might still be a reason for them to not x," e.g. family disapproval and the law. Why wouldn't these be reasons *in addition to* the action in question being immoral? Categorical and non-categorical reasons aren't mutually exclusive, so what on earth does it even mean to choose the "best" response?

**Table S3.4**

*Proportion of items judged categorical and non-categorical in Wright (2018)*

| Issue | Categorical | Non-categorical |
|---|---|---|
| Selling children on the internet | 75% | 25% |
| Eating part of another human being | 69% | 31% |
| Steal money and/or supplies from the large company where you work | 45% | 55% |
| Conscious discrimination based on race/gender | 70% | 30% |
| Having sex with someone other than spouse | 61% | 39% |
| Refusing to provide help to those who need it | 35% | 65% |
| Helping terminally ill patients | 16% | 84% |
| Engaging in prostitution | 43% | 57% |
| Eating your pets (that died from an accident) | 45% | 55% |
| Terminating pregnancy | 23% | 77% |

### S3.10.1.3 Cognitivism vs. noncognitivism

Wright (2018) also introduced two paradigms for assessing whether ordinary people are cognitivists or noncognitivists. Wright employs two paradigms because she draws a distinction between two types of cognitivism:

(1) *Semantic nonfactualism:* "The denial that moral statements express propositions or have truth conditions (i.e., that they are 'truth apt')"

(2) *Psychological noncognitivism*: The denial "that the mental states that moral statements are conventionally intended to convey are beliefs (or other related cognitive mental states)" (p. 131)

These roughly correspond to my distinction between metaethical *commitments* and *stances*, respectively. Wright opted to test each separately.

### *S3.10.1.4 Semantic nonfactualism*

To test for semantic nonfactualism, Wright again adopted a training paradigm, which involved what Wright refers to as an "Introductory Exercise" (p. 131). All participants had to first complete the exercise before proceeding with the study. Participants were told to carefully read a set of instructions in order to "properly attune" them "to the difference between statements that are 'truth-apt' and those that are not":

> Some statements assert propositions that are what we call "truth-apt"—that is, they are meant to reflect matters of fact about the world (though sometimes they may fail to do so), which means they will be either true or false. For example, if I said to someone that "Boston, MA is north of Miami, FL" I would be stating something that is truth-apt—it is either true or false. In this case, we can easily establish whether my statement is true or false (e.g., by looking at a map). And, as it turns out, it is true. If, however, I had stated that "Boston, MA is south of Miami, FL", it would have been false. Either way, the important thing is that there is a fact of the matter (in this case, the geographical relationship between Boston and Miami) that my statement was meant to assert.

> Determining the truth/falsity of statements like the above is relatively easy. But sometimes it isn't easy. Consider, for example, the statement that "The earth is the only planet in our galaxy with life on it". We simply don't know at this point (and, indeed, we may never know) whether this statement accurately reflects a matter of fact (that is, whether it accurately reflects how many planets in our galaxy actually currently support life). So, we have no way of establishing whether the statement is true or false—but, nonetheless, it is still truth-apt. It is either true or false—i.e., either the earth is the only planet in our galaxy with life on it or it isn't. So, if one person said "The earth is the only planet in our galaxy with life on it" and another person said "Earth is not the only planet in our galaxy with life on it", one of these people would be correct and the other one mistaken (even if we can't say at this point which one is which).

> Consider, on the other hand, claims like "Peanut butter ice cream is delicious" or "Jazz music is the best form of music ever invented" or "Riding on the roller coaster at Elitch's is awesome!" Unlike the statements considered above, these statements aren't truth-apt. They are neither true nor false—there isn't a fact of the matter about the world that they are intended to reflect. In other words, there isn't an actual fact of the matter about whether peanut butter ice cream tastes delicious or riding the roller coaster at Elitch's is awesome. Some people enjoy

the taste of peanut butter ice cream, others don't. Some people have a great time riding the roller coaster at Elitch's, others don't. So, if one person said "Riding roller coasters is awesome!" and another person said "Riding roller coasters is absolutely terrifying!" it wouldn't make sense to say that one of the two was correct and the other mistaken. This is because neither of these statements are intended to accurately reflect some fact about roller coaster riding—rather, they are expressions of people's liking/disliking of or approval/disapproval for something (in this case, riding roller coasters). In other words, statements like "Riding on roller coasters is exciting" or "Peanut butter ice cream is delicious" are not truth-apt—they are neither true nor false. Instead, they are expressions of what we call people's "pro/con attitudes" (i.e., their positive/negative feelings, likes/dislikes, approval/disapproval, etc.).

It is important to recognize that truth-apt statements about ice cream and roller coaster riding can be made—for example, "Meredith hates peanut butter ice cream" or "I really love riding the roller coaster at Elitch's" are both statements that are either true or false (either Meredith hates peanut butter ice cream or she doesn't, etc.). To illustrate further: Imagine that Meredith said "I hate peanut butter ice cream". In this case, she'd be stating something that is truth-apt, since her statement asserts a fact of the matter about herself (namely, that she hates peanut butter ice cream). But if instead she said "Peanut butter ice cream is disgusting", she'd be stating something that is not truth-apt, since it is a statement intended to express her dislike of peanut butter ice cream.

For the questions that follow, please keep this distinction in mind, as you'll be asked to identify which statements you think are "truth-apt" (i.e., asserting matters of fact that are either true or false) and which statements you think are not "truth-apt" (i.e., expressing pro/con attitudes, and so are neither true nor false). (pp. 131-133)

After the extremely lengthy instructions (750 words), participants were then given a training exercise that involved categorizing ten statements as either truth-apt or not. Only participants who categorized at least 9 of the 10 items proceeded with the study. 23% of participants failed at this task, and did not continue. This is *a lot*. As Bergenholtz, Busch, and Praëm (2021) point out, "[...] experimental philosophy studies sometimes exclude an alarmingly high number of participants," which, they argue, "threatens the external and internal validity of the conclusions being drawn [...]" (p. 1531). They also reference van 't Veer and Giner-Sorolla (2016), who suggest that around 25% is a rough cutoff for the number of participants who can fail a comprehension check, after which "the instructions and/or scenario should be rewritten and the experiment rerun in order not to put the study's objective in

jeopardy" (p. 1536). At 23%, Wright's study comes perilously close to this cutoff. Bergenholtz et al. also emphasize that it isn't just how many people you exclude, but why you exclude them. As they point out, unlike standard comprehension checks where the correct answer is uncontroversial:

> Philosophical thought experiments often rely on the acceptance of certain key premises that may be regarded contestable [sic], and asking comprehension questions involving such key assumptions could be problematic as that may result in some participants being inadvertently excluded from the study, potentially creating a selection bias. (p. 1531)

In other words, they correctly recognize that comprehension checks, as they are used in folk philosophical research, can exclude people simply because those people disagree with certain philosophical assumptions presumed by the comprehension checks, rather than because they didn't understand the stimuli. As they point out, unless this is one's goal (and there's no reason to think it is in this case), this threatens the internal validity of the study. As they remark later, this problem may be distinctive to folk philosophical research:

> This seems to us to be a rather curious and particular problem for experimental philosophy. On the one hand, we want to ensure that respondents actually grasp the key premises of the presented thought experiment. On the other hand, if we include key premises as part of comprehension questions, we run the risk of excluding people who fail comprehension tests due to strongly held beliefs or intuitions that go against key premises of the thought experiment. (p. 1543)

If participants are excluded because they disagree with the premises presumed by the stimuli, this runs the risk of nonrandom exclusion, resulting in a pool of participants self-selected for particular philosophical positions, which will typically result in a sample unrepresentative of the intended population (e.g., "ordinary people," not "ordinary people who happen to agree with these particular philosophical positions").

Next, participants were given a set of 20 issues and were asked to categorize them as *moral* or *not-moral*. Finally, participants were presented with a sentence stating that "It is wrong to [issue]," for

each issue, e.g., "It is wrong to [sell children on the internet]." For each of these statements, they were asked to judge whether the statement was truth-apt or not with the following response options:

- "truth-apt" (*assertions of matters of fact* that are either true or false).
- not "truth-apt" (*expressions of positive/negative feelings*, pro/con attitudes, etc., that are neither true nor false).

12 of the 20 items were classified as moral by most participants. Out of these 12 items, most participants judged 3 to be truth-apt, 4 to be not truth-apt, and 5 were about evenly split. 76% of participants assigned at least one item to each of the two categories, indicating that a majority of participants were metaethical pluralists with respect to semantic cognitivism.

The length of these instructions is one cause for worry. At 750 words, such extensive instructions provide ample opportunity for participants to become bored, inattentive, confused, or fatigued, and for researcher bias to influence their understanding in ways that biases subsequent results. Participants are at risk of being exposed to conflicting or unclear stimuli, and researchers run the risk of misleading or inaccurate information. For instance, the notion that truth-apt statements "reflect *matters of fact* about the world" could potentially confuse some participants (p. 131, emphasis original). Some participants may interpret facts "about the world" to refer to natural facts, or facts that can be discovered (or only discovered) using scientific or empirical methods. Facts about geography or biology may be facts "about the world," but what about mathematical facts, such as "2+2=4" or "triangles have three sides"? These may be facts, but are they facts *about the world*? Many moral realists reject moral naturalism, and would not think of moral facts as natural facts of the sort one might think are facts "about the world." Rather, they may think of moral facts as more akin to mathematics or *a priori* knowledge. While these non-naturalist moral realists might nevertheless describe such moral facts as facts "about the world," it's not clear that lay people would do so. And if lay people are inclined to think in ways similar to moral realists, they may understand this description

of truth-apt statements in a way more closely approximating naturalism or a distinct claim about what *type* of facts moral facts are.

This isn't limited just to thinking that if moral facts are "facts about the world," that they would need to be natural facts. For example, people could understand "about the world" in a way that implies stance-independence. For instance, suppose an ordinary person endorses subjectivism. They believe moral facts are indexical statements conveying some fact about the speaker's moral standards, e.g., if Alex says, "murder is wrong," this means something like "murder is inconsistent with my moral standards." As such, moral utterances are truth-apt, but they only convey facts about the speaker's subjective standards. Would ordinary people who endorse some inchoate notion of subjectivism still regard such facts as facts "about the world"? It's not obvious that they would. They might take "about the world" to mean something approximating stance-independence. While many philosophers might regard mental states, such as beliefs and attitudes, as facts "about the world," some ordinary people may not. If so, they would interpret these instructions to imply that truth-apt statements aren't merely truth-apt, but stance-independent or nonrelative. This would be an unintended interpretation.

This example serves to illustrate how a single throwaway line in a text could actively serve to cause unintended interpretations. While I grant that this is speculative, it is not idle speculation. As a discipline, a great deal of philosophical work centers on clarity and disambiguation. Philosophers with even a modicum of experience interacting with their colleagues and laypeople alike are likely to recall numerous instances in which turns of phrase led people astray because interlocutors were not on precisely the same page about what a word or phrase means. It is difficult to overstate how central such misunderstandings are to the entire enterprise of philosophy. Such confusions and misunderstandings are so commonplace that some philosophers have floated the possibility that *all* problems in philosophy are rooted in linguistic and conceptual confusions (Gill, 1971; Schlagel,

1974).[149] While this may overstate the degree to which linguistic confusion contributes to the persistence of philosophical disagreement, there is little reasonable dispute that linguistic confusions, e.g., ambiguity, underspecificity, and equivocation play a significant role in misunderstandings; the only question is *how much* of a role do they play, both in general and in any particular case. Conventional social scientific research that involves familiar stimuli may not suffer much from such concerns, but few philosophers would dismiss the possibility that linguistic confusion could account for both the failure of two interlocutors reaching an accord in any given philosophical exchange, and for the general intransigence of at least some significant philosophical controversies. In short, we cannot simply presume that ordinary people would interpret "about the world" consistently and in a way that doesn't threaten their understanding of what it means for a statement to be truth-apt. This is the sort of thing we'd want to pretest, to assess whether people interpret it in a way that doesn't distort their understanding of "truth-apt." Note that this is just one of the ways instructions could mislead participants.

A far more serious problem with Wright's instructions is that it is not obvious, much less necessarily true, that the examples of noncognitivist statements are uncontroversially noncognitivistic. Wright provides the following examples:

> *"Peanut butter ice cream is delicious"*
> *"Jazz music is the best form of music ever invented"*
> *"Riding on the roller coaster at Elitch's is awesome!"*

---

[149] Schlagel (1974) opens his critique of Wittgenstein with the following account:

"There is a doctrine about the nature and function of philosophy which is so prevalent among Anglo-American philosophers today that it deserves to be described as the official theory. This official doctrine, which derives mainly from the later writings of Wittgenstein, goes something like this.'" Most (if not all) philosophical problems are not genuine problems in the sense of arising directly from empirical inquiry or indirectly from conceptual difficulties growing out of empirical inquiries (as in the various sciences), but arise because philosophers misuse ordinary forms of speech or place a strange interpretation on common linguistic uses which results in a distorted way of construing things. While the philosopher believes he is using language to think about the world, actually he becomes so entangled in the grammar of language that he cannot get beyond this structure to anything outside it, mistaking this for the logic of the problem." (p. 539)

Participants are told that "Unlike the statements considered above, these statements aren't truth-apt." Such instructions are not appropriate. It is an open question whether statements about personal taste or other preferences are truth-apt. Although preference claims are not identical to aesthetic claims, they are similar in many respects, yet the most common response academic philosophers provide when asked about their stance towards aesthetics is aesthetic objectivism (43.5%), followed by aesthetic subjectivism (40.6%). What you *don't* see is a significant proportion of aesthetic noncognitivists.[150] It may likewise be the case that many philosophers would object to noncognitivism about preference claims. At the very least, one is not entitled to simply presume noncognitivism is the correct account of preference claims. Indeed, in what may be a twist of irony and coincidence, Kirwin (2021) has recently and explicitly defended a realist account with respect to food preferences. What is especially amusing about this is that the *primary example that used in the article is realism with respect to peanut butter ice cream*:

> Nonetheless, I think that with the value-expertise model in hand, we will find that a comprehensive realism—realism even in those arenas that Loeb considers potentially absurd enough to serve as the absurdum for a reductio of value realism generally—is much more plausible than one might at first suppose. *My preference for peanut-butter-cup ice cream* can be understood, I'll argue, as exemplifying a particular form of gastronomic value-expertise (albeit a localized, small, and fairly unimportant one). My defense of this claim rests on a combination of positive arguments and defusing explanations for the sense of absurdity, as well as some clarifications concerning what the (comprehensive) realist is—and is not—committed to saying about such value. (p. 9)

---

[150] This could be an artifact of the way the question was designed (it asked: "Aesthetic value: objective or subjective?" which may have discouraged people from selecting "other").Yet the fact that prominent philosophers who designed these questions didn't even bother to include noncognitivism or some other way of denying aesthetic claims are even subjectively true or false goes some way in indicating that they didn't consider such a possibility common enough to consider.

Note, also, that other questions on this survey were not limited to just two options, but frequently included three or more. Furthermore, respondents showed considerable willingness to select "other" as an option and to elaborate when given the opportunity. "Other" was a popular choice, frequently compromising 15-35% of respondents. Indeed, for some items, "other" was the modal response, e.g., *Time: A-theory or B-theory?* (Other: 39.5%), *Arguments for theism (which argument is strongest?): cosmological, design, ontological, pragmatic, or moral?* (Other: 44.8%), and in one case even commanded a majority of respondents: Sleeping beauty (woken once if heads, woken twice if tails, credence in heads on waking?): one-third or one-half? (Other: 53.6%). If respondents really did find objectivism and subjectivism unacceptable because of a commitment to noncognitivism, it's unclear why they wouldn't have expressed this by selecting "Other" and offering this as their position. They were perfectly willing to do so for other questions.

Kirwin is arguing not simply that claims about food preferences can be true or false, but that some are true in a fully realist sense. While I am tempted to digress to discuss my own personal love of peanut butter ice cream (with or without the cups), my point here is simply that it is *not* the case that philosophers unreservedly regard claims about what food is good or bad to merely express nonpropositional attitudes; philosophers not only can, but demonstrably do argue that such claims can be true or false. One is not simply entitled to inform participants that such claims are noncognitivist. Wright does acknowledge this. In a footnote, Wright states that:

> This is highly nuanced and philosophically treacherous territory—especially when attempting to guide the "folk" through it. For example, it could be argued that "peanut butter ice cream is delicious!" is truth-apt, just relativized to the speaker. Nonetheless, there is also a reading of it in which it is not truth-apt, and not meant to be truth-apt, which seemed good enough for the goal of creating a way for participants to at least begin to see the distinction between statements of matters of fact vs. expressions of pro/con attitudes. As a first pass, there are likely to be a number of ways this instruction exercise can be improved. Thanks to John Parks for his helpful feedback here. For an excellent review of the issues—and pitfalls—associated with doing empirical research in this area, see Pölzler (forthcoming). (pp. 132-133)

Wright recognizes the possibility that such remarks could convey indexical factual claims. However, Wright's rationale for circumventing this problem is obscure: "Nonetheless, there is also a reading of it in which it is not truth-apt, and not meant to be truth-apt" (p. 133). What does Wright mean? There "is a reading"? Well, sure, that is one way it is *possible* to interpret such remarks. Yet this is trivially true of any normative or evaluative claims, including moral claims. One possible reading of "murder is wrong," is that it merely conveys a negative emotional stance about murder. Merely because some interpretation is possible, and perhaps minimally plausible[151], does not entail that it is correct or in line with the way participants thought prior to participating in the study.

---

[151] Since, after all, we might consider it possible but implausible to the point of dismissing the notion to interpret statements like "It is an objective fact that Los Angeles is north of Chile" as a mere expression of a person's emotion, though I'd contest even this: technically, such a remark could be used as e.g., a code phrase for expressing distress. My own view is that sentences like this don't mean anything outside a context of usage.

For instance, suppose participants were cognitivists about preference claims. Now they're being told that such claims don't have any cognitive content, not simply that this is a possibility. Yet Wright simply states there is "a reading of it in which [an instance of a preference claim] it is not truth-apt." Perhaps there is such a reading among academic philosophers, but these readings may conflict with how ordinary people think about such claims. There's a difference between the fact that some construal of what people mean is conceivable within a particular academic community and it *actually being the case* that a noncognitivist reading actually captures ordinary usage. Once you foist the latter claim onto participants, you are not merely providing them with a training exercise or a set of instructions. You are taking a substantive stance on metanormativity, and attempting to coerce participants into conforming to it. This is not an appropriate procedure if the intent of such instructions is to merely explain cognitivism and noncognitivism to participants. Imagine, for instance, a set of instructions that intended to clarify what moral realism was that simply asserted that "aesthetic claims have relative truth values but cannot be true or false in an objective sense" as though this were simply a fact about aesthetic claims. This would be inappropriate, since it would entangle simply clarifying what it would mean for some normative domain to be realist or antirealist with a substantive claim about which domains were in fact realist or antirealist. Wright's instructions are thus biased in favor of particular metanormative theory, and this bias is no small thing: the instructions don't simply favor a noncognitivist reading of preference claims, but outright assert that it is true in a way that implies no contrary positions or controversies could exist. Wright's elaboration on *why* these statements are best understood in noncognitive terms is also questionable. Wright states:

> Some people enjoy the taste of peanut butter ice cream, others don't. Some people have a great time riding the roller coaster at Elitch's, others don't. So, if one person said "Riding roller coasters is awesome!" and another person said "Riding roller coasters is absolutely terrifying!" it wouldn't make sense to say that one of the two was correct and the other mistaken. (p. 132)

First, it might make sense to say one person is correct and the other is mistaken, if you're a realist about these sorts of claims. So it's questionable whether researchers can just tell people that this wouldn't make sense. Second, as Kirwin (2021) argues, it is possible to be a realist *and* think that people with different preferences aren't necessarily mistaken. More generally, the notion that one person is correct and the other is mistaken can be understood in antirealist (but cognitivist) terms, anyway. For instance, even if we dismiss these as remote possibilities, subjectivism would remain a primary contender for a plausible account of the way ordinary people might use or understand preference claims. Subjectivism offers a cognitivist account of preference claims, yet this seems to be ruled out by these instructions. Notably, even a subjectivist could agree that there are meaningful respects in which one or another of two people who disagree about a matter of preference could be mistaken. In particular, people could be mistaken with respect to what they would themselves enjoy. In fact, this possibility is embedded into the very fabric of American culture: *Green Eggs and Ham*. In the story, Sam-I-Am approaches the unnamed antagonist and asks:

> *"Do you like green eggs and ham?"*

The antagonist responds:

> *"I do not like green eggs and ham. I do not like them, Sam-I-Am."*

The antagonist goes on to make it clear that there are no circumstances in which he would eat green eggs and ham. For instance, he would not eat them with a fox, and he would not eat them in a box. However, Sam-I-Am is persistent, and eventually persuades the antagonist to try green eggs and ham. After trying them, the antagonist announces:

> *"Say! I do like green eggs and ham! I do! I like them, Sam-I-am!"*

There is no presumption in this story that food preferences aren't solely determined by an individual's subjective preferences. Yet we learn that people can be mistaken about their own preferences. Thus, even if you are a subjectivist, you could think that if two people disagree, and someone denies that

"Riding roller coasters is awesome!" that person *could be mistaken* for the simple reason that *people can be mistaken about their own preferences.*[152] Indeed, that appears to be the whole point of *Green Eggs and Ham.* If even children are expected to understand this, it is bizarre that we would presume an adult would not, could not, think this way.

There are yet more problems with Wright's instructions. These particular instructions are intended to represent semantic nonfactualism, *not* psychological noncognitivism. Note that Wright describes psychological noncognitivism as the denial that "the mental states that moral statements are conventionally *intended* to convey are beliefs (or other related cognitive mental states)" (p. 131, emphasis mine). Semantic nonfactualism is explicitly *not* supposed to be about the mental states associated with moral claims. Yet Wright contradicts the distinction drawn between semantic and psychological noncognitivism in her instructions. Participants in the semantic nonfactualism condition are told that someone said the various examples of noncognitivist statements are not "*intended* to accurately reflect some fact about roller coaster riding—rather, they are expressions of people's liking/disliking of or approval/disapproval for something (in this case, riding roller coasters)" (p. 132). Later, participants are told that if someone said, "Peanut butter ice cream is disgusting," they'd "be stating something that is not truth-apt, since it is a statement *intended* to express her dislike of peanut butter ice cream (p. 132). Note the use of language about what moral statements are *intended* to convey. This looks like it's about the mental states of the people making these claims, and not about (or at

---

[152] We could even imagine a story that swaps out "green eggs and ham" for roller coasters:

Antagonist: "I do not like roller coasters! I would not, could not ride on one."
Sam-I-Am: "Would you, could you, in the sun? Would you, could you, just for fun?"
Antagonist: "I would not, could not, in the sun. I would not, could not, just for fun." [...]

Antagonist: "If you will let me be, I will ride one. You will see." [...]

Antagonist: "Say! I do like roller coasters! I do! I like them, Sam-I-Am! And I would ride one in the sun and I would ride one just for fun!"

least not just about) the semantics of their claims. If so, then it would appear that Wright's instructions conflate semantic and psychological noncognitivism.

People can also employ seemingly-propositional language to express non-propositional attitudes. Take, for instance, widespread use of "literally" as a tool of emphasis. If someone says, "It is *literally* a million degrees in here," this is not typically intended to convey a propositional claim that it is in fact a million degrees in here. Rather, it could be used to express the propositional claim that it is very hot. Yet it is not merely that such statements could be used to make hyperbolic but nevertheless propositional claims. Presenting this claim in both hyperbolic and superficially assertoric terms could be used as a means of emphasizing an emotional state, e.g., a negative attitude about the current temperature e.g., "This temperature? Ugh!" Take, for instance, the remark, "I *literally* hate *everything* about you" could be construed as a proposition, but it could just be used to convey a speaker's extremely negative attitude towards someone.

Every expression that isn't truth-apt doesn't need to convey a form of crude emotivism, either. Seemingly propositional claims could also be used as imperatives. Consider a parent saying to a teenager, "You will *not* go to the party tonight" Is this statement truth-apt? It has the *structure* of a truth-apt sentence. But it would be a mistake to interpret it as one. This would entail interpreting it as a truth-apt prediction about what will occur in the future, e.g., "I predict that you will not attend the party." Yet such an interpretation would be strained at best, and even ridiculous in most circumstances. Rather, the most plausible interpretation is an imperative (and perhaps a threat)[153].

Given these examples, it would appear that people use both sentences that are not explicitly truth-apt to make truth-apt claims, and to use seemingly truth-apt language to express various things

---

[153] It may even include propositional content and end up being truth-apt, since it could imply the threat of negative consequences, e.g., "if you attempt to go to the party, I will punish you." This would be truth-apt. However, such implication may not be intended, and the young man in question may not interpret it as a threat, either. Interestingly, if this is the case, the propositional content actually implied by the sentence would not appear in the sentence at all. The sente

that are not truth-apt. This is largely a result of the role pragmatics play in the meaning of ordinary utterances. Wright's attempt to draw a sharp dividing line between propositional and non-propositional claims by leaning on explicit terms such as "I" is a mistake. While it is plausible in some cases (e.g., "Meredith hates peanut butter ice cream" is mostly plausibly interpreted as truth-apt), it won't consistently hold, across different contexts and utterances, such as first-person expressions. Think of a child who yells "I *hate* you!" at their parents. Is the best interpretation of this simply that the child is reporting their mental states, and that one of these is that they hate their parents? This strikes me as, if anything, *less* plausible than interpreting this statement (at least in many contexts) as something closer to the emotivist's interpretation: the child is expressing an extreme negative emotion using superficially propositional language. At the very least, whether people doing this are only expressing an emotion or are also expressing a propositional claim *is an empirical question*; we cannot simply presume all meaning is carried exclusively by the semantic content of the utterance, and if we could, this would be news to philosophers working in metaethics: *all* noncognitivists are aware of the seemingly propositional nature of many moral utterances. They argue that such claims aren't truth-apt *in spite of this fact*. As such, they obviously think the superficial structure of a sentence can differ from what it actually means. In effect, Wright seems to want to offer participants a sort of general principle for knowing when an utterance is truth-apt or not that turns on the explicit semantic content of the utterance. Unfortunately, this just isn't how language works.

Next, let's look at the training exercise that followed these instructions (see **Table 3S.5**). Participants were asked to classify the following statements as truth-apt or not truth-apt. Recall that only participants who categorized at least 9 of these items proceeded to the next phase of the study:

**Table S3.5**

*Examples of truth-apt and not truth-apt statements from Wright (2018)*

| Not truth-apt | Truth-apt |
|---|---|
| Golden retrievers are better dogs than Chihuahuas. | Penguins are birds that can't fly. |
| Heavy metal music sucks! | Golden retrievers are bigger dogs than Chihuahuas. |
| Strawberries are tastier than raspberries. | Water is $H_2O$. |
| Abstract art is a waste of time and space. | Triangles are sturdier for construction (hold more weight) than squares. |
| Walking on the beach at sunset is relaxing. | Benjamin Franklin was the third president of the United States. |

There are many reasons to worry that this training exercise not only failed to instill competence with the distinction between cognitivism and noncognitivism in ordinary people, but may have actively misled or biased their responses to the questions presented after this exercise. All of the items classified as truth-apt are plausibly truth-apt. But note that they all concern similar subject matters: they include descriptive claims about biology, the composition of water, the engineering consequences of particular structures, and historical facts. All involve empirical claims that may be accessible from a third-person point of view, all of them involve facts about physical objections and their relation to one another, and none of them concern psychological states of any kind. Conversely, the statements that are not truth-apt all involve descriptions of psychological states. Note the language involved: "better," "sucks," "tastier," "waste of time and space," and "relaxing." All of these statements concern attitudes or preferences of some kind. Yet cognitivism and noncognitivism is *not* a distinction between psychological and non-psychological facts. Wright has set up a training exercise where these incidentally co-occur across all items. Note, in addition, that all of the statements classified as not

truth-apt convey the speaker's own evaluative standards *in the first person*. This allows the speaker to drop explicit references to themselves. Note Wright's examples:

> *"Meredith hates peanut butter ice cream."*  (truth-apt)
>
> *"I hate peanut butter ice cream"*  (truth-apt)
>
> *"Peanut butter ice cream is disgusting."*  (not truth-apt)

In the latter case, ice cream is shifted from the object to the subject of the sentence. This allows the sentence to drop explicit reference to the speaker, thereby allowing one to exclude any explicit reference to who is making the claim. More importantly, it adopts the structure of a sentence that *could* be interpreted as a claim about something other than the mental states of the speaker. Yet this does *not* entail that such statements aren't intended to convey the speaker's attitudes. Is it obvious, *merely in virtue of the structure of the sentence*, that "Peanut butter ice cream is disgusting" is intended to reflect some factual claim, rather than to convey a nonpropositional attitude? It's not obvious to me. You may have encountered a disgusting or horrifying image in your life, or seen someone else react to one, by declaring

> "That's *disgusting*!"

Note the structure of the utterance. Given Wright's criteria, this would have to be regarded as a truth-apt claim. And perhaps in some cases it is. Yet it strikes me as at least as plausible, if not more so, that such a remark could be used to convey an exclamation, e.g., an expression of disgust, horror, and surprise, but is not intended to convey any propositional claims. Simply put, people *could* use seemingly-propositional language to express nonpropositional attitudes. We are not entitled to simply presume that people mean to express a truth-apt utterance merely by looking at the surface structure of their remarks. Were this the case, noncognitivism about morality could never have gotten off the ground, since there is no controversy about whether people say things like "Murder is wrong," and do not merely say things like "I disapprove of murder."

Wright's emphasis on evaluative claims expressed in the first person as the sole representatives of nonpropositional claims on the one hand, and non-psychological assertions on the other, generates an artificial and misleading pattern of association between claims about mental state attribution and (non-mental) physical state attribution. There are no other relevant cues in this set of items for distinguishing what have been classified *a priori* as truth-apt or not (aside from the inappropriate use of "!" discussed below). They are grammatically similar, in that all adopt a *prima facie* assertoric surface structure (i.e., "P is Q" or "Ps are Qs"). So what considerations determine whether the psychological claims are not truth-apt but the physical claims are? It *could* be because the psychological claims are understood to convey nonpropositional claims, but the non-psychological physical claims aren't. However, Wright has stacked the deck in such a way so as to exclude anyone who doesn't endorse this, *whether or not* it was the case that people regarded psychological claims as uniformly noncognitive, by virtue of excluding anyone that didn't categorize them in this way. This, in effect, could train participants to associate statements to not be truth-apt whenever they are perceived to convey something about the speaker's psychological states. Suppose, for instance, participants were given statements like this:

*"Golden retrievers are more relaxed than chihuahuas."*

Presumably this would be classified as a truth-apt statement, since it would appear to express a claim about the comparative psychological disposition of different breeds of dog. If so, then such statements should have been included in the truth-apt category so as to not confuse participants into thinking truth-apt statements cannot refer to psychological claims. On the other hand, one could see this remark as an expression of a person's emotional or evaluative attitudes about golden retrievers and chihuahuas. If so, it might be classified as not being truth-apt. It's not entirely clear. Note that Wright uses "better" for the non-truth-apt remark comparing golden retrievers and chihuahuas, and "bigger" to reflect a truth-apt claim. In doing so, Wright is training participants to regard evaluative claims as

not being truth-apt! And it is not at all obvious that if a person said golden retrievers are "better" dogs than chihuahuas, that this is merely conveying the speaker's emotions. It could be understood as conveying the speaker's subjective attitudes, and be truth-apt in that respect. But it could also be understood to reflect a stance-independent fact of the matter about some presumptive standard. Suppose two people are discussing which type of dog would be best for families with young children. These families want a dog that is least likely to harm their children. One person suggests chihuahuas, due to their small size. Yet someone else, familiar with the typical behavioral dispositions of dogs, may believe a chihuahua would be more likely to act aggressively towards children, and that in spite of their size, a golden retriever would be more suitable. In response to the question of which dog is best, they might say, "golden retrievers are better dogs than chihuahuas." Yet such a remark does *not* merely express the speaker's nonpropositional attitudes. Instead, it conveys a propositional claim in response to a specific query. It would be best understood as something along the lines of "Golden retrievers are a more suitable choice of dog with respect to the goals and interests of most families with young children." In other words, the term "better" is often used to convey some standard, such as what is true or false relative to some goal, or with respect to some intersubjective shared standards between interlocutors. And this is clear in any context in which there is some shared goal between speakers, or when the goal in question is made explicit. Consider, for instance, a ship's captain and their first mate navigating at sea. They notice a dangerous reef to their left, and clear and safe waters to the right, and their destination is just up ahead, so neither direction will lead them astray:

First mate: *"Cap, betta turn 'ataway, 'void 'at reef there.'"*

Captain: *"Aye."*

Note that the first mate is expressing that going to the right is better than going to the left, with respect to the goal of not wrecking the ship and killing everyone on board, a goal that presumably the captain shares with the first mate. In such cases, evaluative language such as "better" can be understood to

reflect facts about what would or wouldn't be conducive to some goal or end. Such language is *not* merely non-propositional.

Thus, at least one way of using evaluative language in a way consistent with Wright's purportedly non-truth-apt examples is clearly consistent with making truth-apt claims. Stripped of context, this may not be obvious, but why should participants be expected to judge whether a statement is truth-apt or not by assessing it outside of its context of utterance? Wright, like many researchers, seems to treat remarks like those provided in the training exercise as having fixed meanings that don't depend on their context. Without context, it's unclear what participants are doing when responding to this training exercise. Are they *imagining* a context? Are they making some inference or inferences about the typical way they'd expect these utterances to be used? I'm not sure! And without knowing how they are responding to these scenarios, it's unclear whether this training approach is appropriate. It could be distorting how participants think about evaluative and non-evaluative utterances. For instance, it could cause them to spontaneously theorize about the meaning such utterances must have when decontextualized, in a way that doesn't reflect how they'd ordinarily interpret such remarks. That is, most people, in most everyday contexts, interpret what people mean *in situ*. Yet the training exercise Wright has subjected them to requires that they do so *ex situ*. It's unclear whether inferences made in the latter generalize to the former. If so, then people's judgments about the meaning of the toy phrases they are trained on may not be indicative of what people mean when using these phrases in everyday contexts. Such training has the added consequence of presuming that such statements would have a uniform and determinate meaning, i.e., that a sentence like "Strawberries are tastier than raspberries," is either *always* truth-apt or *never* truth-apt, rather than its meaning potentially varying by speaker or context. This is puzzling: Wright's training method seems to serve not simply to familiarize participants with the meaning of nonmoral utterances, and then use this acquired knowledge to assess moral utterances. Rather, Wright's training methods seem to

presuppose a variety of assumptions about the meaning of nonmoral utterances, and to induct participants into adopting or at least utilizing a way of drawing distinctions and making inferences that relies on a distinct set of presuppositions that, in effect, constitute a substantive stance on the philosophy of language. That is, participants *cannot* proceed with the study unless they "correctly" judge the utterances Wright provides in accordance with the instructions they were given, but these instructions presuppose that our utterances have uniform and determinate meanings, that utterances have distinct meanings *ex situ*, and that surface semantics fix the meaning of sentences rather than pragmatic considerations. By requiring participants to adopt a pattern of judgment that presupposes certain substantive empirical presuppositions about language and meaning, this training exercise isn't simply teaching participants so that they can provide their pretheoretical judgments in a different context, it is inadvertently saddling them with the researcher's own post-theoretical presuppositions.

This brings us to the second problem with the set of items Wright used in the training exercise. All of the statements in the not truth-apt category all convey only expressive content (e.g., pro/con attitudes), while none of the statements in the cognitivist category plausibly include any expressive content. Yet truth-apt statements are not necessarily empty of expressive content, nor do cognitivists necessarily expect or presume that they are when it comes to moral and other cognitive moral claims that involve normative or evaluative assertions. Take what Wright's criteria would classify as a truth-apt claim:

*"I really despise murder."*

Given Wright's instructions, this would be truth-apt, yet it unambiguously expresses an emotional attitude as well. Why not give *these* statements to participants in the training exercise? I suspect that doing so would cause many participants to hinge whether they judged statements as truth-apt or not on the artificial and inappropriate inclusion of explicit references to the speaker, such that, "I [...]" would prompt a truth-apt judgment, but its exclusion would not, even when the statements could be

plausibly understood in real world contexts to mean the same thing, with the "I" implicit and conveyed by the pragmatics of the utterance. Note, however, that Wright dropped "I think" or anything similar from statements in both the truth-apt and not truth-apt categories in the training exercise. This results in foisting interpretations onto the pragmatics of the sentences. Take these sentences:

*Strawberries are tastier than raspberries.*     Not truth-apt

*Penguins are birds that can't fly.*     Truth-apt

Each of these statements carries the implicit "I think…" Yet, as per Wright's initial instructions, we're supposed to interpret explicit use of "I think" as an indication that the statement in question is truth-apt. Imagine someone did say, "I think," before each of these statements:

*I think strawberries are tastier than raspberries.*

*I think penguins are birds that can't fly.*

Wright would have us interpret both statements as truth-apt, since both involve an attempt to report a fact about the speaker's mental states. Yet is this the most natural way to interpret these remarks? Imagine the following dialog:

Alex: *"I think penguins are birds that can fly."*

Sam: *"What?! That's ridiculous. You're completely wrong. Penguins can't fly!"*

If we interpreted Alex's remark as Wright would have us, this exchange would be somewhat puzzling. After all, isn't Alex just reporting a fact about what she thinks? *Maybe.* In some contexts, Alex might *merely* intend to report a fact about her own mental states. But Alex would probably not *merely* be doing this. Alex's expression would convey two distinct propositions:

(1) *That penguins are birds that can fly.*

(2) *That Alex believes this to be true.*

Typically, one would not need to explicitly assert both, and the purpose of making an assertion would in many cases be merely to convey (1), with (2) simply coming along for the ride as an implication of

asserting (1). This is because, in most ordinary contexts, asserting that "penguins are birds that can fly," carries the implication that the speaker believes this to be true. Yet it isn't typically the purpose of such remarks to describe one's mental states, but to make the claim itself (for whatever reason one might wish to do so). As such, assertions such as "Penguins are birds that can fly" can be used both to assert that this is true (one proposition) and to convey that the speaker believes this to be true (technically, a second proposition).Yet Wright treats evaluative or preference claims that do not *explicitly* include "I think" or some equivalent that explicitly includes the second proposition as having no propositional content at all, while adding "I think" renders such remarks propositional, not by making it clear that the remark is asserting some proposition about what is true independent of what is true about the mental states of the speaker, but instead to merely describe the mental states of the speaker. That is, to say "I think peanut butter ice cream is disgusting," given Wright's instructions, seems to imply that, on its own, to say that "peanut butter ice cream is disgusting," *just is* to express a nonpropositional attitude, and by explicitly adding something like "I think…" this remark becomes propositional in virtue of conveying a fact about the speaker's mental states. Yet in most contexts in which a person would say "peanut butter ice cream is disgusting," one would *also* be reporting their mental states by implication, in much the same way one is implying a report about one's mental states when saying "penguins are birds that can fly": Both carry an implicit "I think…" or some equivalent implicit indication of the speaker's mental states, or at least they *could* be intended to carry such implication, and interpreted to carry such implication. Wright's instructions present a distorted, awkward, unrealistic, and decontextualized conception of the way first-person evaluative utterances and conventional propositional claims work, in that the former are presumed to only convey emotions and not assertions about the speaker's mental states unless this is made explicit, even though the implication assertions about what is or isn't the case typically pragmatically imply some fact about the speaker's mental states which does not need to be made explicit since this would violate Gricean

norms of saying too much, i.e., the claim that "peanut butter ice cream is disgusting," doesn't need to include "I think" because this is superfluous; if one isn't speaking in the third person, *who else* would one be speaking about aside from themselves?

Since these considerations are excluded from both truth-apt and non-truth apt sentences in the exercise, what are participants supposed to do? *Not* think that when someone says "Penguins are birds that can't fly" that they're expressing what they think? Of course they think this! This results in a set of sentences that all bury "I think" in what is implicit (and pragmatically implied) in the statements. Yet Wright's initial instructions required that people interpret one class of sentences as not being truth-apt when this isn't explicit: statements that involve first-person expressions of one's preferences. Yet now they're presented not only with first-person expressions of preference, they are also presented with first person expressions of beliefs about matters of scientific and historical fact. They've been *instructed* that the former are not truth-apt and the latter as being truth-apt, in spite of their structural similarity. So now we're confronted with a scenario where first-person preference claims are only truth-apt if they explicitly include "I think…" or "I love…" or in some other way explicitly reference the speaker, while this is *not* true of sentences that do not express first person preference claims. In other words, participants are being told not only how to interpret the semantics of these sentences, but how to interpret the pragmatics of these sentences as well. Wright's instructions are not merely informing people about how language works. They're taking a substantive (and, I believe, highly questionable and probably descriptively inaccurate) stance on how people use language, which is the very thing the study is supposed to reveal, not presuppose!

After this, participants are presented with moral claims, e.g., "murder is wrong." Yet there is a serious problem: someone who thinks these statements are truth-apt (i.e., a cognitivist) *does* think that these statements *also* involve, or at least could involve first-person expressions of the speaker's evaluative attitudes. Consider a cognitivist interpretation of, e.g.:

*Murder is wrong.* | Truth-apt

The cognitivist may interpret this to mean a variety of truth-apt assertions, e.g., a relativist may interpret it to mean "Murder is inconsistent with my moral standards" or "I think murder is wrong," or they could be a realist, and interpret "Murder is wrong," to mean something like "it would violate the objective moral rules to commit murder." Yet such assertions are not mutually exclusive with, nor even in any tension with, this assertion *also* conveying the speaker's pro/con attitudes. That is, someone who said, "murder is wrong," could intend to convey both that it is a fact that murder is wrong, *and* to express their dislike of murder. By decoupling evaluative claims from non-evaluative descriptive claims, and providing no instances that could plausibly be used to express both, Wright is further stacking the deck in favor of noncognitivism.

Yet this is not made clear to participants, and it is not reasonable to expect them to exhibit the kind of sophistication necessary to understand this when categorizing moral claims. Whether they could do so isn't the central problem, however. The problem is that the training exercise was presented in a way that biases them in favor of noncognitivism. This is because the initial instructions and the subsequent exercises make it seem as though first person evaluative claims are not truth-apt *unless* they include explicit references to the speaker, e.g., "*I* love peanut butter ice cream," yet *this is the very thing that a cognitivist would dispute*. It's *just not true* that if someone says, "Peanut butter ice cream is delicious," that it is an uncontroversial fact, given the structure of the sentence, that it isn't truth-apt. Perhaps, absent instruction, ordinary people would interpret nonmoral first-person evaluative claims to pragmatically convey "I think," or "according to my subjective standards," in a way that *would* make them truth-apt. If so, then Wright's instructions aren't simply clarifying what these statements mean, and simply making it clear to participants how competent speakers *do* interpret such statements; rather, Wright would be *causing* people to interpret statements in this way, whether or not they did prior to participating in the study. That is, Wright would be causing participants to interpret first person

preference claims as being truth-apt when they include an explicit indexical, but not when they don't, *even if this is not how ordinary people think outside the context of the study*. And, once induced to think this way, this could push many participants to interpret moral claims as first-person preference claims, and thereby classify them as not truth-apt, simply in virtue of the instructions they were given. After all, even a cognitivist could agree that "murder is wrong" does, in part, express some nonpropositional content (e.g., disapproval or an imperative to not commit murder), in much the same way "Heavy metal music sucks!" does. It's just that they *also* think it expresses a truth-apt claim. Any participants led to believe claims that involve first-person expressions of one's evaluative attitudes are thereby not also making a truth-apt assertion could be misled by the instructions and training exercise for the simple reason that these instructions strongly imply that first-person expressions can't *also* be truth-apt. This, of course, is the very thing cognitivists dispute! In short, Wright's instructions seem to push participants towards drawing a dichotomy between first-person expressions of pro/con attitudes and truth-apt statements, and to regard the former as not being truth-apt. Yet since cognitivists are perfectly willing to regard moral statements as *both* first-person expressions of pro/con attitudes *and* truth-apt statements (indeed, some might think they *must* be both, or that ones that aren't both are somehow deficient or strange), these instructions bias participants in favor of noncognitivism. As a final example to illustrate this point, consider if participants were presented with a claim like the following:

*"Sam is a thief and a liar."*

This is expressed in third-personal terms, so it doesn't fit the pattern of statements Wright uses as examples of sentences that aren't truth-apt. Yet it also isn't a mere descriptive claim, since it plausibly conveys the speaker's pro/con attitudes. Items like this could make it clear that a statement can convey both, and still be truth-apt. Since participants were given no items that plausibly convey a speaker's pro/con attitudes, while remaining truth-apt, this creates the misleading impression that truth-apt

statements don't have nonpropositional content. Worse, however, is that noncognitivists don't simply maintain that statements like "murder is wrong," aren't truth-apt; they would also maintain that when moral claims explicitly borrow the structure of propositional (truth-apt) claims, they are *still* not truth-apt. This is because the meaning of these sentences is not born purely by their structure, but by the meaning *implicit* in the utterance. What Wright is essentially doing is instructing participants to regard statements that express emotions as not being truth-apt, then asking if the same is true of moral statements. Yet since it is reasonable to regard moral statements as having emotive content, this would incline many participants to judge moral statements to not be truth-apt whenever they interpret those statements to convey emotional content, but to not do so when they don't. Wright's instructions may therefore give participants the misleading impression that assertions that do not explicitly include reports about the mental states of the speaker but include evaluative or emotional content are not truth-apt, even though the inclusion evaluative and emotional content is consistent with cognitivism. That is, participants are given a training exercise where all items either convey nonpropositional content, and *must* be classified as not truth-apt, or they include no such content, and *must* be classified as truth-apt. Yet in practice, cognitivists tend to regard moral claims as assertions that express *both* nonpropositional content *and* propositional content.

Researchers are not entitled to presume that sentences with normative or evaluative implications are truth-apt or not truth-apt merely in virtue of their superficial grammatical or semantic structure. There is no serious dispute in descriptive metaethics about whether "murder is wrong," superficially appears to be a propositional claim; the only question is what people mean when they make such claims in practice. It is stacking the deck against the cognitivist to present first-person evaluative claims outside the moral domain, such as food and music preferences, as lacking propositional content.

Whatever the nuances of these instructions, *all* of the ostensibly non-truth-apt claims are claims about preferences, and *all* of the ostensibly truth-apt claims are not. This generates an artificial pattern of association between truth-aptness and assertions about what the world is like on the one hand, and first-person evaluative utterances as merely expressions of nonpropositional attitudes on the other.[154]

Another minor issue is the inclusion of an exclamation point in one of the non-truth-apt items: "Heavy metal music sucks!" The use of an exclamation could inappropriately convey added emotional content behind the remark, which could inflate categorizing such remarks as noncognitive. An exclamation *could* be added to the truth-apt assertions as well, yet it would not render these statements no longer truth-apt. The use of exclamation marks further reinforces the misleading impression that anything asserted with an emotional valence to it isn't truth-apt.

Next, note that 23% of the participants failed to classify the items in accordance with Wright's instructions. This could introduce additional methodological worries. While the purpose of excluding participants who failed at this task is, ostensibly, to eliminate analysis of responses that could be attributed to inadequate attention or competence with terms and concepts relevant to the study. For instance, if you wanted to estimate the proportion of people who believe in God, and you asked people whether they believed "theism is true," you would not want to include responses from people who did not know what the word "theism" meant, since their responses would not tell you whether they believed in God. You'd simply be left with a noisy estimate. Yet excluding participants from analysis based on the procedures used in the study risks biasing subsequent results, since such exclusions may

---

[154] In addition, note that semantic cognitivism is misdescribed. It excludes deflationary accounts of truth. By their own admission, Pölzler and Wright have been unable to determine whether people have any determinate notion of truth, and if so, whether it corresponds to the correspondence theory of truth (2020b; Patterson, 2003). Since ordinary people may have indeterminate or variable conceptions of truth, or may reject a truth correspondence theory of truth, another element in establishing folk cognitivism or noncognitivism would involve ensuring that participants also have the appropriate stances or commitments towards other philosophical issues presupposed by and embedded in stimuli used to assess their views towards cognitivism and noncognitivism.

inappropriately exclude legitimate responses. Suppose I wanted to know whether people preferred classical music over heavy metal, but I insisted on excluding anyone who played the electric guitar. This would obviously bias results against people with a preference for metal.

Montgomery, Nyhan, and Torres (2018) emphasize this by noting that excluding participants in this way can lead to *nonrandom attrition*. What if, in Wright's studies, participants who disproportionately failed the training exercise did so because they were warring against their cognitivist intuitions about nonmoral preference claims? That is, suppose you're a cognitivist about all normative and evaluative claims. You are then asked to participate in a study that requires you to judge such claims as not being truth-apt, even though you don't think this is the case, or you aren't intuitively disposed to judge such claims in this way. If so, may *disagree* with the classification scheme mandated by Wright, or simply be at greater risk of confusion or performance error, than someone who is more receptive to the noncognitivist depiction of such claims. If so, you'd be more likely to be excluded from subsequent analysis. Yet suppose those who are cognitivists about preference claims are also more likely to be cognitivists about moral claims. By disproportionately excluding such participants from analysis, the resulting pool of respondents may be skewed towards greater noncognitivism than in the absence of such exclusions. In other words, you'd be curating a pool of people predisposed to adopt one position rather than another when the very task is to estimate the proportion of people who endorse each. For comparison, suppose there is a correlation between believing in psychic powers and believing in ghosts. A researcher runs a study with the goal of estimating the proportion of people who believe in ghosts. However, as one of their exclusion criteria, they eliminate any participants who claim to believe in psychic powers. The resulting pool of participants would consist, not of random

members of the population of interest, but a *nonrandom subset* that exhibited certain, distinct characteristics that rendered them unrepresentative of the target population.[155]

Finally, note that Wright asked participants to make the same judgment for each item. A considerable majority judged some moral issues to be truth-apt, and others not to be. They appear to be metaethical pluralists, while only 24% consistently favored cognitivism or non-cognitivism for all moral items. While this does provide support for metaethical pluralism, it *is* a puzzling finding. Note that in Wright's instructions, participants were expected to classify all first-person evaluative claims as not truth-apt, even if conveyed in the indicative mood. Note the operative term here: *all.* That is, the instructions suggested that all statements of a particular form were to be interpreted in the same, uniform way. And yet when participants were asked to evaluate moral claims, the majority did not do so in a consistent way. While there is no logical inconsistency with doing so, nor any obvious philosophical mistake in thinking that some moral utterances are truth-apt and others are not, we may still reasonably wonder *why* people would judge some moral claims to be truth-apt and others not to be. This could reflect a genuinely pluralist metaethical position. But it could just as readily reflect a sensitivity to subtle sensitivity to perceived differences in the likely meaning of such remarks for reasons unrelated to the metanormative properties of moral claims as a category. Note that participants were asked, for each item, "It is wrong to *x*," where *x* was one of the moral actions. Participants judged items such as the following:

> *It is wrong to sell children on the internet.*
>
> *It is wrong to watch pornographic videos.*
>
> *It is wrong to cheat on an exam.*

---

[155] Out of the twelve items most participants categorized as moral, a majority of participants judged four to be not truth-apt, and about an even number of participants judged five of the moral items to be truth-apt, with only three of the twelve items commanding a majority of truth-apt classifications. This strikes me as a fairly high proportion of noncognitivist responses, but it seems fairly consistent with studies that include a noncognitivist option.

Note, firstly, that these are all expressed in the indicative mood, yet they exhibit a structure quite unlike the structure of the items used in the training exercise. Several of those items involved comparisons between one thing and another, and most did not even include the term "is." Why were participants trained with a varied set of items, then presented with a set of utterances with a uniform structure? I'm not sure, but this may have influenced their judgments in unintended ways. Another difference is that most of the items in the training exercise that were not truth-apt involved richer and more personal evaluative concepts: "sucks," "tastier," waste of time and space," and "relaxing," with the sole exception being the generic "better." Yet these items simply present the notion of something being "wrong." One potential biasing factor with the training items is the use of psychologically richer and more personal content; "wrong," like "better," is generic, and may have failed to readily prompt the perception of emotional attitudes behind the remarks. Furthermore, the term "wrong," is polysemous, and is also used to convey that something was a mistake or error, such as when an answer on a test is marked "wrong." If ordinary people lack the training and sophistication to respond to the term "wrong," in context-sensitive ways, this could enhance the rate of performance errors. That is, because the English language uses the term "wrong," in both a moral context, and in a nonmoral context that is unambiguously truth-apt, inadequately trained participants may struggle with the cognitive load of drawing this distinction despite the use of an identical term.

Yet the issue I want to draw attention to is the notion that participants judged *some* of the statements above to be truth-apt, and others not to be. Again, this *could* reflect different metaethical standards towards different moral issues.[156] But is this the most plausible reason why people would categorize some moral claims as truth-apt and others not? Perhaps not. Consider how *bizarre* the task in question is. The participant is presented with a whole host of sentences:

---

[156] or, more aptly, different metanormative standards towards different normative issues; since these items were only *dominantly* classified as moral issues, some participants who judged them to be truth-apt or not truth-apt may not regard them as moral issues.

*It is wrong to sell children on the internet.*

*It is wrong to watch pornographic videos.*

*It is wrong to cheat on an exam.*

…and their task is to judge whether such sentences are truth-apt or not. For most people, some are judged as truth-apt, and some are not. The metaethical pluralist would have us interpret this as a fairly sophisticated, if implicit, philosophical position. For an ordinary person to be a metaethical pluralist, they would have to recognize all such utterances to belong to the same normative domain, i.e., the moral domain. If they don't, then they are not pluralists proper. After all, if some of these issues aren't moral issues, then the participant can't be a *moral* cognitivist or noncognitivist about them. And if the participant does not consider or treat these sentences as belonging to the same category, then it's unclear in what respect they'd be a pluralist. A pluralist is, by definition, someone who adopts a different metanormative stance towards two or more normative issues *in the same domain*. Already, then, we have to accept at face value that participants treat all of these issues as belonging to the same domain. Wright is aware of this, and goes much further than most researchers by asking participants to classify issues as social, moral, or personal.

But this is already a strange task. What, exactly, does it even mean? What moral issues aren't also either social or personal? What does it mean to say that an issue is "personal" rather than "moral"? And if a participant judges an issue to be "personal" rather than "moral," what do *they* think this means? Do different participants draw the distinction based on the same conception of what the distinction entails? I doubt it. This is far from a clear and well-defined task. It's not quite the same as sorting objects as red or blue, or deciding flavors one likes or dislikes. The participant is asked to put a host of claims into categories: personal, social, moral, but again, *what does that even mean*? Why should we assume this is an especially meaningful task, and that the proportion who fall into each category for every item are especially reflective of e.g., whether participants treat the issue in question as a

member of the "moral domain"? While Wright, myself, and others may have a sense that there is something distinct and meaningful about what it means to belong to the moral domain, it's not clear that this notion (or notions, as our own positions probably do not perfectly overlap) is shared by ordinary people. Consider some of the results: Only 36% of participants considered getting an abortion in the first trimester to be a moral issue, while 63% judged it to be a personal issue and 1% considered it a social issue. What did these people have in mind when judging it to not be a moral issue? I don't know (and more importantly, neither does anyone else). But there is a significant conceptual difference between thinking that abortions are morally permissible, but that whether you get one or not is a personal choice and thinking that the decision to get an abortion simply isn't the kind of consideration that one should consider a matter of moral concern at all. One *could* think this, and perhaps some people do, but in the context of a society in which *every* participant is aware that many people consider abortion immoral, it would be strange to think that the question of abortion isn't even a legitimate moral question, rather than one for which there is a clear and definitive answer. Without knowing more about what participants take their categorization of issues to mean, it's hard to know how to interpret the results of this task. Yet there are other puzzling results. 17% of participants judged "taking things that don't belong to you" and "forcing someone else to have sex" to not be moral issues at all. I don't know how you interpret these phrases, but I take them to refer to stealing and rape. In what possible universe would one out of five people not consider stealing and rape to moral issues? These are paradigmatic instances of moral transgressions. If nearly 20% of your participants aren't classifying these items as moral issues, it is more likely that there is something wrong with your task than that there is something wrong with your participants. What is taken to be variation in participant's views about whether issues are moral or not, where this is understood to mean something specific and meaningful to researchers, may be better explained by variation in *how participants are interpreting the items and the task in general*. Note, for instance, that 8% of participants did

not judge selling children on the internet to be a moral issue. *What were they thinking?* Are we to imagine that these people do not find child trafficking morally objectionable? I'm not prepared to do that, and I don't think researchers should be, either. I find it far more likely that they did not interpret the task or the item (or both) the way I did. And the same could hold for other responses.

This categorization task feeds into the broader issue of understanding just what participants are thinking when responding to the tasks in this study. Returning to the issue of classifying stating that "It is wrong" to engage in these tasks as truth-apt or not, what are we to make of the extraordinary intrapersonal variation across items? Why would people think "It is wrong to…" is sometimes truth-apt, and sometimes not truth-apt? One key problem with these statements is that they are *not real moral utterances.* They are not made by an actual person in a real-world context, with all the rich contextual information and auxiliary details that would be available under such circumstances. Every actual moral judgment, when expressed, is expressed with the intent to achieve one or more goals by the speaker. Assumptions about those goals, the motivations of the speaker, who they are speaking to, and a vast array of other factors are all relevant to assessing what that person means in that particular case. But this is not what we are given. We are given with impoverished, decontextualized "moral utterances" that aren't made by anyone in particular in any particular circumstance. To highlight just how strange it is to assess the meaning of such remarks, extracted from the appropriate ecological conditions in which such utterances actually occur, consider what it would be like to ask someone this:

*Alex threw a rock at someone. Why did Alex throw the rock?*

This question is *unanswerable.* There simply isn't enough information to know why Alex threw the rock. There are many possibilities. Perhaps Alex is trying to injure the other person, and it is an act of aggression. Or perhaps Alex is hiding, keeping an eye on a fellow soldier, and is trying to surreptitiously warn them of an imminent ambush. Or maybe Alex is infiltrating an enemy compound and is throwing

a rock to create a sound that will distract guards. The possibilities are plentiferous,[157] and participants simply cannot know which of these possibilities is the case for Alex. And there's a good reason for this: there is no fact of the matter. This is an entirely imaginary example. Alex doesn't exist, and no rocks were thrown. Any inferences about Alex's intent would require the participant to "fill in the blank" by *imagining* a context. How participants fill in the blanks may involve a variety of psychological processes.[158] Perhaps participants imagine the most typical, or most salient situation in which a particular action or utterance would occur. This could involve a panoply of heuristics, stereotypes, or schemas. Each individual may draw on relevant cultural knowledge and other background assumptions that could vary between participants, and across different imaginary situations. To a soldier, throwing rocks could be a common occurrence in guerilla warfare, and this might be where the mind goes. A child may think of games. A police officer might think of rioters. A fisher may think of people skipping stones across the water. Each of us may be primed to imagine a different context.

Even if most people imagined a similar context, it might still be based on various shared stereotypes and schemas distinct to that particular population. Researchers who present participants with toy sentences stripped of all context foist the burden of filling in the necessary context themselves, prompting participants to rely on availability, salience, and other heuristical tools to provide meaningful responses. The result may be that substantial interpretative variation, both within

---

[157] I am deliberately using a fake word to illustrate a point: despite not even being a real world, I am confident readers have little trouble understanding what I mean: that there are *many* possibilities. We don't need specific words to convey what we mean. Background assumptions and surrounding context do the heavy lifting, and the meaning of any particular word is little more than one tiny cog. We can express what we mean even with a few damaged cogs or worn out springs. Just the same, interpreting what a person means when making a moral claim is mostly inferred holistically. Philosophers mistakenly focus too much on a blinkered assessment of the semantics of isolated and decontextualized sentences. This is a mistake that has unfortunately been recapitulated in the instructions and stimuli used in research on the psychology of folk metaethics.

[158] The fact that *I* don't know what exact processes involved is hardly a problem for me: neither does Wright, or anyone else conducting this research, and the likely candidates for what is going on aren't promising candidates for vindicating the validity of the typical procedures used in these studies. That is, if we don't know what processes participants are employing, this is one reason to doubt the results of these studies. But if we do eventually find out what is going on, it's not likely to support the validity of these studies, anyway. If, for instance, participants are relying on a typical context for assessing each claim, such typicality may reflect culturally contingent assumptions about what people in one's societies would likely intend with a particular remark, *not* that remarks about that particular issue are consistently truth-apt or not.

and between populations, could drive much of the variation across responses, rather than variation with respect to the task in question. That is, participants who would in principle provide the same response to a particular set of stimuli *if* they interpreted that stimuli in the same way as someone else interpret it differently *because* they interpret it differently than others. For comparison, imagine we asked people to think about their "favorite food." Each of us is going to think of something different, and there would be population level differences in the typical foods people imagined, as well. People in Finland are not going to tend to imagine the same kinds of food as people from India. While asking people about their favorite food is an extreme example of interpretative variation, it illustrates one end of the spectrum of specificity. At the other end, we'd have a realistic situation with all the relevant details fleshed out in such a way that interpretative variation was minimized. Achieving this ideal is, in practice, too demanding. Participants cannot be expected to read thirty pages of detailed descriptions, and there'd be significant trade-offs that would make the cure worse than the disease: participants would be unable to remember all the details, would get bored or angry, and would be exposed to many opportunities to interpret details in unintended ways (whether due to performance error, ambiguity, or some other cause). We have to compromise by using stimuli that fall somewhere between these two extremes. Unfortunately, researchers have leaned too far in the direction of impoverished stimuli. Such stimuli *may* work, but unless we supplement these measures with additional measures designed to assess interpretative variation, or even methods for mitigating it, researchers will in many cases be flying blind, unable to distinguish with confidence patterns in their data that result from genuine variation with respect to the construct of interest, or superficial variation due to differences in how people interpret the stimuli.

### S3.10.1.5 Psychological noncognitivism

Psychological noncognitivism is the claim that moral utterances don't express beliefs, i.e., mental states about what is true or false.[159] Wright recruited 116 participants for this study and gave them an extraordinarily long set of instructions. I'm hesitant to share a quote this long, but it's important for anyone assessing the validity of these instructions to see them. I suspect the sheer length of the instructions is adequate, on its own, to raise doubts about how plausible it is that participants could really have internalized all of this and understood the relevant concepts:

> People make different kinds of statements—some of which assert beliefs, others of which express feelings. Consider, for example, if I said to someone that "Boston, MA is north of Miami, FL". What I am doing is expressing my belief that something is the case—namely, that there is a fact of the matter about the geographical relationship between Boston and Miami. My intention is to assert a belief, which in this case turns out to be true. But, there are also times when the beliefs we assert with our statements are false, like if I would have said "Boston, MA is south of Miami, FL" instead. But that does not change the fact that such statements assert a belief about something being the case. For our purposes, it doesn't matter whether the beliefs being asserted are true or false—all that matters is that we sometimes make statements that are intended to assert beliefs about things that we take to be matters of fact about the world.

> The same goes for statements that involve beliefs whose truth/falsity cannot be established. For example, I might state something like, "The earth is the only planet in our galaxy with life on it". This isn't the sort of belief that can currently be established as true or false—we don't know at this point (and, indeed, we may never know) whether my belief accurately reflects a fact of the matter about life in the galaxy or not. But, nonetheless, my objective in making this statement is to assert a belief about something I take to be true, even if it can't be established for sure whether or not I'm correct.

> Consider, on the other hand, my statement that "Peanut butter ice cream is delicious!" or "Jazz music is the best form of music ever invented" or "The roller coaster at Elitch's is terrifying!" Here, these statements are not intended to be assertions of beliefs about matters of fact—i.e., that peanut butter ice cream is the sort of thing that is, in fact, delicious or that riding the roller coaster at Elitch's is the sort of activity that is terrifying. Rather, they are expressions of positive and/or negative feelings and attitudes that I have about the subject matter (in this case, really liking peanut butter ice cream and not liking the roller coaster at Elitch's).

---

[159] Pölzler (2018b) has already raised a handful of objections to this particular study as well.

When I make these sorts of statements, I am fully aware that they aren't true or false (like the statements considered above). While it may be true that I like the taste of peanut butter ice cream and don't enjoy riding on the roller coaster at Elitch's, there isn't actually a fact of the matter about peanut butter ice cream being delicious or the roller coaster being terrifying— after all, it would make perfect sense for someone to reasonably state the opposite and neither of us would be mistaken. In other words, the objective of statements like "Riding on roller coasters is terrifying" or "Peanut butter ice cream is delicious" is to express our positive/ negative feelings (pro/con attitudes, liking/disliking, approval/disapproval) about something, not to assert beliefs about things that we take to be true.

Of course, I can believe (i.e., take it to be true) that I or someone else really likes peanut butter ice cream and doesn't like riding the roller coaster at Elitch's and my statements can assert such beliefs—such as if, for example, I were to say "Meredith really loves peanut butter ice cream" or "Peanut butter ice cream is my favorite". These statements involve beliefs about Meredith and myself that are either true or false. But statements like "Peanut butter ice cream is disgusting!", on the other hand, are not.

> To further illustrate, consider the following two statements:
>
> • Larry loves Bon Jovi
>
> • Bon Jovi rocks!

The first statement involves the assertion of a belief about Larry (namely, that he loves Bon Jovi—which could be true or false); the second, on the other hand, does not assert a belief (there is no fact of the matter about Bon Jovi "rocking" that can be established as true or false) but instead expresses a person's positive attitude (their appreciation, enjoyment, approval) toward Bon Jovi.

For the questions that follow, please keep this distinction in mind, as you'll be asked to identify those statements you think were intended to assert beliefs about matters of fact, those intended to express positive/negative feelings, attitudes, etc. about a topic, and those intended to do both.

That's *really* long. Once again, participants were given the same set of ten items that were used before, which they were forced to "correctly" categorize. 16% failed to do so. This isn't as bad as the previous study, but it's still a troubling proportion. Next, they categorized a set of 20 items as moral or nonmoral, and were then asked whether each sentence was intended to:

- Assert *beliefs* about matters of fact

- Express positive/negative feelings, attitudes, etc.

- Both

For instance, a participant may be asked about the sentence "It is wrong to cheat on an exam." First, I personally struggle to make sense of this task. How can any of these *sentences* be *intended* to assert a belief or a feeling? Sentences don't intend anything. *People* intend things. Presumably, the questions are asking about what people would mean *if* they said these sorts of things. But why think there's any single fact of the matter? Why wouldn't it depend on what that person meant, and what a person saying something intended to communicate could vary from person to person. This is certainly how I think language works, but it's not, as far as I can tell, how many academic philosophers seem to think language works. Already, we can see that the very way these questions are framed makes certain contestable philosophical assumptions about the way language works (assumptions that I incidentally reject, rendering me incapable of answering these questions since I don't have the response option "there is no way to answer this question").

Also, note that utterances can do more than assert feelings or express facts. Wright excludes the ability to express imperatives, even though this represents another form of noncognitivism (van Roojen, 2018). There are also other forms of expressivism, quasi-realism, and hybrid accounts that may reject the crude emotivism entailed by Wright's measures. None of these response options are available to participants. And the response option beliefs *about matters of fact* may conflate cognitivism with realism, whatever the instructions might say. This is exacerbated by the instructions, which state that "all that matters is that we sometimes make statements that are intended to assert beliefs about things that we take to be matters of fact about the world" (p. 135). About *the world*? Technically, a subjectivist thinks moral facts are facts about our personal moral standards. Are these standards facts "about the world"? I'm inclined to think so, but ordinary people may not. They may instead think

facts about the world can't be reducible to people's mental states in this way, and may instead associate such facts with scientific or discoverable facts, not private mental states.

Yet one of the more serious problems is that the examples of noncognitivism that participants are given are, at best, controversial, e.g., "Peanut butter ice cream is delicious!" Why are we told that such a statement isn't intended to be an assertion of belief? There may be view proponents of Loeb's (2003) *gastronomic realism*, but subjectivists would regard such a statement as being true or false. As such, Wright's examples rely on a controversial claim about the meaning of such utterances. We're also given, once again, the weird case of some declarative sentences being declared nonpropositional but others being treated as propositional. Why is "Peanut butter ice cream is delicious!" only intended to express an emotion, while "Peanut butter ice cream is my favorite" isn't? The latter will, in many contexts, be interchangeable with the former. If someone took a bite of ice cream, not knowing the flavor, and discovered it was their favorite flavor, they might exclaim, "Peanut butter ice cream? This is my favorite!" This seems like an equally good candidate for a nonpropositional assertion as "Peanut butter ice cream is delicious." Wright's examples seem somewhat arbitrary and forced. For instance, statements about evaluative judgments (e.g. "delicious") are treated as nonpropositional when they don't explicitly reference the speaker, but they are propositional if they explicitly refer to the speaker or someone else. This is a questionable criterion to employ, and many philosophers may reject it. It seems strange, then, to employ instructions that take a substantive philosophical stance, and even require participants to conform to it. This is even more of a problem when participants are required to categorize a set of items "correctly," where "correct" just means "in accordance with the criteria provided in the instructions." I don't accept these criteria, so I would "fail" these "comprehension" checks, and be excluded from analysis for allegedly being too incompetent to engage with the stimuli. If researchers with specialized training would fail your comprehension checks because they disagree with you, there's probably something wrong with your comprehension checks.

Unfortunately, these instructions are not adequate, nor are the training exercises that accompany them. However, Pölzler and Wright (2020a; 2020b) have devised new instructions and training exercises, along with an assortment of novel paradigms, that avoid at least some of the shortcomings of these studies.[160]

### S3.10.2 Pölzler & Wright's Training Paradigms

Pölzler & Wright (2020a; 2020b) describe, over two papers, what are without a doubt the most sophisticated and well-designed folk metaethics studies to date. I feel some remorse in directing criticism at these studies. I confess to a degree of envy at the creative paradigms they devised. They are innovative, thoughtful, and go a long way in circumventing the shortcomings that appeared in previous scales.

Nevertheless, I am not convinced that any of these measures succeed where previous measures failed. My primary objection appears in **Chapter 3**, where I argue that extensive instructions and detailed response options risk training participants so much so that they are no longer ordinary people, which threatens the external validity of the findings. Here, I want to assess the specific measures that were used. Measures were divided into abstract and concrete measures. Abstract measures ask about morality in general, while concrete measures address specific moral issues (e.g., abortion). This leaves us with a total of seven measures:

*Abstract measures*
1. Theory task
2. Metaphor task
3. Comparison task

---

[160] Pölzler (2018b) also raises some novel objections to Wright (2018). Pölzler claims that Wright's findings "may have biased subjects towards non-cognitivism" (p. 661). This is because many of the examples used to illustrate noncognitivism are "practically normative," that is, they are about "the goodness of actions" rather than "the goodness of beliefs" or "about descriptive facts" (p. 661). Pölzler claims that this could have prompted participants to incorrectly associate noncognitivism with practical normativity, which could have inflated the degree to which they likewise regarded moral claims as not being truth-apt. Pölzler also points out that, by drawing attention to the existence of disagreement about matters of taste, Wright may have given the impression that disagreement implies noncognitivism, even though this isn't true (or, at the very least, this would be a controversial stance to take, and would not be appropriate to include in instructions that are intended to be theoretically neutral).

4. Disagreement task
5. Truth-aptness task

*Concrete measures*
1. Disagreement task
2. Truth-aptness task

### S3.10.2.1 Abstract theory task

This task begins with an explanation of the central questions in dispute in contemporary metaethics:

*(1) Do moral sentences intend to state moral facts?*

*(2) If yes, do these facts exist?*

*(3) And if yes, are they independent from what anybody thinks about them?*
(Adapted from Pölzler & Wright, 2020b, p. 60)

Note the similarity to Davis's (2021) flowchart method. Here, participants are not asked to proceed through a flowchart, but are given a description of metaethics that characterizes the central issues in a flowchart-like fashion. Note that there is nothing theoretically neutral about this. If one's goal were to prompt spontaneous theorizing, recapitulating the central disputes in contemporary metaethics in this way is precisely what you'd want to do.

Note the first question: moral sentences are endowed with intentions, as though sentences are intended to mean things, rather than the people uttering those sentences. This is already a point of contention that bakes an orthodox philosophical view in the philosophy of language of how language works, one I happen to reject. I don't think it makes any sense to speak of the intended meaning of a sentence. Sentences don't intend anything, *people* do. Granted, ascribing intentions to sentences may be a loose or nonliteral way of speaking, and what this is intended to convey is a question about what people who employ sentences mean. Yet if this is what's meant, why not say this? Why couch it in metaphors?

Many philosophers (perhaps most, though I'm not aware of any data on the matter) would think that sentences do have meanings that are determined at least partially by factors independent of

the speaker. Note, as well, that this set of questions presumes the uniformity and determinacy of folk moral sentences, which is an odd presumption to make if the researchers conducting the studies think most ordinary people are metaethical pluralists. This is a bit like asking people whether "fruit is good," despite a wealth of data indicating that most people think some fruit is good and some isn't. It's bizarre to simultaneously think that ordinary people's metaethical commitments incline them towards realism for some moral issues and antirealism for others, or that people's explicit stance is that the meaning of moral sentences varies, yet still present the issue in the way philosophers have traditionally framed the matter, such that all moral sentences share the same metaethical presuppositions. These instructions also strongly indicate that folk semantics is deeply relevant to metaethics, even though this is likewise a questionable philosophical thesis. Both of these assumptions are mirrored in the response options, which *require* a participant to select a response option that pairs a semantic thesis with a metaphysical thesis. I'm an indeterminist. I reject all the semantic theses. I have no way to respond to this question. Likewise, Kahane (2013), who rejects the notion that moral realism requires a semantic claim, may lack any meaningful way to respond. While we may be in the minority, it's not appropriate for researchers to simply presume ordinary people *must* think in accordance with mainstream academic philosophy. Once instructions were given, participants were asked to choose from among the following response options:

> *[SECULAR REALISM]*
> When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they are independent from what anybody thinks about them. For example, an action that is morally wrong is wrong no matter what anyone thinks. So it would still be wrong even if you yourself, or the majority of the members of your culture, thought that it is not morally wrong.

> *[THEIST REALISM]*
> When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they depend on God's will. For example, an action is only morally wrong if God forbids us to perform the action. If God did not forbid us performing the action, it would not be wrong.

*[CULTURAL RELATIVISM]*
When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they depend on what the majority of the members of her culture think about them. For example, an action is only morally wrong if the majority of the members of your culture believe that it is wrong. If the majority of the members of her culture did not believe the action to be wrong, it would not be wrong.

*[INDIVIDUAL SUBJECTIVISM]*
When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they depend on what individuals think about them. For example, an action is only morally wrong if you yourself believe that it is morally wrong. If you did not believe the action to be wrong, it would not be wrong.

*[ERROR THEORY]*
When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts do not exist. Thus, it is never the case that something is morally right or wrong, good or bad, etc. No such moral statement can be true.

*[NON-COGNITIVISM]*
When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts do not exist. Thus, it is never the case that something is morally right or wrong, good or bad, etc. No such moral statement can be true.
When a person says that something is morally right or wrong, good or bad, etc. she does not intend to state a fact. Instead, she intends to communicate/express her feelings, emotions, intentions or attitudes about it. For example, by saying that some-thing is wrong, you only express feelings of disapproval towards it (and that is the only thing you are doing). Moreover, there are no facts about what is morally right or wrong, good or bad, etc.

These are shockingly long and complicated response options. I have little confidence ordinary people could understand even half of what's stated in each of these. Note, as well, that participants are asked to theorize about *what other people mean when they make moral claims*. This is *not* a measure of the participant's own commitments, but a measure of their theory *about* how people speak and think, i.e., their *metalinguistic* intuitions. It remains an open question how well ordinary people's metalinguistic intuitions accurately capture actual usage (Martí, 2009). Such judgments may or may not be reliable,

and even if they're reliable with respect to some features of the way people speak, this may not generalize to others.

There are a variety of oddities and complications with these response options, all of which could throw off participants in ways that would not be readily captured by simple comprehension checks. However, these concerns are mostly minor nitpicks. The secular realism condition tells participants that moral facts "exist," yet it's unclear what it would mean for them to exist. Note, as well, that ordinary people may not appreciate the instructions as indicating stance-independence in the sense entailed by realism. People could conflate stance-independence in the realist sense with constructivist notions, which entail that there are procedures we could employ for constructing moral rules and institutions; such rules would not be *directly* dependent on our standards, or the standards of any particular person or culture, but indirectly dependent. It's not clear ordinary people would appreciate that this would still be considered a type of antirealism. And such a conflation is not without precedent. When asked about moral objectivity, findings in **Chapter 4** showed people asked about moral objectivity often make reference to whether a question can be resolved by some publicly available measure that, if we agree to use it, would furnish us with some quantifiable or at least non-subjective standard. This is a bit of a stretch, though. It may be that while *some* people conflate the secular realist question with some type of constructivism, it's not clear this would be a frequent enough concern to undermine the validity of the study. My more general concern is that the set of response options here are simply too complicated, and that people cannot be expected to have a robust understanding of stance-independence, such that they could appropriately respond to the question.

The theist realism condition has an ambiguity. It states, "When a person says that something is morally right or wrong, good or bad, etc. she intends to state a fact. Such facts exist – and they depend on God's will." Participants may interpret the latter part of this description to suggest that people intend to state facts about what's consistent with God's will. And if people don't think this is

the case, they may reject this option as a result of an unintended interpretation. In other words, this item does not make clear whether people are trying to refer to moral facts, without explicitly trying to refer to God's will, or whether they are trying to refer to facts about God's will.

The cultural relativism condition states that moral facts "depend on what the majority of the members of her culture think about them" (p. 60). It's not clear this is a necessary condition for cultural relativism. Moral facts may be determined by authorities within a community, or some more complicated considerations, not necessarily a simple majority. The error theory condition is extremely complicated, and I have serious doubts most people would understand it. And the noncognitivism condition could confuse participants who recognize that moral claims may be used to express emotions *and* state facts. While the response says, "instead," this only implies, indirectly, that the response option in question treats claims as nonpropositional. It also pragmatically implies that if moral claims are intended to state facts, that they aren't also used to express emotions, even though moral claims expressing both facts and feelings are consistent with cognitivism. This could have inappropriately inflated noncognitivist responses.

### S3.10.2.2 Abstract metaphor task

The next task presents people with the following instructions, with accompanying response options:

> This task is about moral facts. Moral facts are facts about what is morally right or wrong, good or bad, virtuous or vicious, and so on. For example, it could be a moral fact that it is (or is not) wrong to break promises, or that the US has (or does not have) a duty to reduce their greenhouse gas emissions. Below moral facts are explained in terms of several metaphors. Which of these metaphors seems most appropriate to you?
>
> *[SECULAR REALISM]*
> Moral facts are "discovered". They can be discovered in the same way in which we discover other facts about the objective world.
>
> *[THEIST REALISM]*
> Moral facts are "divine commandments". They are introduced and determined by God.

*[CULTURAL RELATIVISM]*
Moral facts are "cultural inventions". They are introduced and determined by cultures.

*[INDIVIDUAL SUBJECTIVISM]*
Moral facts are "individual inventions". They are introduced and determined by individuals.

*[ERROR THEORY or NON-COGNTIVISM]*
Moral facts are "illusions". While it may seem to us that they exist they actually do not exist at all.

Discovery is an epistemic term, which could have prompted some confusions or conflations with the first response option. There is also the issue that ordinary people don't appear to understand the term "objective" in a consistent way, or in the same way as the term is used in academic philosophy. A more serious issue, however, is simply that it's vague. What does it mean to discover facts in the same way we discover facts about the "objective world"? It's left to each participant to fill in the blank as to how it is we discover facts about the "objective world," and it's possible many of these interpretations would prompt people to avoid this response option because they imagine that discovering facts about the objective world means something like discovering them *empirically*, or *scientifically*. In fact, this strikes me as the *most* plausible interpretation. Yet moral realists don't think we discover moral facts empirically or using science. Given this concern, it's possible many participants would interpret this metaphor in unintended ways. Also, if people think some moral facts are obvious or self-evident, they may think we don't need to discover them.

Proponents of theistic realism do not necessarily think God determines or introduces moral facts. The notion that God "determines" facts could be interpreted in antirealist terms, and the idea that God "introduces" such facts suggests that they may not be eternal. It's also not clear why the term "divine commandments" is in scare quotes.

I have little to say about the cultural relativism and individual subjectivism conditions, but it does strike me as odd to say they're "introduced" by cultures or individuals. Note, as well, that

"determined" may have a clear meaning to philosophers, but ordinary people could understand this as an epistemic concept that means something like "discovered." Regardless, it's not clear what "determined" or "introduced" mean in these response options.

The last item is double-barreled and isn't appropriate as a measure of noncognitivism. It's not clear noncognitivists would agree that it seems to us that there are moral facts. Also error theory doesn't necessarily require denying that people have cognitivist phenomenology, so this is a strange way to frame error theory, too. Also, error theory typically denies stance-independent moral facts, not the notion that there are any moral facts at all. This lack of specificity may have discouraged some participants from selecting it. Finally, it's strange that "illusion" appears in scare quotes. Why is it in scare quotes? I understand the goal of consistency across response options but it strikes me as a bit strange.

Overall, I suspect many participants may have favored the relativist response options to this question because the realist and error theory/noncognitive response options are framed in a strange way that may have led them to be unappealing. There is also the possibility that participants are thinking in descriptive terms, and favoring relativist options because they think these items accurately capture the genealogy of moral beliefs: that our moral beliefs tend to come from our personal values and our cultural standards.

### S3.10.2.3 Abstract comparison task

Next, we have a comparison task. Participants were given the following instructions:

> Below morality is being compared to various types of matters. Please indicate which comparison seems most appropriate to you.

They were asked to choose one of the following response options:

> *[SECULAR REALISM]*
> Morality is akin to science or mathematics. There are objective facts about what is right or wrong (facts that are independent from what anybody thinks about them). We cannot change these facts, we have to discover them.

*[THEIST REALISM]*

Morality is akin to religion. What is morally right or wrong is determined by what God wants us to do. Individuals cannot, by themselves, change the moral facts.

*[CULTURAL RELATIVISM]*

Morality is akin to social conventions. In each culture different things can be morally right or wrong. Cultures determine the moral facts. Individuals within cultures cannot, by themselves, change those facts.

*[INDIVIDUAL SUBJECTIVISM]*

Morality is akin to personal taste or preferences. For each person different things can be morally right or wrong. The individual determines the moral facts.

*[ERROR THEORY]*

Morality is akin to superstition. It is based on a fundamental error. It assumes things exist (namely facts about rightness and wrongness) that do not actually exist.

*[NON-COGNITIVISM]*

Morality is akin to exclamations (such as "Yeah!" or "That sucks!"). We use terms such as "right" and "wrong" to express our feelings, emotions, intentions or attitudes, but that is all. There are no moral facts.

"Akin" is not a common word, so it's a bit strange to use it in a questionnaire when one's concern is with participants not understanding the questions. Why not use "like"? In any case, it's strange that the secular realism condition compares morality to science or math. Like them *how*? Many philosophers that endorse moral realism endorse moral non-naturalism, which is virtually, by definition, the view that moral claims are in certain fundamental respects *unlike* scientific facts. Given the prominence of non-naturalist realism, it's strange to presume ordinary people would be disposed towards thinking moral facts might be like scientific facts. Non-naturalist realism is not a theistic or religious position; it's prominent among *secular* moral philosophers. It's just as troubling to say morality is akin to mathematics. Ordinary people aren't necessarily mathematical platonists; they may have no philosophical stance on the matter, or not know how we discover mathematical facts, or think math is disanalogous to morality in unknown ways, or they could be implicitly committed to nominalism or

some other view of math where they take some kind of antirealist or anti-platonist stance towards it, in which case judging that morality was like math would not lend itself to realism. Why does this item presume that if morality is like math that therefore it's stance-independently true? Also, by saying "science or mathematics," this introduces another concern. Are these presented as similar to one another? Or different? If different, now participants are presented with a strange disjunct. What if I think morality is like math, but not science?

The theistic realism item is especially strange. It begins with the claim that morality is "akin to religion." This is *extremely* open-ended. Akin to it *how*? Religions could be seen as culturally constructed systems, in which case choosing this option could imply a form of antirealism. Someone could even choose this option because they think it's a descriptively accurate account of how others think about morality, even if they don't themselves endorse it. Also, strictly speaking, if moral facts are determined by what God *wants* us to do, this is a type of stance-dependence, which would indicate a relation-designating nonrelativist form of antirealism, not realism.

Next, we have "Morality is akin to social conventions." This seems highly susceptible to descriptive conflations. In a certain descriptive sense, our moral standards are much like social conventions, even if some of those conventions are stance-independently true. Even if there are stance-independent moral facts, there are still the actual moral standards people abide by, which someone could readily regard as social conventions (because they quite literally are). This calls attention to an ambiguity: even in a world populated by stance-independent moral facts, there are at least two distinct moral phenomena in the world: the moral facts themselves, and the moral systems different human populations abide by. Such systems are cultural constructions that may or may not reflect the actual moral facts (if they exist), but even if they do, there is a difference between the moral fact that "it's wrong to hurt others for fun," and the internalized social norms, reflected in the attitudes, beliefs, and behaviors of individuals in a particular community who *believe* it's wrong to hurt others for

fun, and act accordingly. The former may be some kind of *a priori* moral fact, while the latter is a fact about the psychology and sociology of a population. The latter exists even in a world with the former. But it may not be obvious to ordinary people which of these researchers are interested in. Lastly, ordinary people may reject elements of this item even if they generally subscribe to cultural relativism. The last part states that "Individuals within cultures cannot, by themselves, change those facts." Yet cultural relativism is open-ended with respect to how the moral rules of a culture are determined. It is not necessarily a simple majority; the standards of elders, or legal procedures, or influential minorities may determine the facts. For instance, many ancient civilizations had dominant populations who subjugated minority populations or held large numbers of slaves. Yet the cultural and moral institutions in force in such societies were those of dominant social class, *even if that class were not a majority of the population*. Individuals or social groups within a society may very well have been endowed with the functional capacity to change the norms or institutions of that society, whether in spite of the desires of the population (in the case of tyrants and monarchs), or in accordance with it. In the latter case, people may also believe that individuals have the power to persuade or change the minds of others. People might think that, for instance, individual civil rights activists could cause a shift in their society's moral standards. The problem here is that the notion that individuals "cannot, by themselves, change the moral facts," is intended to mean something very specific and technical: that the truth of moral claims doesn't depend on the beliefs of any particular individual, such that if that individual's moral beliefs change, the moral facts change. Yet a person could "change the moral facts," in other ways, such as persuading others or acquiring political power.

Next, we have the claim that morality is "akin to personal taste or preferences." Some participants may have eschewed this response option because the notion that morality is merely a matter of "taste," could imply not simply that it's made true in a stance-dependent way, but that it is *merely* a matter of taste, i.e., it's in some sense trivial or unimportant. In my experience, this is an

incredibly common canard hurled at the moral antirealist: that their moral standards are somehow arbitrary, or lack meaning, or force, or importance. This simply isn't true. Regarding your moral standards as stance-independent does *not* entail that they're as trivial and arbitrary as your favorite pizza toppings. An antirealist could be indifferent or at least not overly concerned if their taste in food or music changes, but many would rather die than have their moral standards change significantly. This item also includes a second, unfortunate phrase: "For each person different things can be morally right or wrong." This is ambiguous. There is a literal interpretation that things can be *literally correct* relative to an individual's moral standards, which would be the intended metaethical meaning, or it could be the descriptive claim that different people *regard* different things as right or wrong. Ordinary people frequently reference the latter when responding to questions about metaethics, and will say things like "it's true *for her.*" This *could* be a form of relativism, but often people's comments suggest or clearly indicate that people mean that "it's true *according to her*," which is not subjectivism, it's just a descriptive claim about people's moral standards. In addition, the notion that "For each person different things can be morally right or wrong," implies *agent* relativism, not *appraiser* relativism. As such, this item may misleadingly imply a very specific form of relativism.

Next, we have the claim that morality is "akin to superstition." again, akin *how*? I don't believe in stance-independent moral facts, but I wouldn't think of morality as like "superstition." I think of superstitions specifically as mistaken beliefs about supernatural or paranormal phenomena, such as luck. I don't think morality is like that. Ordinary people may likewise be disinclined to select this option due to the various associations that come to mind when thinking about "superstition." They may think of superstition as something ignorant and foolish people believe in, whereas morality is ubiquitous and important to our everyday lives. In other words, superstition may have negative connotations that morality doesn't, and it would be going too far to say morality is like superstition since this would imply that people who held moral beliefs were ignorant or foolish. I deny stance-

independent moral facts, and in that respect am very similar to error theorists. Yet I don't think moral beliefs are "superstitions," with or without negative connotations associated with superstition. It's not reasonable to expect ordinary people to set such connotations aside when considering response options. It's also not clear what it means to say morality is based on a "fundamental error." *What* error? What's meant by "fundamental"? Is superstition based on a fundamental error? Finally, note that this item does not make clear that it involves only the denial of stance-independent facts, even though this would be necessary to clearly express error theory.

Finally, we have the claim that morality is "akin to exclamations (such as "Yea!" or "That sucks!"). This only represents a crude form of emotivism that arose last century and fell out of prominence. It does not reflect prescriptivist accounts of noncognitivism, nor does it accurately represent contemporary forms of expressivism that are more sophisticated. This isn't to say any of these theories are plausible, but that the comparison task represents noncognitivism with a caricature of noncognitivism, not the real thing. Furthermore, while philosophers may understand the comparison to exclamations to capture the notion that moral claims express only nonpropositional content, it's not clear that ordinary people would recognize this. Moral claims aren't always made in contexts where we would make exclamations in particular, *even if* their primary or central role were to express nonproportional content. In other words, even an emotivist whose views most closely resembled this response option would not think that moral claims uniformly or exclusively function as *exclamations*. Think of it this way: when you express emotions, do you *only* express them as exclamations like "Nooooooo!" and "Wow!"? No. People express emotions in contexts that are more somber or have a different emotional valence, and the same could certainly be true of moral claims. Depicting noncognitivism as the "Boo-Hurrah" theory is largely tongue-in-cheek. Unfortunately, this response option takes that unserious way of characterizing noncognitivism far too literally, resulting in a highly unattractive item that does a very poor job of representing noncognitivism. This is mitigated

somewhat by the follow-up clarification: "We use terms such as 'right' and 'wrong' to express our feelings, emotions, intentions or attitudes, but that is all" (p. 62). Yet this may be too little, too late. Imagine you are taking a survey. You are given a massive list of response options. Some include multiple sentences. What are you most likely to focus on? I suspect the answer would be the first sentence. And in any case, that's likely the first part of a response option you'd read. If that part seems unappealing or inconsistent with your views then, in order to favor that response, you'd now have to overcome this initial disinclination. That is, a person may think "I don't think morality is like exclamations…well, okay, I see, maybe this is just an example, and it really means they just express emotions. Maybe that's true. Still, this seems to focus on morality as exclamations, and I don't think it's like *that*. Let me find a better option." Participants either have to agree with *all* the content of the response option, in which case the mitigating clarification may be inadequate, or they'll ignore elements of the response option, in which case their response isn't valid anyway since it's not actually reflecting agreement with the item. That calls attention to yet another problem: all of these items are multi-barreled. Participants must agree with *all* or *none* of the content of any given response option, even when it includes multiple elements. Finally, note something strange about this item: it states that *morality* is akin to exclamations. Yet this isn't what noncognitivists think. Noncognitivists think that moral *claims* express nonpropositional content (such as emotions and exclamations). It does not follow that noncognitivists think morality is like exclamations. Morality can be described at multiple levels and in different ways simultaneously, even on the same overarching account. The noncognitivist may think that the function of moral claims is to express emotions, but morality at a societal level may reflect a complex web of language and emotional expression that facilitates cooperation within a society. In other words, noncognitivists may still make descriptive claims about what morality is akin to that are not fully reducible to the fact that they regard the function of moral claims to be expressive. In other words, morality is about more than just moral claims. Participants may likewise recognize the

descriptive inadequacy of comparing morality to exclamations. Even if moral claims did only express exclamations, the collective interaction of many people making distinct kinds of exclamations may not be reducible to this fact alone, since such exclamations could play important sociofunctional roles. If so, even someone who'd regard moral *claims* as exclamations may not agree that morality itself is akin to or reducible to exclamations. As such, this item does not accurately reflect noncognitivism, since noncognitivism's semantic thesis is exclusively concerned with explaining moral *claims*, not morality as a whole.

### *S3.10.2.4 Abstract disagreement task*

Next, we have the abstract disagreement task. This is a new version of the disagreement paradigm, updated to minimize some of the problems with earlier versions. Rather than present a specific moral issue, participants must adjudicate a dispute about a moral issue that's left unspecified:

> Consider the following scenario. Two people from the same culture are evaluating the exact same situation and utter conflicting moral sentences about it. One person says that what happened is morally bad (wrong, vicious, etc.). The other person says that what happened is not morally bad (wrong, vicious, etc.). Which interpretation of this disagreement seems most appropriate to you? (p. 62)

Note the improvements. They specify that both people are members of the same culture, minimizing evaluative standard ambiguity. This item also minimizes the risk of attributing the source of moral disagreement to something other than a difference in moral value, by specifying that they're referring to the "exact same situation." Here are the response options:

> *[SECULAR REALISM, CULTURAL RELATIVISM or THEIST REALISM]*
> One of these two people is right and the other one is wrong (Please note that this could be the case for several reasons: for example, because the truth of moral sentences is objective, or because it is determined by the dominant moral beliefs in their culture, or because it is determined by the commandments of God).

> *[INDIVIDUAL SUBJECTIVISM]*
> Both people are right (because the truth of moral sentences is determined by the moral beliefs of individuals).

*[ERROR THEORY]*
Both people are wrong (because although moral sentences intend to state moral truths, there are no such truths).

*[NON-COGNITIVISM]*
Neither person is right or wrong (because moral sentences do not intend to state moral truths, and are therefore neither true nor false).

There are many strange things about this study, especially the follow-up questions. It's less than ideal to begin with a response option that is consistent with multiple, conflicting metaethical positions and expect people not to balk at choosing it despite its transparently underspecified nature. Note that while cultural relativists might believe that if two people from the same culture disagree, at least one would be incorrect, this requires a sophisticated recognition that both people are indexing the same moral standard. Even if an ordinary person with a rudimentary commitment to cultural relativism would recognize this on reflection, it's not obvious that this would be salient when responding in the middle of a study; they may take "right" or "wrong," to mean right or wrong *simpliciter*, and not have the indexing element of the truth claims a salient factor. In which case, they may perceive this response option to reflect something more like realism than relativism. In other words, performance errors could bar the cultural relativist from favoring this response option. Social desirability may also discourage them, since it may feel unappealing to select a response option that could signal intolerance or rigidity; this is especially plausible in this case since the item was deliberately designed to lump the cultural relativist in with realists, and even includes specific remarks that indicate that one of the people could be mistaken because of God's commands. So from the very outset, cultural relativists who want to respond accurately would be *required* to choose a response option that lumps them in with theistic realists, even though this may be the view they are most diametrically opposed to.

The individual subjectivist item seems fine for the most part, though there's still a worry that ordinary people won't interpret the notion that moral truth is "determined by the moral beliefs of

individuals" in the way philosophers intend. The error theory item is also fine, though error theory remains a complicated synthesis of semantic and metaphysical claims that it's unlikely ordinary people would understand; finally, the noncognitivism item is also reasonably adequate. Overall, these response options are quite minimal, but do a good job of reflecting the metaethical positions. As critical as I am, I should give credit where it is due: these are well-crafted response options. I doubt ordinary people interpret them as intended anyway, but that's no fault of researchers.

The trouble mostly concerns the first response option. Since the first response option doesn't distinguish between realism and relativism, participants who select this response option are directed to a set of follow-up questions. This involves presenting them with a disagreement between people from different cultures. Participants who judge that both are correct in this situation are interpreted as cultural relativists (p. 63). This seems fine, though the problem that normative considerations would influence how people respond persists. By not presenting participants with concrete moral issues, even this is kept at a minimum. So the problems with the disagreement paradigm are largely minimized at this point, even if nonspecific matters of interpretation remain.

The real worry emerges with the next question. If participants persist in judging that only one side of the dispute is correct, they are given a third scenario:

> Those who again choose "One of the persons is right and the other one is wrong" are presented a third scenario in which the disagreeing parties are subject to different commands by God, with each of their moral judgements corresponding to these commands. (p. 63)

Those who still judge that one person is incorrect are classified as realists, while those who judge that both are correct are classified as theistic realists. This is bizarre. First, suppose you're a theist. While you might believe God issues different commands to different people (e.g., Noah was ordered to build an ark, but you and I weren't), this scenario requires that people *disagree because of those commands*. This requires either that these people disagree, even though their commands don't conflict with one another, in which case these people are confused and are not engaged in a genuine disagreement about

conflicting moral standards, in which case responses to this scenario wouldn't reflect one's metaethical views, or they really are subject to conflicting moral commands, in which case we'd have to imagine that God deliberately issued conflicting commands to two different people. For many theists, this may be literally impossible: God simply *would not* issue genuinely contradictory commands to people, especially on the view that one is morally obliged to comply with God's commands. So from a theist's point of view, they may be asked to respond to a question that is inconsistent with their background beliefs. As such, any response they give would be a forced choice between answers that don't reflect what they think. At best, such participants could entertain a counterfactual where they consider what they *would* think if *hypothetically* God were to do something like this, yet it's unclear whether ordinary people responding to this question would actually engage with this counterfactual. More importantly, it's not clear whether their response to a counterfactual kind of this would serve as a valid measure of their actual metaethical position.

Note that this isn't even the most troubling part of this scenario. Think about what participants are being asked: God issues command A to one person, and command B to another person. These people disagree about what should be done morally, and the participant in the study is asked whether one of these people is mistaken *even though God commanded them*. This requires not only rejecting DCT, it also requires the judgment that God can issue mistaken commands! To many theists (perhaps most, or nearly all), this is *literally impossible*. God can't be mistaken about what's morally right or wrong! And believing God cannot be mistaken about moral claims does not require endorsing DCT. Someone who believes moral facts are true independent of God's commands can (and probably would) still believe God cannot be mistaken. The reason would simply be that (a) God is morally perfect, and would not lie or issue immoral commands to people and (b) God is omniscient and incapable of error, and would thus have perfect epistemic access to the moral facts. Given (a) and (b), if God issued a command, it is morally obligatory to follow the command. If God issued two people commands and

they disagreed about what was morally correct to do, this disagreement *could not* be due to a genuine contradiction between those commands, since this would be logically impossible. In short, regardless of the theist's position on DCT, the response options they're given would either not reflect their metaethical views because the disagreement is due to errors on the part of one or more of the disputants rather than a genuine moral conflict, or the situation is logically impossible, in which case they are forced to choose from among a set of response options that they don't endorse. In short, participants may be forced to choose from among response options that require them to entertain impossible scenarios. This is not an appropriate way to measure people's metaethical views.

The scenario is at least as strange for the nontheist or the theist that doesn't endorse DCT. In this situation, they're required to consider what they *would* think if, *hypothetically* there were a God, and God issued conflicting commands. Note how demanding this is. The participant was already asked to entertain a hypothetical disagreement about an unspecified moral disagreement between two anonymous people. Now they're being asked to consider what they *would* think under some *other* hypothetical considerations. In other words, they're being asked to entertain a hypothetical scenario *within* another hypothetical scenario, scenarios that are extremely abstract, since they are almost totally devoid of any of the context that would ordinarily be necessary to provide a meaningful response. The movie *Inception* became the butt of jokes for how cognitively demanding it was for viewers to keep track of dreams within dreams within dreams, yet here we have a scenario in which study participants are expected to keep track of hypotheticals nested within other hypotheticals. This is hypothetiception! We should be wary of responses to conventional hypothetical scenarios. We should be even more wary of responses to situations as complicated as this.

### S3.10.2.5 Abstract truth-aptness task

Next, we have the truth-aptness task. Participants are given an explanation of truth-aptness. This includes being told that truth-apt sentences "express beliefs about facts," and that they remain beliefs

about facts even if we don't know what those facts are. This was coupled with training exercises.

Finally, they were presented with the following instructions and response options:

> Think about moral sentences (sentences that express that something is morally good or bad, right or wrong, virtuous or vicious, and so on). Are these sentences truth-apt or not truth-apt?
>
> *[Cognitivism]*
> Yes, moral sentences are "truth-apt" – that is, they intend to express how things are; what is the case (either with regard to the objective world or with regard to what particular individuals, cultures, etc. think about morality). Thus, these sentences are either true or false
>
> *[Noncognitivism]*
> No, moral sentences are not "truth-apt" – that is, they do not intend to express beliefs about objective or subjective facts, but rather only express feelings, emotions, intentions or attitudes. Thus, these sentences are neither true nor false.

These response options are clear and well-written. Thus, there is little concern that the question and response options weren't framed appropriately. Perhaps they were. The main methodological concerns will thus hinge on *other* problems, e.g., whether people still interpret the instructions or response options as intended, whether the training they undergo is successful, whether training changes participants such that they're no longer ordinary people, and whether the training induces them to engage in spontaneous theorizing. This problem also persists in presuming that ordinary people must think that *all* moral claims are *always* used to convey propositions or to convey emotions. That is, the question presents a forced choice that requires participants to respond in ways that presume uniformity and determinacy, even though the researchers who conducted this study reject uniformity and even though they've presented no good evidence against indeterminacy, either. Once again, it's bizarre to *require* participants to express views that your own research shows that they probably don't think. That is, most studies suggest that ordinary people endorse realism for some moral issues and antirealism for others, cognitivism for some moral claims and noncognitivism for others. So why are they forced to express uniformity?

Finally, note that this study only assesses people's *meta*linguistic intuitions; it does not evaluate actual usage. As such, it at best can only directly measure people's metaethical stances, and not their commitments. Stances are especially vulnerable to spontaneous theorizing, and may fail to reflect what people are actually doing when they make moral claims. That is, simply because ordinary people think that when people make moral claims that they are making truth-apt claims, or aren't doing so, it does not follow that people are in fact doing so. What these results tell us isn't whether ordinary moral claims are truth-apt or not, but whether ordinary people *think* that they are.

### S3.10.2.6 Concrete disagreement task

Next, we return to the original concrete version of the disagreement paradigm. One problem with standard versions of the disagreement paradigm is that participants may not regard a disagreement as a moral issue. If so, then their response won't reflect their metaethical stance, since the issue isn't one they themselves regard as an ethical dispute. To circumvent this, Pölzler and Wright (2020a) introduce instructions explain the moral/nonmoral distinction "in the most general and non-biasing terms" (p. 64):

> The main point of some sentences is to make moral evaluations (i.e., evaluations about a moral matter, evaluating something as being morally right or wrong, good or bad). Here are some sentences that one might think belong to this category: It is wrong to break promises. The US has a duty to reduce its greenhouse gas emissions. Hitler was morally depraved. Parents should be willing to make sacrifices for their children. The main point of other sentences is to make other sorts of (nonmoral) evaluations (i.e., evaluations that don't have anything to do with morality, evaluating something as being correct or incorrect). Here are some sentences that one might think belong to this category: You put your shoe on the wrong foot. That chocolate ice cream tastes good. It is illegal for you to park on campus without a permit. It is rude to talk with your mouth full. (p. 64, footnote 15)

I don't see how they could possibly maintain that these instructions were presented in the "most general and non-biasing terms." Their examples recapitulate exemplars of moral and nonmoral claims *as contemporary analytic philosophers conceive of the distinction*. There's nothing general or non-biasing about this; if ordinary people were already disposed to think of the moral/nonmoral distinction in the same

way as philosophers, we probably wouldn't have to "tell" them about it. We *might* think that ordinary people are implicitly predisposed to draw the distinction in this way, and that these instructions simply make a preexisting disposition salient, but is that what these instructions are intended to do? Unless we employ additional empirical methods to determine whether this is the case, we can't know if we're simply rendering an implicit distinction people already make more salient, or if our instructions aren't *informing* participants of a distinction, but *inducing* participants to draw the distinction in line with philosophers. There's nothing unbiased about that. Keep in mind that participants are given instructions that encourage them to draw the distinction in the philosophers whose ways of thought are already unrepresentative of most of the world since the analytic tradition arose in a highly parochial and idiosyncratic community that is not only mostly composed of people from WEIRD populations, but specifically among highly educated academics in the Anglophone world. If these instructions are ineffective, then we can't be sure the measures are valid. If they are effective, then these instructions may have simply caused participants to think more like philosophers, and to the extent, the study would be causing participants to stop being ordinary people. In short: if our goal is to find out how nonphilosophers think, it makes no sense to prime them to think like philosophers.

All of the examples they use are reminiscent of the tendency for some populations to draw a distinction between moral and nonmoral social conventions, and other normative domains. Yet as I argue in **Appendix D**, and as critics have observed, the moral/conventional distinction may be culturally distinct, and may not reliably emerge across cultures (Machery & Stich, 2022). More generally, there's little indication people from other cultures think about moral norms in the same way as people from broadly WEIRD populations (Berniūnas, 2020; Dranseika, Berniūnas, & Silius, 2018; Machery, 2018; Stich, 2018), and even some religious subcommunities within WEIRD populations may not think about moral norms in the same way (Levine et al., 2021). At present, evidence suggests that people from different cultures and communities do not appear to distinguish moral from

nonmoral norms in the same way as one another. If so, then there would be no way for the instructions provided in this study to be general or non-biased, since they reflect a culturally specific way of thinking about moral versus nonmoral norms. And even if these measures functioned in line with the way some populations think about morality, this wouldn't necessarily translate well to other populations, limiting the generalizability of whatever findings might be obtained with this measure.

Lastly, while some of the changes made to this study may minimize *some* of the methodological shortcomings of previous versions of the disagreement paradigm, they don't eliminate all of them. It remains an open question whether the handful of issues that were minimized or eliminated are enough to provide a valid version of the disagreement paradigm.

### S3.10.2.7 Concrete truth-aptness task

The concrete truth-aptness task is similar in many ways to the abstract version, and is generally subject to most of the same concerns.

### S3.10.2.8 Excessive exclusion rate

32% ($n = 55$) of participants were dropped from analysis due to a variety of exclusion criteria. As noted in previous sections, extremely high exclusion rates can threaten the validity of a measure (Bergenholtz, Busch, & Praëm, 2021; van 't Veer & Giner-Sorolla, 2016). Although there is no clear cutoff, Bergenholtz et al. maintain that: "comprehension failure rates substantially higher than 10% (certainly if they are 25% or higher) should be a cause for concern, since higher numbers increase the risk of participants not being excluded at random" (p. 1536). While I don't endorse these percentages as especially meaningful, the general problem is that as the proportion of participants excluded from a study rises, the greater the risk that the remaining sample no longer represents the population they were drawn from.

*S3.10.2.9 Anomalous findings*

One potential weakness with many of the sets of response options is that they include a disproportionate number of antirealist options (four, as opposed to just two realist options). Recognizing that this could skew participants towards antirealism, Pölzler & Wright (2020a) employed a variety of methods that purportedly reduced insufficient effort, which they claim, "makes it more likely that those who opted for anti-realist options really felt drawn towards these options" (p. 69).[161] Yet they also claim that:

> Moreover, in an independent study we also confirmed that our disagreement tasks deliver plausible results for non-moral domains (scientific statements were dominantly rated as realist, and statements about social conventions and personal preferences were dominantly rated as anti-realist). (p. 69)

Unfortunately, they provide a brief discussion of this study that, if anything, raises more worries than it resolves:

> In this independent study, the statement „The earth is flat"was [sic] rated as realist by 55% of subjects, the statement „Boston (Massachusetts) is farther north than Miami (Florida)"was rated as realist by 63% of subjects, and the statement „The chemical formula of water molecules is H2O"was rated as realist by 60% of subjects. These numbers are of course lower than we would have hoped. That said, based on different measures, previous studies on folk moral realism found high proportions of scientific anti-realists too. In Nichols 2004 studies, for example, 13%, 23%, 22%, 23%, and 18% of subjects responded as anti-realists about facts (even though Nichols' measures likely considerably exaggerated the proportion of realists by invoking first-order intuitions, see Pölzler 2018a, 2018b). This suggests that a considerable proportion of the population may genuinely hold that scientific facts are non-objective. (p. 69, footnote 19)

They suggest that perhaps we should accept these findings at face value. Perhaps, in other words, about half of the people in these studies think there's no stance-independent fact of the matter about whether the earth is flat.

---

[161] This is questionable, given that nearly a third of participants were excluded for various reasons.

I don't believe that. It strikes me as far more plausible that an extremely high proportion of these participants did not interpret the question as intended. This seems far more plausible than the notion that nearly half of the people in the sample are radical antirealists about basic descriptive facts. These findings are so startling that, at the very least, they call for an explanation. These findings almost seem like they'd be better suited as a demonstration of the invalidity of the measures reported in the main study. If a secondary study using similar methods suggests a radical and bizarre outcome like this, it seems strange to casually float the possibility that we simply accept the findings at face value. While possible, I would have thought that these findings would prompt one to revisit the validity of the measures.

### S3.10.2.10 Instructions

The various paradigms above were all conducted between-subjects. Yet before participants proceeded to these paradigms, they were given a general set of instructions intended to clarify what Pölzler and Wright were asking and to familiarize them with metaethics. Once again, these instructions don't present ethics in a philosophically neutral way, but instead present moral philosophy in accordance with the standard distinctions employed in academia:

> Normative sentences about morality express moral judgments. In uttering these sentences we evaluate something morally; we indicate that we regard something as morally right or wrong, good or bad, virtuous or vicious, and so on.
> […]
> Meta-ethical sentences about morality do not express moral judgments. In uttering them we remain evaluatively neutral. Instead, we are making claims about the nature of morality itself. (Pölzler & Wright, 2020b, p. 59)

This isn't *too* terrible a distinction to draw attention to, though it's worth noting that participants are already being primed to draw distinctions even if they hadn't previously done so. Even if this distinction strikes us as benign and reasonable, the mere act of drawing such distinctions is a central component of analytic philosophical training. We simply don't know if ordinary people are disposed

to pull apart metaethical and normative considerations when engaged in everyday moral judgment. It could be that both metaethical and normative concepts are already present in folk thought, but are so intertwined that people don't make the distinction, or there are missing elements from folk thought, in which case these instructions are introducing new concepts and therefore prompting spontaneous theorizing. It's possible the distinction exists in some implicit and nascent form, and researchers are simply rendering it salient, but we're not entitled to presume without evidence that ordinary people are implicitly disposed to draw the same distinctions as contemporary analytic philosophers without evidence. Again, it's *extremely* peculiar that researchers simply assume that the distinctions drawn in a highly sophisticated academic field would be reflected in the unconscious psychological processes and linguistic practices of ordinary people.

In addition, note that participants are introduced to the way contemporary moral philosophers discuss normative ethics by referencing things as "right or wrong," "good or bad," or "virtuous or vicious." These pairings echo normative, evaluative, and virtue theoretic positive and negative pairs, respectively. Once again, the instructions may seem completely benign, but participants are being taught to think specifically in the way that contemporary moral philosophers think. And by contemporary, I do mean *contemporary*. Take the reference to "virtuous and vicious." This alludes to virtue ethics, which emphasizes the cultivation of virtue and focuses more on positive and negative character traits than on the analysis of right and wrong action (Hursthouse, 2002). Virtue ethics dominated Western moral philosophy, only to fall out of favor in the West for a few centuries, then enjoy a renaissance as recently in the 1950s (Hursthouse, 2016). Yet despite its scholastic dominance for most of the history of Western philosophy, virtue ethics still plays second fiddle to deontology and consequentialism in contemporary normative ethics, and is often mentioned as an afterthought when it is mentioned at all.

The same is true of these instructions, which, while they reference virtues, still focus on evaluating "something" morally (not *someone*). In addition, three of the four examples that they provided focus on actions:

> *It is wrong to break promises*
>
> *The US has a duty to reduce its greenhouse gas emissions*
>
> *Hitler was morally depraved*
>
> *Parents should be willing to make sacrifices for their children* (p. 59, footnote 9)

This emphasis on actions is reflected in response options and even the design of the studies themselves. Note, for instance, that the disagreement paradigm focuses almost exclusively on the analysis of moral *actions* and not on the assessment of *virtues*. Participants are asked to judge *events* rather than *people*, and the concrete moral disagreements focus primarily (though not exclusively) on disagreement about actions. One commendable feature of Pölzler and Wright's items is that they *do* incorporate virtue theoretic evaluations:

> *Men who violently physically punish their children are cruel.*
>
> *Martin Luther King was a righteous man.*

This is an excellent improvement over previous studies, and displays genuine insight into the deficiencies of previous studies. But it's not enough. Tossing in a handful of virtue theoretic items doesn't do justice to the vast gulf between thinking in largely characterological terms and thinking in terms of actions and principles. We don't know what role or proportion each plays in characterizing folk moral thought, and the same paradigm may not be ideal for measuring both simultaneously. Pizarro and Tannenbaum (2012) argue that character evaluation is central to ordinary moral thought[162], and that moral psychological research is often deficient in its failure to adequately address the

---

[162] At least with respect to moral blame.

importance of character evaluation. I see little reason not to think such concerns could extend to folk metaethics research. Perhaps it's a mistake to focus so much on actions, and so little on character. The problem isn't that we know this is the case, but that we *don't* know. The very content and structure of folk metaethics research reveals, in its very design, the distinctive cultural background of the researchers conducting the studies, a background steeped in a culturally distinctive emphasis on action over character, even though this may not represent ordinary moral thought. In other words, the very paradigms themselves are thoroughly saturated in a relatively recent emphasis on moral actions and principles over character traits. Even these improved forms of the disagreement paradigm are exemplars of precisely the ways in which parochial conceptions of a topic can bleed into the design of a study without anyone noticing, or at least adequately appreciating the degree to which the studies themselves may represent a parochial way of thinking about morality.

There is another problem with introducing virtue theoretic terms in these paradigms. Terms such as *cruel*, *depraved*, and *righteous* are *thick* moral concepts, and as such they may incorporate a much greater degree of descriptive content than thin moral concepts (Väyrynen, 2021). This risks amplifying the risk of descriptive conflations (see **Chapter 2**). When we think of someone being *cruel*, for instance, this is accompanied by certain substantive descriptive traits or behaviors, which may be true of a person regardless of our moral evaluations of their actions, i.e., there is a certain degree of non-normative descriptive content to such evaluations. These items may enhance the generalizability of findings by introducing a broader array of moral concerns, but compromise validity in doing so, by enhancing the risk of unintended interpretations.

The problems don't end there. Participants not only have to understand the distinction between metaethics and normative ethics, they also have to recognize that the questions in the study are about metaethics, and they have to not allow the intrusion of their normative standards to influence the way they respond. For instance, people with a strong normative moral opposition to an action

may judge that if two people disagree, the one who holds a moral stance inconsistent with their own is "wrong," not because realism is true, but because the participant wants to express their *normative* objection to that person's position. Even if people are aware of the distinction, it may be difficult for people to fully suppress the influence of normative moral standards. To minimize this risk, Pölzler and Wright (2020b) provide the following instructions:

> Given that we are interested in your intuitions about meta-ethical sentences, we ask you to "bracket" your views about the normative sentences that we will present you (to ignore these intuitions or put them to the side). For the purposes of this study it does not matter whether, for example, you judge that breaking promises is wrong, that the US has a duty to reduce their greenhouse gas emissions, and so on. (p. 59)

This is not adequate. It's not plausible that participants can fully suppress their normative attitudes when responding to questions simply because researchers ask them to. Concrete moral issues presented in the study include abortion, adultery, and child abuse. Can ordinary people simply *turn off* their normative attitudes about these issues because they're asked to do so?

### S3.10.2.11 Comprehension checks

Participants were given a theoretical question as a comprehension check. They don't provide details of what the question is, but it appears to be a question that checks whether they understood the distinction between normative ethics and metaethics, since the correct answer is:

> *Normative sentences express moral judgments and meta-ethical sentences make claims about the nature of morality itself* (Pölzler and Wright, 2020b, p. 59)

97.4% ($n$ = 114) participants got the question correct on the first attempt, and the remaining 2.6% ($n$ = 3) got it right on the second attempt. While this is good evidence that people were capable of responding to the question adequately, this appears to be a multiple choice question. When presented with a distinction, followed shortly thereafter by a multiple choice question, such questions serve more as an attention or memory check than a *comprehension* check. For example, if I tell you:

> *All zorps are blorps, but not florps.*

Then later ask you:

> *Which of the following is true of zorps?*
> ○ *They're all blorps*
> ○ *They're all florps*

I suspect most people could get the correct answer. Should we interpret this as an indication that people *comprehend* the distinction between blorps and florps? No, because this is literally impossible. These terms are all meaningless nonsense I made up for the purposes of illustration. The comprehension check Pölzler and Wright employed presents such a low bar that it illustrates almost nothing, other than the minimal capacity for participants to repeat information they were just given. And they don't even have to recall the wording, but were simply given the options to choose from a list. This is an extremely superficial test of comprehension. In fact, it's so superficial I'm hesitant to call it a comprehension check at all. This is indicated by *literally all participants passing it*. While this is evidence of some minimal comprehension, there's a tradeoff. Ideally, all participants would illustrate comprehension. However, in practice, when everyone does so, this may serve as evidence that people have comprehended to the level indicated by the test, but this may indicate that everyone crosses an extremely low threshold of comprehension. A good comprehension check is not merely one that most people succeed at, but one that is sufficiently robust that success means something. The rest of the comprehension checks are not presented in the article, so it's not possible to assess their quality. They employed several other comprehension checks that were not described in the study, so I cannot assess them. However, they do describe their instructions and comprehension checks for training participants on the truth-apt/not truth-apt distinction. This consisted of four true/false questions:

A false sentence cannot be truth-apt. [False]

Truth-apt sentences can express beliefs about facts that are subjective, that is, facts that are determined by the moral beliefs of individuals, the dominant moral beliefs in cultures, and so on. [True]

Truth-apt sentences only express feelings, emotions, intentions or attitudes. [False]

Even if we do not know whether a sentence is true or false it can still be truth-apt. [True] (Pölzler and Wright, 2020a, p. 12)

Again, these are extremely simple comprehension checks that seem to require little more than that the participant attend to and recall the instructions they were given. Getting these questions correct immediately after reading instructions is easy and does not require much in the way of significant comprehension. This is, if anything, more a test of reading comprehension or memory than genuinely internalizing and understanding philosophical concepts. This is very far from what Kauppinen (2007) has in mind when insisting that genuine reflection and competence would be necessary for studies among ordinary people to genuinely reflect the relevant kinds of philosophical judgments. It's absurd to think people understand complicated philosophical topics because they can respond mostly accurately to a handful of true/false questions.

### S3.10.2.12 Classification task

Participants were also given a classification exercise, which required them to judge which statements were truth-apt or not:

> **Truth-apt**
> My pencil is sharp.
> She was very sad about what happened.
> Walking in the street is generally safer than running.
> Garlic lowers cholesterol.
> John believes that it was fun storming the castle.
>
> **Not truth-apt**
> Don't run in the street.
> Yikes!
> Bummer!
> Have fun storming the castle.
> Be happy about what happened!
> (Adapted from Pölzler & Wright, 2020a, p. 12)

One point of concern is that the instructions and response options largely describe noncognitivism by referencing emotions and attitudes, yet the "Not truth-apt" category includes imperatives. This is an inconsistency that may have confused some participants. Yet this is a minor concern. A far more serious concern is that these instructions are extraordinarily biased against noncognitivism. All of the truth-apt sentences are expressed in the indicative mood, while none of the non-truth-apt sentences are. This creates the impression that we can determine whether an expression is truth-apt exclusively by examining its grammatical structure: if it's in the indicative mood, it's truth-apt. Yet no noncognitivist denies that moral claims are presented in the indicative mood. None deny, that is, that people say things like "murder is wrong." Their whole point is that in spite of this apparently propositional structure, they nevertheless fail to express propositions. By training participants to associate truth-aptness with the grammatical structure of a sentence, the instructions train participants to adopt a view of linguistic analysis that effectively begs the question against noncognitivism. And it does so in a way no noncognitivist would accept. The instructions effectively conflate grammar with semantics. This is bad enough, yet this isn't even accounting for an equally serious problem, which is that all of these instructions completely ignore the role of pragmatics. Indeed, at the beginning of their article they include a footnote which states:

> As we define non-cognitivism as belonging to moral semantics and philosophical psychology, our considerations are not meant to apply to pragmatic versions of non-cognitivism (according to which moral speech acts are not assertions). (pp. 2-3, footnote 1)

They refer to an older version of the *Stanford Encyclopedia of Philosophy* entry on moral-antirealism, which discusses pragmatic versions of noncognitivism by Joyce (2007). It's interesting to have a look at what this article says, though I can't be sure whether this is the precise passage they were referencing:

> It is impossible to characterize noncognitivism in a way that will please everyone. Etymologically speaking, moral noncognitivism is the view that there is no such thing as moral knowledge. But it is rarely considered in these terms. Traditionally, it is presented as the view that moral judgments are neither true nor false. This characterization is indeterminate and

problematic in several ways. First, it leaves it unclear what category of thing a "moral judgment" is; in particular, is it a mental state or a linguistic entity? If moral judgments are considered to be mental states, then noncognitivism is the view that they are a type of mental state that is neither true nor false, which is equivalent (most assume) to the denial that moral judgments are beliefs. There are at least two ways of treating a moral judgment as a type of "linguistic entity": We could think of it as a type of sentence (generally, one that involves a moral predicate, such as "…is morally good" or "…is evil") or we could think of it as a type of speech act. On the former disambiguation, noncognitivism is the semantic view that moral judgments are a type of sentence that is neither true nor false, which is equivalent (most assume) to saying that the underlying grammar of the sentence—its logical form—is such that it fails to express a proposition (in the same way as, say, "Is the cat brown?" and "Shut the door!" are sentences that fail to express propositions). On the latter disambiguation, noncognitivism is the pragmatic view that moral judgments are a type of speech act that is neither true nor false, which is equivalent (most assume) to the denial that moral judgments are assertions (i.e., the denial that moral judgments express belief states). (For discussion of the semantic/pragmatic distinction, see the entry on pragmatics, section 4.) In all cases, note, noncognitivism is principally a view of what moral judgments are not—thus leaving open space for many different forms of noncognitivism claiming what moral judgments are.

Note that the semantic view involves the use of moral predicates. Even if Pölzler and Wright want to focus exclusively on semantic versions of noncognitivism, their instructions still present the misleading impression that if a sentence includes a moral predicate, it is *ipso facto* truth-apt given that fact alone, because it is the *grammar* of the sentence that determines whether it's truth-apt. Yet Note Joyce's characterization of semantic views of noncognitivism:

> On the former disambiguation, noncognitivism is the semantic view that moral judgments are a type of sentence that is neither true nor false, which is equivalent (most assume) to saying that the underlying grammar of the sentence—its logical form—is such that it fails to express a proposition

If semantic accounts of noncognitivism involve the denial of the propositional status of moral claims in virtue of the underlying grammar of the sentence, why are participants being instructed to judge truth-aptness in such a way that sets them up to adopt a cognitivist stance about the grammar of all sentences with a structure similar to prototypical moral sentences (e.g., "murder is wrong")? This is more or less instructing participants to endorse cognitivism. And all that is completely ignoring that

the entire study only addresses semantic accounts of noncognitivism, while ignoring pragmatic accounts. Again, note how the very structure of the study sidelines unpopular philosophical positions, even if they represent legitimate possibilities for how ordinary people could think. Why are participants *forced* by the structure of the study to adopt a particular stance towards language and meaning that downplays or ignores the role of pragmatics? I suspect many researchers are oblivious to this problem, while others may see little problem with it because they don't think pragmatic accounts are plausible or worth studying. Yet this is not a legitimate stance to take when studying folk metaethics. It reflects little more than the philosophical predilections and preferences of researchers to focus exclusively on mainstream approaches to metaethics that dismiss the role of pragmatics. This is deeply troubling to me, because I not only don't discount such accounts, I consider them the *only plausible accounts*. I think it's absurd to ignore pragmatics, since I think pragmatics plays a central role in understanding moral thought and discourse. Pragmatics isn't some irritating parasite that complicates everything and serves only to muddle things. And it isn't something we should either deal with by removing it, or (as is usually the case) simply ignoring it. I regard it as a substantive and central element of moral thought, language, and practice. It makes no more sense to ignore pragmatics when studying folk metaethics than it does to ignore water when studying marine ecology.

## S3.11 General problems with training paradigms

One serious worry with training paradigms is that, in practice, they will be far too inadequate to instill sufficient comprehension of the relevant concepts and distinctions to be confident people's responses genuinely reflect their metaethical stances. Simply because you attempt to train people in the relevant distinctions doesn't mean you've succeeded at instilling genuine competence in the relevant metaethical concepts and distinctions. P&W's efforts to instill competence in their participants are genuinely impressive. They survey the minefield of potential conflations and unintended interpretations that impeded earlier research, and carefully circumvent these difficulties.

Yet as impressive as their clarifications and distinctions may be, metaethics is a difficult, abstract subject that trained philosophers struggle to fully grasp. I have serious doubts that the instructions and exercises employed by their studies can readily instill an accurate understanding of the relevant concepts in a few *minutes*. This is especially implausible given that study participants may be far less motivated to learn about metaethics than people who intentionally study the topic. All else being equal, it is plausible that less motivated people would be less likely to acquire competence via instruction.[163]

Even under optimal conditions, the disambiguations P&W provide do not cover many possible ways participants could interpret questions about metaethics in unintended ways. For instance, their only disambiguating instructions involve instructing people in the distinction between metaethics and normative ethics. Although this may minimize conflations between metaethics and normative ethics, there are many other ambiguities and conflations that could reduce rates of intended interpretations, in addition to other methodological concerns that go unaddressed or are not adequately addressed. For instance, their instructions do not (i) distinguish metaphysics from epistemology, even though epistemic conflations are very common, (ii) do not explain or train participants in the distinction between realism and absolutism or universalism (iii) do not explain or train participants in the distinction between relativism and contextualism or descriptive claims, (iv) do not adequately address classification inconsistency and do not test for a shared conception of the moral domain[164], (v) do not explain or train participation to avoid evaluative standard ambiguity or

---

[163] Of course, you could sample from populations that want to learn about metaethics. Yet this would be a failure from the outset: such populations may already possess some prior knowledge, and in any case wouldn't be representative of people in general. For comparison, it would make little sense to recruit people who wanted to become professional MMA fighters, train them to fight, and then generalize from the fighting styles they develop to how people fight in general. It is pretty obvious that such a population could be plausibly expected to be much better at fighting in the first place, and to more readily improve in response to training, than the general population, and it is far from clear that whatever styles they develop would be the same as those uninterested people already have or would develop in response to training.

[164] They *do* attempt to address this, but not in a way I find satisfactory. In Pölzler and Wright (2020b), they attempt to explain the distinction between moral and nonmoral norms. Yet this consists of little more than recycling moral terms like "morally right" and "morally wrong," and providing exemplars of putatively moral terms. This is unlikely to be successful,

abstract norm ambiguity, (vi) do not minimize or eliminate the potential influence of reputational concerns associated with selecting particular responses, (vii) still presume a correspondence theory of truth.

Some measures *do* make an effort to minimize particular conflations. For instance, in setting up their version of the disagreement paradigm, participants are given the following instructions:

> Consider the following scenario. Two people from the same culture are evaluating the exact same situation and utter conflicting moral sentences about it. One person says that what happened is morally bad (wrong, vicious, etc.). The other person says that what happened is not morally bad (wrong, vicious, etc.). Which interpretation of this disagreement seems most appropriate to you? (p. 62)

This is a fantastic effort at minimizing the risk of certain conflations. By mentioning that both people are from the same culture, this description minimizes the risk of evaluative standard ambiguity. And by specifying that the two people are evaluating "the exact same situation," and referring to a *specific* event by referring to "what happened," this minimizes the risk that participants would attribute the disagreement to nonmoral differences. I do not wish to understate just how much of an improvement this is over earlier measures: the strides made to improve folk metaethics measures are truly remarkable.

Yet I don't think they're good enough. Adequate interpretation still requires that participants circumvent all the interpretative problems that the instructions and training exercises *don't* address. Unfortunately, training participants to distinguish metaethics from normative ethics may give the appearance of success: if participants pass comprehension checks and succeed at training exercises, this can give the misleading appearance that their responses to items reflect a genuine understanding of the relevant metaethical concepts and distinctions. But the distinction between metaethics and normative ethics is a very basic distinction that a student would learn about in the first week of a

---

especially among populations who have fundamentally different conception of paradigmatic moral norms or whose native languages don't lexicalize equivalent terms in a straightforward and culturally salient way (e.g., Berniūnas, 2020).

course on ethics. Even if their instructions and exercises did succeed at training participants in this distinction, this would barely move us any closer to ensuring intended interpretations, for the same reason that teaching people the distinction between philosophy of mathematics and applied mathematics would enable them to understand questions about mathematical Platonism.

In addition, clarifying the moral standards of two people who disagree is only relevant to agent relativism. This item, and items like it, are incapable of detecting appraiser relativism. This may seem like a minor problem, but it isn't. Quintelier, De Smet, and Fessler (2014) found that the *appraiser's* moral standards influence people's moral standards, observing that "People are more likely to say that a moral statement is true when the message is in line with the agents' moral frameworks compared to when the message is not in line with the agents' moral frameworks" (p. 226). Yet since the appraiser in these studies is the participant themselves, and they are not presented with the appraisal of another third-person evaluator, the study cannot even test for appraiser relativism. Thus, there is already empirical evidence that appraiser relativism is relevant to ordinary people's responses to metaethical questions. This is not surprising, since appraiser relativism is the more common of the two forms of relativism in metaethics. In fact, it is so common that it is treated as the *default* form of relativism in the *Stanford Encyclopedia of Philosophy's* entry on moral relativism, a resource often taken to be a definitive, or at least highly prominent academic resource for philosophical concepts. In the entry, Gowans (2021) states that between agent and appraiser relativism, "Appraiser relativism is the more common position, and it will usually be assumed in the discussion that follows." In short: P&W's disagreement task is only capable of measuring one form of relativism, and it may be the less common of two candidate forms of relativism.

Furthermore, instructions and response options still rely on a host of terms that may have technical and precise meanings to philosophers but may be difficult to interpret as intended by participants, and if so, in a reliable and consistent way as one another. For instance, participants are

told that these people "utter conflicting moral sentences" about a moral issue. Philosophers may understand "conflicting" to mean that they are the logical negation of one another, e.g., Alex asserts "P" and Sam asserts "not-P," but ordinary people may not interpret "conflicting" in this way. Colloquial uses of "conflicting" allow for oblique statements that are not necessarily the precise logical negation of one another. Yet the precise interpretation is *necessary* for the validity of the study. Response options also make reference to the truth of moral sentences being "objective," despite my findings indicating that people don't clearly interpret this to mean stance-independence (see **Chapter 4**), and to the truth of moral sentences being "determined" by e.g., cultural consensus or God's commands. What does it mean for the truth of a sentence to be "determined" in this way? Do ordinary people interpret it as intended (e.g., to concern how God, culture, etc. *ground* moral facts)? I don't know, and they don't mention directly testing this. This is no idle concern, since this the way in which certain groundings "determine" moral facts is a highly technical topic of philosophical inquiry (see e.g., Cohen, 2021).[165]

These concerns illustrate a more general concern that the training and instructions P&W are simply *inadequate*. This seems especially likely given how high a bar they've set for themselves. P&W hope to devise materials that meet Kauppinen's (2007) stringent standards, which "suggest that

---

[165] For instance, Cohen (2021) characterizes the dispute between realists and antirealists as one over "what noncausally makes it the case that some moral facts (or standards) obtain" (pp. 181-182). This is often characterized as "grounding," and, in addition to being noncausal, these grounding relations are generally thought to be *asymmetric*, *irreflexive*, and *transitive* (p. 183). Yet in spite of all the rich literature on grounding, the precise way in which culture, God, or some other account of the grounding of moral facts is obscure and is often underdeveloped (or simply absent) from discussions about realism. Indeed, Cohen argues that there are *two* distinct ways of cashing out the notion of "objectivism" in common use. Namely, there is a distinction between *grounding* and *dependence*, and this furnishes us with at least two ways of construing the dispute between realists and antirealists, one which exclusively emphasizes grounding, and a second that emphasizes both grounding and other dependence relations (p. 189). If so, then there may be ambiguity and underspecificity that makes it unclear what philosophers *themselves* mean when they refer to "determining" the moral facts. I reference all these complications to illustrate that a term like "determined," may seem simple on the surface, but it masks a dialectical *ocean* of philosophical discussion. Experts variously grapple with or ignore it, but ordinary people are expected to be responsive to a reasonable approximation of some adequate notion of "determine" that captures distinctive dependence relations with no formal training, no familiarity with the technical use of these terms, and no knowledge of what's philosophically at stake. Consider, for instance, that one colloquial use of "determine" is simply to "decide" or "settle on." A culture could agree on what the moral facts are without this having any meaningful metaethical implications.

ordinary speakers' intuitions must fulfill three additional conditions in order to be relevant to moral semantics and philosophical moral psychology" (p. 6):

(1) *Reflection*: Judgments about the application conditions of the relevant concepts must be made under "sufficiently ideal conditions" (p. Kauppinen, 2007, p. 101)

(2) *Competency*: Participants must be competent with the relevant terms and concepts and must therefore not be subject to performance errors

(3) *Semantics*: Participants must be responsive exclusively to semantic considerations, and must therefore not be influenced by pragmatic considerations[166]

It is difficult to overstate just how stringent these criteria actually are. Kauppinen (2007) does not appear to believe conventional social scientific methods (like those employed by P&W) are capable of meeting these conditions *in principle*: "intuition statements cannot be interpreted as straightforward predictions, and therefore *cannot, for reasons of principle*, be tested through the methods of non-participatory social science, without taking a stance on the concepts involved and engaging in dialogue" (p. 97, emphasis mine). Of course, P&W are free to reject Kauppinen's claim as overly pessimistic. Indeed, Pölzler (2018a) explicitly rejects Kauppinen's claim that social scientific methods cannot shed light on folk metaethics. Yet even if we reject Kauppinen's claim that such methods could not succeed *in principle*, we may still ask whether they succeed *in practice*.

Unfortunately, I do not believe P&W's studies come close to meeting any of these criteria in practice, even if they could in principle. Kauppinen insists that his conditions are only met "when failures of competence, failures of performance, and influence of irrelevant factors are ruled out" (p. 101). P&W seem to accept this challenge. One immediate problem is that P&W don't appear to adequately characterize Kauppinen's criteria. According to P&W, meeting the criteria for reflection involves accessing people "how speakers are disposed to apply moral concerns having thoroughly

---

[166] These aren't the criteria Kauppinen lists as his "three" criteria. Kauppinen states that "I identify three characteristic assumptions that philosophers implicitly make about the responses that count as revealing folk concepts—competence of the speaker, absence of performance errors, and basis in semantic rather than pragmatic considerations." P&W seem

thought about the case at issue," (Pölzler & Wright, 2020a, p. 6). This is, at best, vague. Even if someone thoroughly thinks about an issue, they could still be systematically disposed to make a mistake. Kauppinen explicitly points this out, stating that:

> A tempting response is to say that correct applications are those that one is disposed to give under suitable conditions. This allows for the possibility of mistakes, since it can be true that I am disposed to do something I do not actually do. It is, however, clear that at least simple forms of dispositionalism do not solve the problem, since, as Kripke points out, we can also be disposed to make mistakes. (p. 102)

According to Kauppinen, proper application of a concept would require meeting certain *normative* conditions. And these normative conditions may only be met by participating in philosophy, not simply thinking thoroughly. As Kauppinen points out, "the question about who is a competent user is a normative question, a question about who gets it right, and it is very hard to see how one could answer it from the detached stance of an observer" (p. 105). Setting this aside, I wish to simply examine whether any plausible conception of the conditions for reflection, competency, and semantics were met using P&W's methods.

### S3.11.1 Reflection

First, do P&W succeed at prompting responses from people who have "thoroughly thought about the case at issue"? This strikes me as very unlikely, under any reasonable construal of "thoroughly." The instructions P&W provide are accurate, clear, and well-written, but they are sparse. They are the sort of descriptions you'd get from the first page of an introductory textbook or the slides in the PowerPoint of the first lecture of a course. They are very far from providing the level of detail and explanation necessary for someone to understand the relevant philosophical issues. Understanding metaethical concepts, like many issues in philosophy, typically requires understanding not only some superficial definition of the concept, but understanding its relation to other philosophical concepts, the philosophical implications of the concept, and how the concept holds up under demanding counterfactual conditions, among other requirements. This is why the Socratic method is so powerful

a dialectical tool: it allows the questioner to reveal inconsistencies and contradictions in a person's philosophical positions that, once revealed, prompt reflection and encourage the interlocutor (or victim, as the case may be), to move towards a state of reflective equilibrium by resolving inconsistencies, abandoning defective concepts, and updating their stance in light of the revealed inadequacies of their responses.

Anyone familiar with Plato's dialogues will know that they typically begin with Socrates asking someone about some concept, such as love or justice. This person is tasked with providing a definition, or account of the concept. Then, Socrates subjects this person to a devious series of questions that reveal that the person's account was flawed. Every effort to revise the concept in light of these revelations is subjected to the same scrutiny, and likewise falls short. The conversation ends, typically without any satisfactory resolution. Whatever Plato's aims in writing these dialogues, there is little doubt that they demonstrate how a mere surface definition masks an entire unexplored world of nuance. Just so with the surface level accounts of various metaethical positions or distinctions between different branches of philosophy. Contemporary philosophers have the benefit of over two millennia of hindsight to refine their distinctions and streamline their concepts. Yet clear and competent discussion of these concepts still relies on a highly specialized vocabulary, a distinctive set of methods, and a vast body of knowledge that only people with an education in philosophy possess. Researchers who offer definitions in the midst of a study may find the definition adequate, since it accurately reflects the relevant concepts and distinctions. But the sense of adequacy such definitions may induce in experts is not an adequate guide to judging whether ordinary people understand instructions in the same way. As Moss and I suggest, we suspect that "researchers may systematically face a 'curse of knowledge,' whereby they are unable to appreciate the extent to which untrained individuals may struggle to correctly grasp these positions" (Bush & Moss, 2020, pp. 7-8; see also Camerer,

Loewenstein, & Weber, 1989). In short, accurate but brief descriptions do little to move people towards a state of having "thoroughly" reflected.

None of the rest of P&W's methods plausibly ensure that people have thoroughly reflected. For instance, they excluded 32% of participants from one of their studies for either (a) failing attention checks, (b) completing the study too quickly, (c) failing comprehension checks, or (d) providing "confused or irrelevant verbal explanations of their responses" (Pölzler & Wright, 2020a, p. 11). This is, itself, a significant methodological problem. If our goal is to study how ordinary people think about metaethics, and we exclude nearly a *third* of our participants, we are tossing out a *huge* number of people. Since such people are not excluded randomly, the remaining participants may not represent the population they were sampled from, i.e., "ordinary people." Perhaps one of the empirical facts about ordinary people is that one third of them are so confused about metaethics that they fail comprehension checks and provide confused or irrelevant responses to questions about metaethics. Why aren't these people a legitimate category of ordinary people? Why ignore them? Excluding them by design results in a sample whose members who, by definition, have demonstrated especially high levels of competence with metaethical concepts. Why would what amounts to a semi-elite group of participants be an appropriate sample for drawing inferences about ordinary people as a whole? And how is this an appropriate move to make when one of the hypotheses on the table holds that people simply have a determinate stance on these matters? Such people may be disproportionately likely to fail various checks, and land in the exclusion group, resulting in a sample whose participants are skewed away from a true representation of the target population. Surprisingly, even given their modest criteria, nearly one third of participants still failed in one way or another. This should perhaps give us some pause as well. If this many people struggle with P&W's tasks, it at least hints at the difficulty of metaethics.

P&W employed a variety of other methods intended to prompt reflection. For instance, they told participants that they "'cannot speed through' the study," that "the study involves various comprehension checks," and that they "are only looking for people 'who will be serious and conscientious about reading through answering questions carefully and honestly'" (p. 11).

I have great respect for P&W's work, but these are utterly feeble efforts to prompt adequate reflection. Recall that the reflection in question involves "having thoroughly thought about the case at issue" (p. 6). Think about what would be involved in *thoroughly* thinking about e.g., free will, or your purpose in life, or the existence of God. I imagine long walks deep in thought, discussions with friends and family, and perhaps even watching YouTube videos, reading articles online, or picking up a book or two. For some people, it may involve speaking to a priest or a rabbi, consulting scripture, or meditation. It could even involve taking psychedelics or traveling to see the world. And in all of these cases, one might spend considerable time mulling over one's thoughts and discussions and experiences before reaching a more mature perspective on the matter. *That*, at a *minimum*, is what I would consider having "thoroughly" thought about a philosophical topic. Maybe P&W imagine a far more modest standard, but no reasonable standard of having "thoroughly" thought about a philosophical topic could be achieved by the kind of run-of-the-mill stern warnings they employ, warnings that are routinely employed in surveys and that people's eyes likely glaze over, barely registering to conscious awareness. It's like imagining that those annoying safety videos they play at the start of every flight reliably induce passengers to reflect deeply on the importance of ensuring one's own safety before you aid others. If anything, it's worse, since those videos provide rich, extended, multimedia engagement and are mirrored by flight attendants in real time. And yet we still all recognize this is little more than safety theater. Just the same, it's not plausible that the inclusion of mundane appeals to take a study seriously would contribute in any meaningful way to thoroughly reflecting on abstract philosophical topics. This is methodological theater.

Now imagine a group of students or MTurk participants who are participating in a study for a modest sum of cash ($7.25, in this case), or course credit. Do you think warning them that the study will have comprehension checks, and telling them they can't speed through the study, and that you only want participants who will be serious would be enough to get a significant number of these people to genuinely reflect on the nature of morality, and to do so *thoroughly*? I would say I'm incredulous at the suggestion, but I don't have enough incredulity to do so.

Their other methods of inducing reflection (as well as addressing competency and semantics) are better, but far from sufficient. Participants went through the entire set of steps (1) - (9) outlined above, including instructions explaining the difference between normative ethics and metaethics, telling participants the study was about metaethics, providing an explanation of the available metaethical positions, and putting the participants through a series of training exercises. However, these instructions and exercises are fairly minimal, and while they go some way towards prompting participants to reflect on the relevant philosophical distinctions, it is not at all clear whether one-paragraph descriptions, short training exercises, and a simple quiz, all of which are designed simply to train participants to understand the basic outlines of a distinction, come anywhere close to causing people to reflect on the substantive philosophical theses themselves.

For comparison, even if you succeeded at getting people to understand the difference between libertarian and compatibilist concepts of free will, it would not follow that you thereby succeeded at getting those people to thoroughly reflect on each of these positions in such a way as to provide a genuinely considered judgment. In other words, all of their training focuses on the superficial task of simply understanding putatively conflicting positions, but it goes no further in prompting them to reflect on, e.g., the respective merits or implications of those positions. In short, whatever the merits of P&W's enhanced instructions and exercises, they focus exclusively on training participants to develop a kind of minimal competence to understand what they're being asked. Of course, *my* primary

objection to folk metaethics is that participants do not interpret questions as intended. At the very least, P&W's methods do a far better job of achieving that more modest goal, though as I discuss below at too high a cost.

### S3.11.2 Competence

In any case, I am skeptical that the participants in these studies really do develop genuine competence with metaethical concepts. The training is sparse, and the exercises are incredibly easy. For instance, participants are simply asked to select which of a set of four statements are correct. If they fail, they're given a second chance. It is fairly easy to do well a second time merely due to the process of elimination. Yet even if participants genuinely recognize the correct answer, this is not good evidence that they genuinely reflected on and understood the concepts. Keep in mind that participants are given simple instructions that draw a distinction between e.g., two concepts, then ask people questions that repeat many of the same words that appear in the descriptions of those concepts. This is a task one could potentially pass through shallow processes like memory and word association. Imagine presenting someone with the following instructions and questions:

> *Today, you will learn about several important concepts discussed amongst Jabberwockyians, a group of philosophers dedicated to the study of the Jabberwocky. These concepts are slithy toves, borogroves, and Tumtum trees.*
>
> *If it is brillig, slithy toves gyre and gimble in the wabe. However, borogroves are always mimsy, regardless of whether it is brillig.*
>
> *Although slithy toves normally gyre and gimble in the wabe only if it is brillig, there is an exception to this rule. If one rests by a Tumtum tree, slithy toves will gyre and gimble even if it isn't brillig. However, it will have no effect on borogroves, which remain mimsy regardless of whether one rests by a Tumtum tree.[167]*
>
> *Please select all of the following statements that are correct (you can select more than one)*
>
>     o   *Slithy toves only gyre and gimble in the wabe is if it is brillig*
>
>     o   *Borogroves remain mimsy even after one rests by a Tumtum tree*

---

[167] The nonsense terms used here come from Lewis Carroll's (1871/2022) poem "Jabberwocky."

- *If it is both brillig and one rests by a Tumtum tree, slithy toves will gyre and gimble in the wabe*
- *Sometimes borogroves are not mimsy*

Were you able to select the correct answers and identify the incorrect answers? See this footnote[168] to check how you did. I am willing to bet people would have little trouble with this task, even though the concepts are complete nonsense. One *could not in principle* thoroughly reflect on slithy toves or borogroves. And yet this task is modeled after the task P&W employed to test whether participants understood what a truth-apt sentence is.

This silly example is not intended to illustrate that P&W's instructions and questions didn't succeed at facilitating increased competence with metaethical concepts. Rather, it is to illustrate that one can easily learn patterns and correctly respond to rules one has just been taught even if the substantive content of those ideas remains a complete mystery: indeed, even if it is *complete gibberish*. The kind of task P&W employed is one that could be passed with little or no actual comprehension of the relevant concepts, but simply a capacity to recall information and repeat patterns. This is not the kind of task one would ideally employ if one's goal is to get people to comprehend, much less thoroughly reflect on abstract philosophical questions about the nature of morality.

The training exercise goes some way towards genuinely developing people's ability to employ the relevant distinction, although I question whether this task succeeds, either. A handful of extremely simple examples and a second chance at completing the task are hardly demanding conditions for prompting one to reflect on or develop competence with the relevant concepts. In general, my concern with P&W's exercises is that they are *too simple*. They may confer some competence with metaethical distinctions, but is it enough for us to conclude that participants are competent with the relevant concepts? I have serious doubts. P&W also set the bar *very* high for themselves. For instance, they

---

[168] (1) Incorrect. Slithy toves gyre and gimble if one rests by a Tumtum tree (2) Correct. Borogroves are always mimsy. (3) Correct. So long as either condition is met, slithy toves gyre and gimble in the wabe (4) Incorrect. Borogroves are always mimsy.

state that "competency must be specified in uncontroversial and theoretically neutral ways" (p. 6). At a minimum, this would mean that whatever competence people develop, they must do so in a way that doesn't bias them towards one or another of competing metaethical positions. Yet they must also be free of performance errors, such as conflating questions about metaethics with normative or other non-metaethical considerations. Unfortunately, as I've already pointed out, P&W only include stimuli that (if successful) would minimize *some*, but not all potentially distorting influences on participant's judgments. Thus, the training provided is at best incomplete. P&W would therefore not only have to adequately induce competence with respect to the conflations and unintended interpretations they do address, but competence with respect to these other conditions, to genuinely establish competence. At best, participants who receive their instruction may demonstrate more competence than those who haven't received it. But *more* isn't the same as *sufficient*. Yet there is a deeper worry about their approach, which I develop in the next section. Here, I will note, briefly, that if their goal is to present participants with a theoretically neutral description of the relevant concepts, then they have not succeeded. While their descriptions of the various metaethical positions are accurate, and may not be controversial among most philosophers, they are *not* theoretically neutral: they take on board the distinctive assumptions of the mainstream philosophical and metaphilosophical presuppositions of contemporary analytic philosophers.

Finally, note that competence with some of the distinctions necessary to interpret questions as intended does not guarantee or necessarily help much with competence evaluating the meaning of response options or potentially confusing or ambiguous terms that go unexplained. Response options still employ a variety of obscure, ambiguous, or technical terms that P&W do not mention training people to understand.

### 3.11.3 Semantics

Finally, we are told that "ordinary speakers' intuitions about cases may sometimes be explained by their assumptions about the context of these cases or the intentions of the characters" (Pölzler and Wright, 2020a, p. 7). P&W provide an example: someone may judge two people who disagree about a moral issue are both correct because they suspect one of the people is insincerely contradicting the other person in order to antagonize them. Appropriate analysis of moral disagreements requires us to set aside such possibilities, since such an assumption would entail that at least one of the disputants was insincere, in which case we wouldn't know whether there was a genuine moral disagreement. Since judging both people to be correct only implies realism if people have a genuine disagreement over the normative moral facts, assuming this isn't the case would render judgments about the scenario irrelevant to assessing one's metaethical stances or commitments. In short: *all* pragmatic considerations that could threaten the intended interpretation must be excluded, or else a person's judgment simply isn't about the concept of interest. As they put it, "Intuitions that are grounded in such pragmatic considerations must be discounted as well" (p. 7).

Unfortunately, this requirement is almost completely ignored by P&W. Since participants were given both abstract and concrete measures, one of the things they did to induce semantic considerations was to ask participants to "explain any inconsistencies between their abstract and concrete responses within their concrete responses" which "was supposed to engage in reflection and allow testing the COMPETENCY and SEMANTICS requirements" (p. 13). Perhaps this could be used to assess competency, though it's not clear to me how. It's not at all clear, on the other hand, how this could be used to test semantics. While participants *could* be excluded if their explanations explicitly drew on contextual or pragmatic considerations that are irrelevant to the purpose of the study, such considerations may not be sufficiently salient to factor into people's explanations, people may not be consciously aware of them and thus couldn't report them, or their responses may not have

been coded adequately. In any case, relying on obscure and inadequately described open response data is hardly a robust method of meeting one of the *requirements* for the relevance of one's results. It also seems to be, at best, a feeble method. It *may* be useful in excluding participants subjects who clearly appealed to pragmatic considerations, but it (a) won't catch participants whose judgments were driven by pragmatic considerations but didn't mention these in their responses and (b) still relies on excluding people after the fact, rather than ensuring proper understanding in advance. (b) is especially worrisome, given that nearly a third of participants were excluded from analysis for failing various checks, a factor that can compromise the validity of the study by leaving researchers with a participant pool that doesn't represent the target population.

Almost nothing else is done to ensure that the semantic condition was met. I consider it fair to say that this condition was effectively ignored or at best downplayed to the point of de facto irrelevant for the purposes of this study. P&W don't train people in the distinction between semantics and pragmatics and don't employ exercises to train people to avoid pragmatic influences. They do virtually nothing specifically to minimize the influence of pragmatics. And by virtually nothing, I mean that they included the following instructions in the study, "We instructed them to 'focus on the information given by the sentence and […] not introduce additional assumptions or details about what happened or may have happened, or why'," which, they claim, "was supposed to decrease pragmatic influences" (p. 13).[169] I struggle to see this as anything other than a desperate act of handwaving at the issue of minimizing the influence of pragmatics. One cannot induce ordinary people to reliably ignore

---

[169] Again, I must emphasize my immense respect for the strides taken in this study to address methodological concerns. I don't say this to express empty platitudes. I've spent most of the past decade griping about methodological concerns in this kind of research. And without my prompting, other researchers noticed the problems, provided detailed accounts of them, and have gone out of the way to attempt to correct them. This is almost never done in any other area of research. The dedication to taking methodological worries seriously and getting innovative about how to avoid them is commendable.

semantics by simply *asking them to do so*. Competence at minimizing the influence of pragmatic considerations requires intense training and is difficult even for seasoned philosophers.

This is no small concern, either. Recall that they claim that ordinary people's intuitions may *sometimes* be explained by assumptions about context or intentions. *Sometimes*? This strikes me as an extraordinary understatement. Under what circumstances would we expect ordinary people to *not* think the context in which a morally relevant event takes place, and the intentions of the people involved, is relevant? Under what circumstances would ordinary people *not* make assumptions about the intentions of speakers? It's not even clear it makes sense to require people to ignore such considerations.

Contextual considerations may be *necessary* to express a meaningful moral judgment about a given case. It's not even clear that we have the psychological capability of imagining cases without imputing any assumptions about context, or to suppress any presumptions we may have about the intentions of the characters involved in scenarios that involve agents; nor is it clear, if we did so, that the situations would be intelligible any more, and, if they still were, it's not clear whether our judgments about such rarefied scenarios would be meaningful or bear any relation to our practical deliberations in actual cases in the real world. In fact, if our goal is to solicit judgments about what people mean when they make moral claims, it's not even clear it makes sense to attempt to isolate semantics and ignore pragmatics. This relies on the assumption that the meaning of ordinary moral claims is exclusively determined by considerations that have nothing to do with the intentions of speakers or the context of utterance. Why should we suppose that that's the case? I don't think this is how language works in general, much less with respect to moral claims. Why should we *require* ordinary people to adopt a view of the nature of language that may not even be very clear, or even true? While I share many of Kauppinen's (2007) reservations about experimental philosophy, I *do not* agree with Kauppinen's broader views about language. In short, P&W conform the requirements for their study

to a standard they probably don't even meet, but even if they could meet that standard, the standard turns on assumptions about the nature of language that reflect a substantive philosophical stance.

Can we simply *presume* ordinary people must conform their judgments to this philosophical position for the purposes of a study? We don't know if this position is true, we don't know if it reflects how ordinary people are already disposed to think, and if they aren't, we don't know how readily they could think in terms of such a standard. Finally, *even if they could*, this would at best only tell us what their metalinguistic position on the meaning of moral claims was *under counterfactual conditions in which they adopted a particular philosophical stance about language*. If they don't endorse these philosophical assumptions, this may not tell us what their actual position would be when they weren't required to adopt a philosophical position. And if they do employ this position in the course of a study in such a way that they endorse it, this could itself be a secondary form of spontaneous theorizing. That is, successfully inducing participants to focus only on semantics and not pragmatics could require training participants who did not previously draw the distinction, much less exhibit any competence with it, to do so, immediately before testing them. This would effectively involve teaching them to think, yet again, like a philosopher, even if they didn't think this way prior to participating in the study. And if the only way to accurately measure their metaethical position requires inducing them to think this way, then we're once again receiving a response from participants who have received distinct and unusual philosophical training that may not reflect how ordinary people think.

In short, the semantic condition isn't met by their study, but even if it were, the effort necessary to meet this standard may induce spontaneous theorizing and may require training participants so extensively that they were no longer ordinary people, in which case their responses would no longer generalize to ordinary people. Yet such concerns are, in this particular case, moot. They don't explain the role of pragmatics nor train people in how to avoid pragmatic influences. I'm not even sure this is feasible. I don't think I can avoid pragmatic influences in how I interpret sentences, and I'm more

familiar with their potential distorting or misleading influence than almost everyone in the world. Yet I'm supposed to believe participants can manage this because they are told "not introduce additional assumptions or details about what happened or may have happened, or why""? With respect for the many fantastic features of the study, which has done vastly more than anyone could have reasonably hoped for to minimize distorting influences and unintended interpretations, this is so far from adequate as a means of eliminating pragmatic influences that it's not much better than conducting a survey that requires people feel like they are motionless while on a rollercoaster, asking participants to "ignore your sense of motion," and expecting this to work.

Yet there are deeper, and subtler concerns about the way these questions are framed. Among philosophers, the metaethical commitments implicit in ordinary moral thought and language are taken to fix the referents of moral terms, which in turn plays some role in determining which metaethical account is correct. Yet the goal of this paradigm is to determine what the participant's *metaphysical* views are: i.e., do they think there are stance-independent moral facts or not? P&W's measures address this question in a mostly indirect way focusing on the participant's position on *descriptive* metaethics, alongside or in lieu of the participant's metaphysical position. In doing so, these measures presume that ordinary people share with philosophers the presumption that the correct metaethical account is predicated on what people mean when they make moral claims, and that we may therefore infer what participant's position on the nature of morality is by evaluating their stance on descriptive metaethics. Yet this is, itself, a substantive philosophical position.

Some of P&W's response options make exclusively semantic claims about what other people mean when they make moral claims. These elements of a response option concern the participant's judgments about ordinary folk semantics (such as their noncognitivist response option). Yet other items include metaphysical claims as well; for instance, their response option for error theory states that both people are incorrect "because although moral sentences intend to state moral truths, *there are*

*no such truths*" (p. 63, emphasis mine). Whether they include metaphysical claims or not, the set of response options participants are given do not directly or exclusively test whether they, personally, think that there are stance-independent moral facts. Rather, they indirectly assess participant's metaethical stances/commitments by inviting participants to select among response options that actually reflect that participant's stance or commitment towards *descriptive metaethics*, i.e., the participant's views on *what people mean when they make moral claims*. Yet why should we suppose that ordinary people think, as philosophers do, that whether realism is true or false depends on what other people mean when they make moral claims?

Take, for instance their noncognitivist option, that "Neither person is right or wrong." It includes the parenthetical "because moral sentences do not intend to state moral truths, and are therefore neither true nor false." This is an apt description of noncognitivism. Yet this response option presumes that *if* you believe ordinary moral claims do not express propositions, that *therefore* there are no stance-independent moral facts, but this is how this response option is interpreted: as evidence that the participant is a moral antirealist. Yet I don't think that the latter follows from the former. That is, I don't agree that if ordinary moral claims are nonpropositional, that *therefore* there are no stance-independent moral facts. This is because I don't accept that stance-independent moral facts depend for their existence on our account of what ordinary people intend when they make moral claims.[170]

I'm not alone in thinking this. In a recent article, "Must Metaethical Realism make a Semantic Claim?" Kahane (2013) answers the titular question with a definitive *no*. As Kahane argues, "Robust metaethical realism is best understood as making a purely metaphysical claim. It is thus not enough for antirealists to show that our discourse is antirealist. They must directly attack the realist's

---

[170] In case this seems like a convenient way to continue to endorse moral realism by divorcing metaphysics from folk discourse, note that I am *not* a moral realist. It would be all too convenient to appeal to existing empirical evidence, which seems to favor folk antirealism, or to my own account, indeterminacy, and make a direct inference to antirealism. Yet I just don't think whether there are stance-independent moral facts has much to do with what people are doing when they make moral claims.

metaphysical claim" (p. 148). If Kahane is correct, then the presuppositions implicit in these questions rest on a mistaken conception of the relationship between language and metaphysics, a conception so deeply embedded in the way the questions are framed that the questions *make no sense* if that assumption is not granted. And yet *I* don't grant it. In fact, I'm quite confident it's completely mistaken. This is not as bizarre or unconventional a position as one might suppose, if one takes a bird's eye view of the history of philosophy. The focus of contemporary metaethics that culminated in descriptive metaethics serving front and center in the analysis of moral thought is a legacy of the *linguistic turn*, a period in Western philosophy that began in the early 20th century that led to a pronounced emphasis on the relation between language and reality. The recent focus of every prominent metaethical position has been a largely descriptive enterprise, with philosophers debating what ordinary people mean when they make moral claims. This is reflected in the three straightforwardly antirealist options P&W provide: individual subjectivism, error theory, and noncognitivism.

Ironically, as an aside, responses to these questions *require participants to express a uniform and determinate stance towards the meaning of moral claims*. In other words, despite Pölzler and Wright's own findings consistently suggesting metaethical pluralism, participants themselves have no way to express such a view, but must necessarily express a stance about the meaning of ordinary moral language that their own studies suggest is probably false. Why would we design studies that force participants to choose from among only options that those very studies suggest are mistaken?[171] This observation serves to reinforce more or less the same point: by enriching their study with additional instructions,

---

[171] This is a more complicated version of asking participants to choose whether (a) *everyone* likes pineapple on pizza or (b) *nobody* likes pineapple on pizza. It's absurd in this case. It's equally absurd for research on folk metaethics, since it requires us to presume ordinary people are committed to *precisely* the mistaken meta-semantic errors that Gill (2009) attributes to 20th academic metaethicists. Yet this is the very article that served as one of the primary inspirations for conducting this research in the first place! It is baffling that, in suspecting philosophers share catastrophically mistaken assumptions about folk morality, that we would attempt to demonstrate this by restricting response options only to positions that share those same (likely mistaken) assumptions.

training, and clarifications, Pölzler and Wright are effectively training participants to adopt and think in terms of the distinctive metaphilosophical presuppositions that characterize contemporary analytic philosophy. That is, the only way to clarify the relevant metaethical concepts and distinctions effectively *requires training participants to do analytic philosophy*. What's more, it requires them to think in accordance with mainstream about the relation between language and metaphysics. There's nothing theoretically neutral about that.

## SUPPLEMENT TO CHAPTER 4

**S4.1 Additional discussion about coding procedure for interpretation rates**

In general, coding is a difficult process that requires an understanding of researcher intent[172], the relevant metaethical distinctions, the ways these distinctions might be conflated with other, similar distinctions (e.g., stance-independence is *not* the same as universalism), a sensitivity to the way nonphilosophers might express metaethical concepts without technical jargon (e.g., if a participant says that moral facts "don't depend on us," this is fairly close to stance-independence), and a reasonable degree of charitability.

To complicate coding further, a competent coder may also require experience both with coding responses to questions about metaethics *and* knowledge of empirical data about folk metaethics. To illustrate why, consider instances in which a participant states that "morality is relative" or "morality is objective" without any additional context or details. Should these responses be interpreted as clear instances of intended interpretation? It might seem uncharitable to critics to code these responses as unclear intended interpretations (1|0) rather than clear ones (1|1). However, when I explicitly asked people what it meant to say that morality was objective or relative, virtually no participants clearly interpreted either term in a way consistent with how academic philosophers use them. Given these findings, it is difficult to see why we should presume anyone who invokes terms used among academic philosophers understands those terms in the way academics do.[173]

---

[172] Which can make coding responses in a way blind to hypotheses very difficult, since elements of researcher intent could give away researcher concerns or expectations.

[173] We could lean towards charitability in interpreting these questions as clearly intended (1|1). However, doing so would involve coding responses in a way intended to mitigate researcher bias, rather than because the coding accurately reflects our honest assessment of whether the participant interpreted the stimuli as intended. I do not believe it would be best to deliberately code data in a way that I think is inaccurate merely as an attempt to mitigate my biases, since such coding would no longer reflect a genuine attempt to interpret participants, but would instead inextricably entangle coding of their responses with my own concerns and anxieties about my personal biases. Coding should be based on our sincere judgments about whether the participant interpreted questions as intended, and either did so clearly or unclearly. It should not incorporate heuristics intended to mitigate our own biases that results in coding items in an insincere way.

Nevertheless, a reasonable case could be made for coding these items as clearly intended interpretations. One reason is that invoking considerations that are themselves the result of the very methods I am employing could be seen as circular, self-justifying, and far too contingent on heuristics about base rates that depend on my assessment of a narrow slice of data. Perhaps other studies would show that a higher proportion of people interpret metaethical terms in a way consistent with the way they're understood by academics. Or perhaps my coding of how people interpret explicit metaethical data grossly underestimates the true proportion of people who interpret these terms consistent with researcher intent.

These are reasonable concerns, but if we take them *too* seriously, this would only raise the bar for coders even further, since the implication might be that we need much more data about how people interpret terms like "objective" and "relative" before we can even begin to code responses that use these terms. This raises a more general worry: *any* terms participants use could be understood in a variety of ways. Do we need comprehensive data on how each respondent understands each word they use? Or at least some of the key terms? Once we open Pandora's box with concerns about variation in meaning, the deluge of data necessary for fine-grained coding would require knowledge and experience beyond what is available to mortals.

Unfortunately, I am mortal, and any coders I could recruit would likely be mortal, too. We need to make judgment calls, but any judgment or cutoff criteria we employ may seem arbitrary or unprincipled. The necessity to make decisive judgments in the absence of perfect access to what participants mean points to a more general consideration: each coder brings their own priors, background assumptions, and knowledge to bear on the data, and these and other factors all play a role in their judgments. And since variation between coders is not directly accessible (at least not in a precise and comprehensive way), both my own coding and anyone else's will invariably introduce noise into the way we code. Some of this noise may manifest as variation between coders, but noise

might also reflect systematic inaccuracies shared between coders. For instance, if I primarily recruit coders with particular educational backgrounds or shared theoretical assumptions, consistency between coders might misleadingly suggest accuracy in our coding, when we could simply be miscoding the data in the same way as one another.

Some noise in our coding (whether shared or unshared) will result from performance errors, but some will reflect differences in our degree of *charitability*. How charitable should a coder be? One way to circumvent these concerns is to simply code all such responses as clearly intended interpretations. This could even be part of a general strategy: since I expect (and, in truth, *hope*) to find that few participants interpret metaethical stimuli as intended, I am undoubtedly subject to confirmation bias. As a precaution, I could adopt formal coding rules that would limit the degree to which biases could result in underestimating the true proportion of intended interpretations. However, I opted against this. The goal of coding ought to be to approximate, as closely as possible, how participants *actually interpreted stimuli*. Adopting precautionary rules might serve to mitigate anxieties about my own biases, and mollify skeptics of my results. Yet it would achieve these results by (from my perspective) strategically misrepresenting the data.[174] Whatever my biases may be, the best corrective against them would be to identify adversarial coders who hold contrary expectations, and have them code the same data. In publishing these results, one of my goals is to invite just this type of scrutiny, and if my conclusions seem shocking or dubious to others, perhaps this will provoke critical reevaluation of the raw data.

If we do not adopt a principle of maximal charity, then how charitable should we be? Unfortunately, any such answer will involve theoretical and normative assumptions that fall outside

---

[174] For comparison, if a well-designed quantitative study resulted in an unexpectedly large effect size, it would be inappropriate to eschew reporting these results, and to instead begin reanalyzing the data in order to identify a result with a lower, more plausible effect size merely to report a less surprising result that the researcher would expect critics to find more palatable.

the scope of a direct emphasis on the descriptive facts themselves. In short, all coding must be filtered through particular theoretical lenses and normative considerations. I have, in ruminating on the matter of how to best code responses, simply recapitulated the fundamental limitations of qualitative research. At the very least, in reflecting on these matters I hope to at least demonstrate my awareness of these issues and the assumptions motivating my approach to the data.

Aside from the difficulties of coding items in accordance with my quaternary scheme, there is also the question of why I opted for this approach. The goal of this coding scheme is to balance simplicity, flexibility, and to serve as a buffer against coder bias by conserving information that could conflict with my expectations.

In principle, I could present only two categories: whether items were interpreted as intended or not. This approach would have the advantage of being maximally simple, and many responses do appear to clearly reflect an intended or unintended interpretation. However, there are at least as many items that are much harder to code. Simplicity must be weighed against explanatory adequacy, and in this case I judged that using only a binary coding scheme would compress findings in a misleading and unhelpful way. This is because, when reviewing early datasets, I found that a substantial proportion of responses were vague, ambiguous, or borderline to such an extent that I could not reasonably include them in the same category as responses that more clearly reflected an intended interpretation. Many items looked like they *might* show that the participant interpreted the question as intended, but there was no way to be confident. In other cases, they did not show any explicit indication that they interpreted stimuli as intended, but they made remarks that hinted at an unintended interpretation. It did not seem appropriate to lump ambiguous or borderline cases in with responses that did seem to more clearly reflect an intended or unintended interpretation. Doing so would dilute the informational force of both clearly intended and unintended interpretations.

Categorizing items as clear or unclear can be difficult. There is no bright line that divides clear cases from unclear cases, and there are numerous responses that reasonable people could categorize differently. Yet the existence of borderline cases does not obviate the need for drawing distinctions where such distinctions are warranted. Just as clinicians may use cutoffs when diagnosing patients with mental illnesses, despite symptoms existing along a spectrum, coders must also draw the line somewhere when assessing the clarity of a response. While this introduces an unavoidable degree of error, enough responses are sufficiently clear or unclear that they appear bimodally distributed, and, in most cases, there is little difficulty in categorizing a response as clear or unclear.

It may be helpful to review concrete examples of responses I coded into each of the four categories, which may be seen on **Table S4.1**. This will serve both to provide a general sense of the data and to explain my reasoning in advance of the analysis provided for any particular study. For each of the four categories, I have selected three examples. These will appear on the left column, along with a description of the rationale behind my coding on the right. All examples were taken from studies that asked participants to state, in their own words, what it meant either to say that moral truth is objective or relative or that a concrete action (murder or abortion) was objective or relative.

**Table S4.1**

*Examples of responses coded into each category*

| Clear intended interpretation   1|1 | Explanation for coding |
| --- | --- |
| *(1) It means that it is not based on anything other than cultural ideas about the value of individual human life. It is an idea borne of human society and not some law of the universe.* | Clearly contrasts culturally constructed (and thus stance-dependent) standards with "some law of the universe," which plausibly represent stance-independent standards. |
| *(2) You are saying it is an objective fact murder is wrong. An objective fact is one which is not based on human judgement or belief.* | Uses explicit metaethical language ("objective") and characterizes it as stance-independent. |

| | |
|---|---|
| *(3) Right and wrong is not subjective* | Participants sometimes used explicit metaethical language consistent with the intended interpretation. These were coded as clearly intended even though my findings suggest most people don't understand metaethical terms like "objective" and "subjective" in the same way as researchers (i.e., to reflect stance independence/dependence). |

## Clear unintended interpretation   0|1

| | |
|---|---|
| *(4) The meaning of moral truth being objective is when something is already considered by the major majority of the masses to be the moral truth. Thus being considered this by the major majority it becomes no longer subjective by an individual.* | Interprets objectivism to refer to strong consensus. Strong consensus does not entail stance-independence. Note that studies have found a strong correlation between "objectivism" judgments and perceived consensus (Goodwin & Darley, 2008; 2012) |
| *(5) It means what you believe is your truth and when you use your moral judgement it is unbiased* | "Objective" is a polysemous word that also refers to rendering a judgment in an impartial or unbiased way. This is distinct from "objective" as it is used in metaethics (Oxford University Press, 2021). |
| *(6) It means that whether an act is right or wrong can be determined precisely or in a binary manner based on the facts. And there is no continuous measure of morality or exceptions to the rules.* | This response construes objectivism as the view that morality is categorical and binary (e.g., whether a number is prime or not prime) rather than continuous (e.g., height). This is a coherent way of distinguishing moral claims, but it is unrelated to realism/antirealism. |

## Unclear intended interpretation   1|0

| | |
|---|---|
| *(7) What is moral? It can be relative because morals are relative to each individual.* | Response is consistent with an intended interpretation and shows marginal familiarity with the intended notion of relativism with the phrase "relative *to each individual*," but is insufficiently clear. Responses like these often rephrase what was asked using the terms in the question, creating the superficial appearance of an intended interpretation. |
| *(8) Moral truth is objective because something is either true or false* | Something being "either true or false" is consistent with antirealist positions. Most closely resembles cognitivism or the notion that truth |

| | and falsehood are categorical rather than continuous. |
|---|---|
| *(9) Murder is morally wrong relative to the viewpoint of an individual. Every person has a different set of moral principles.* | This is a borderline case. It appears to reflect an intended interpretation in the first sentence, but the second sentence suggests a descriptive interpretation. |

**Unclear unintended interpretation   0|0**

| | |
|---|---|
| *(10) I believe moral truth is relative. We must be morally truthful in our everyday lives. That is how we should be. I would be involved with someone who is morally truthful.* | An intended interpretation would involve explaining what relativism means. However, this respondent stated that they agree with relativism, then made several comments unrelated to the question. While irrelevant, these comments do not show that they did not interpret "relativism" as intended. |
| *(11) being for the better good of the world* | It is unclear what this response means. There is something normative or evaluative about it, but it is not obvious that they interpreted the question in an unintended way. Some responses may suggest low engagement with the question. In these cases, it is hard to judge the reason for the unintended interpretation. |
| *(12) YES, BECAUSE THE COURT IS THERE FOR THE PUNISHMENT GAVE.* | Some responses do not appear to engage seriously with the question. This may involve not answering, writing (e.g., "n/a,") writing a single word (e.g., "MORALITY"), or, in this case, writing something that is only vaguely relevant to the question.[175] |

---

[175] The rationale behind coding these responses as unclear unintended interpretations is a bit more complicated than some of the other coding decisions. The goal of clear unintended interpretations is to capture instances in which the participant appeared to have seriously considered the question, but interpreted in an unintended way. I distinguish this type of unintended interpretation from those where the participant plausibly did not interpret the question as intended simply because they didn't consider the question (or at least did not consider it sufficiently) at all. The goal of identifying clear unintended interpretations is to assess whether people struggle to interpret questions about metaethics even when they are genuinely attempting to do so. If they explicitly make a remark that suggests that their interpretation is not consistent with researcher intent, this provides evidence (if only a single "unit" of it) that some people actively interpret metaethical stimuli in unintended ways. When a participant's response is nonsensical, uninterpretable, or suggests low engagement with the task, the participant plausibly did not interpret the question as intended. However, this may have little to do with the substantive content of metaethical stimuli as such. Sometimes people are distracted, bored, or dislike an experimental task. Yet this can lead to irrelevant responses even for questions a person would not plausibly interpret in an unintended were they engaged with it. For comparison, suppose you asked participants to "please describe a bank." Suppose you intended for them to interpret this as a question about financial institutions. If a participant interpreted it as a question about riverbanks, this would be evidence that your question was ambiguous. If someone simply did not answer your question, or said something that suggested they were not engaged with the question, such as "I hate open response questions" or "I like cheeseburgers," this would provide little information about whether the stimuli in question were ambiguous. Such

Another reason to distinguish clearly unintended interpretations from unclear interpretations is that instances of the former provide greater evidential support for or against my hypothesis. If many participants offered ambiguous answers or did not answer at all, we might suspect that their failure to clearly express an interpretation consistent with researcher intent could be attributable to some other cause than an unintended interpretation. For instance, they may have lacked the motivation to provide a clear response or to even respond at all, since doing so may seem tedious and require more effort than responding to a multiple choice question. Or they may have interpreted the stimuli as intended, but have difficulty articulating their understanding clearly. Either way, a significant proportion of people who interpret stimuli as intended may fail to clearly demonstrate that they did so when presented with an open response question. If many participants provided unclear responses like these, there would be less concern that validity was compromised by low rates of intended interpretations.

If, on the other hand, there are many instances of clearly unintended interpretations, this would provide a more direct challenge to a study's validity. Rather than simply not knowing whether they interpreted stimuli as intended, we would have direct, positive reason to believe they *did* interpret in a way inconsistent with researcher intent. In the preceding chapters, I highlighted several distinct ways participants conflate questions about metaethics with questions about other distinctions. If participants often respond in a way consistent with these predictions, this would provide especially strong evidence that they did not interpret the question as intended, since such interpretations would not be random or unprincipled, but reflect distinct and plausible reactions to ambiguous stimuli that are theoretically grounded in preexisting explanations of the ways we should expect people to interpret metaethical stimuli in unintended ways.

---

disengagement may result from finding the question ambiguous or confusing, but it could be a general phenomenon that has nothing to do with the particular stimuli you are asking about. When this occurs, we have little direct evidence about the distinctive features of the stimuli (e.g. ambiguity), and cannot be confident that the response raises methodological concerns about the stimuli being used.

For instance, suppose researchers asked participants about a "bank," and had financial establishments in mind. Yet they happened to present their questions to a population that lived along a river. We may be unsurprised if many of these participants interpreted the question to refer to a *riverbank*, and if they did, we could confidently conclude that their interpretation of the question was *not* what researchers had in mind. In such circumstances, it would make sense to ask participants "What does 'bank' refer to?" If most participants described a riverbank, this would be better evidence that they interpreted the question in an unintended way than if they didn't answer at all. Likewise, identifying distinct ways participants interpreted questions about metaethics in unintended ways provides stronger evidence than focusing exclusively on the proportion who did interpret questions as intended.

While there are recognizable advantages to distinguishing clear unintended interpretations from unclear interpretations, I could have employed a trinary system: clear intended, clear unintended, and unclear. However, failing to distinguish between unclear interpretations that lean more towards an intended interpretation from ones that lean more towards an unintended interpretation would give up potentially valuable information without justification. One of the greatest limitations with these findings is that they rely on my own subjective evaluations. Since there is no way for me to be blind to my own hypotheses, this exposes results to considerable risk that my judgments are biased in favor of my hypothesis.

For instance, confirmation bias could lead me to uncharitably miscategorize participants who plausibly did interpret stimuli as intended as "unclear" without qualification. By distinguishing between unclear responses that *may* reflect intended interpretations from those could not plausibly be interpreted as clear interpretations (including, e.g., people who simply did not respond at all), I have erred on the side of caution by preserving any instance in which I believe a reasonable person could have coded a response as an intended interpretation from those where this would not be reasonable.

This allows me to place a tentative upper bound on the number of participants who could plausibly have interpreted stimuli as intended. Of course, even this qualification has limitations. I could still miscategorize some responses as seemingly unintended interpretations even when this is not the case. Hopefully, the inclusion of unclear but seemingly intended interpretations will have at least some palliative[176] force with critics.[177]

In sum, a quaternary system allows me to balance optimal informational resolution against the need for simplicity, and serves to partially mitigate concerns about underestimating clear intended interpretations. Finally, by explicitly distinguishing clear instances of unintended interpretation, I can provide stronger evidence for the predicted low rates of intended interpretation and support the contention that metaethical stimuli are frequently conflated with unintended concepts and distinctions. Any advantages to a simpler coding scheme would be easy to achieve by collapsing categories into a trinary (clear intended, clear unintended, unclear) or binary (clear intended, not clearly intended) scheme, anyway, allowing for ease of presentation in any context in which this would be appropriate.

**S4.2 Additional commentary on general procedures**

*S4.2.1 General methods*

In the main text, I reference the absence of significant qualitative research in experimental philosophy. I find the lack of such studies somewhat surprising. Experimental philosophers have acknowledged

---

[176] I chose the term palliative despite the fact that the term ameliorative immediately came to mind. The latter is overused, and I dislike that philosophers make excessive use of shibboleths that function more to signal their affiliation with their elite colleagues than to express themselves clearly or make use of the typical and more desirable degree of terminological variation that would characterize good writing. For instance, philosophers frequently use the term 'gloss' in strange and unconventional ways.

[177] Worryingly, the thought occurred to me that including this more complicated coding scheme may be unconsciously motivated by a desire to discourage deeper scrutiny by signaling awareness of my biases and efforts to correct for them. Even this comment may reflect a meta-signal that I am aware I may be signaling. There is no escaping the recursion of potentially Machiavellian motives, so the best I can do is encourage critics to ruthlessly evaluate my claims and the data itself.

the importance of utilizing a broader range of methods (Nadelhoffer & Nahmias, 2007), while others have explicitly drawn attention to the importance of augmenting research in experimental philosophy with qualitative methods (Allen et al., 2020; Andow, 2016; Moss, 2017; Thompson, 2022). Moss (2017) in particular has advocated the use of qualitative methods specifically for studying folk metaethics, in part precisely *because* of the methodological difficulties discussed here (see also Bush and Moss, 2020). Why have no prior studies approached folk metaethics in the way I do here? Perhaps the stigma against qualitative research and the difficulty of publishing these findings would discourage researchers. Furthermore, this research is incredibly time-consuming and requires distinct and narrow expertise in whatever area of inquiry one is studying, limiting the generality of the method for any given research. For instance, while I may be an expert judge of responses to questions about metaethics, my competence would drop off a cliff the moment my attention was turned to most other topics.

Not only is it time-consuming, but my findings are limited, in most cases, by the fact that I was the only one to code them. The ideal approach to assessing qualitative data would be to have at least two coders. And since coding responses properly would require both a familiarity with and motivation to engage in psychological research, *and* competence with metaethical positions, there have been few people with the time and ability to actually assess items like these. However, the main reason this research has not been done may be that not much time has elapsed since researchers began identifying severe methodological problems with research on folk metaethics. Even those researchers who have previously identified shortcomings with existing research (e.g., Pölzler, 2018b) may not share my pessimism about the extent and scope of the problems this research faces. Finally, the psychology of metaethics has yet to reach a broad audience. Few researchers study the psychology of metaethics, and there has been little commentary among philosophers or psychologists about its findings. Taken together, it is easy to see why no prior studies have adopted the approach taken in this chapter. The work is laborious, disincentivized by the current publication environment, unfamiliar

to most people, and unlikely to seem necessary even to those familiar with empirical research on folk metaethics.

### *S4.2.2 General procedures*

For all datasets, each response was extracted from either a Qualtrics file (when data was collected by myself or my colleagues) or from its original source if it was provided by another researcher (e.g., an Excel file). Responses were then posted to a separate Google Sheets document. It is possible that this process introduced errors or altered text, but no such instances were discovered. All text was converted to Calibri 11-point font. All responses were presented in the order they appeared in the original data file in a single column. This was followed by a column for whether the interpretation was intended ("1") or unintended ("0") and a second column that indicated whether the item was clear ("1") or unclear ("0"). All coding sheets included column headings in the first row. The second row included reference information for the coder's convenience. This included the original question presented to the participant, and an informal question to guide how coders were to code items for that dataset. When a dataset included multiple items, these items appeared in bold in the main text of the column itself, rather than in the reference row (that is, the second row). I have included an example dataset; see **Figure S4.1.**

# Figure 5.1

| Item | Response | Interpretation | Clarity | Code | Notes | theme_1 |
|---|---|---|---|---|---|---|
| | QUESTION: In your own words, what does it mean to say that the truth of the moral claim "[statement]" is [objective/relative]?<br><br>CODE: Does response clearly demonstrate an intended interpretation? (any metaethical interpretation) | 1 = intended<br>0 = unintended | 1 = clear<br>0 = unclear | | | |
| | **Concrete \| Objectivism \| Murder** | | | | | |
| Item #1 | **In your own words, what does it mean to say that the truth of the moral claim "murder is morally wrong" is objective?** | | | | | |
| 1 | It means that all people everywhere should generally agree that murder is wrong and that this statement is essentially based on principles or facts that are beyond question. | 1 | 1 | 11 | Universalism, restatement | correct |
| 2 | It means that "murder is morally wrong" is a truth recognised by all individuals to be morally wrong. | 1 | 0 | 01 | Consensus, descriptive | universal |
| 3 | It strives for a goal. By saying it is morally wrong, it is something that people will live by. | 0 | 0 | 00 | Prescriptive, ideal, goal | prescriptive |
| 4 | It's a grey area. It can't be defined so generally | 1 | 0 | 01 | opposite | black_and_white |
| 5 | It means that it is up to the beholder to determine the truth the statement and that it differs. | 1 | 0 | 01 | opposite | opposite |
| 6 | it may be objective because to someone murder may be justified to re-pay for another murder that was done. | 0 | 0 | 00 | justification | context |
| 7 | It means there exists an absolute system of morals by which we can measure actions, and murder is wrong by that system | 1 | 1 | 11 | absolutism | correct |
| 8 | What is morally wrong or right is based on society's social norms and culture of the time. | 1 | 0 | 01 | opposite | opposite |
| 9 | Because each society and culture has their own sense of morals | 0 | 1 | 10 | Descriptive | descriptive |
| 10 | Humans are evolved so that a psychologically healthy human will feel guilt when they murder, because if murder went unchecked, entire groups of humans would be killed and the rate of survival would decrease considerably. | 1 | 0 | 01 | irrelevant | etiology |
| 11 | No matter who or where or what is happening, it is bad to kill on purpose. | 1 | 0 | 01 | irrelevant | normative |
| 12 | You are saying it is an objective fact murder is wrong. An objective fact is one which is not based on human judgement or belief. | 1 | 1 | 11 | Mind-independence | correct |
| 13 | Morally speaking one shouldn't do something that causes harm or pain to another person. "Do unto others..." That would include murder unless it was an accident. | 1 | 0 | 01 | Normative | normative |
| 14 | Every society everywhere has laws against murder. It's a very anti-social act, if not the ultimate anti-social act. If people run around killing each other, there won't be many people left. | 1 | 0 | 01 | Irrelevant | normative |
| 15 | In my own words, murder IS morally wrong. There is no objection when it comes to it, a human should not take the life of another human no matter what the circumstances are. I personally do not believe in the death sentence due to this. | 1 | 0 | 01 | Irrelevant | normative |
| 16 | I think the statement is objective because there are ways to murder with cause that is not morally wrong (war, defending self, etc.). | 1 | 0 | 01 | opposite | context |
| 17 | According to the laws, and the bible, it means it is not right to murder another person, or no one has the right to murder someone else. | 0 | 1 | 10 | theological, unclear | normative |

I reference these details because it is possible that the way researchers organize and present data influence how items are coded. For instance, the inclusion of the question "Does response describe an objectivist/relativist interpretation?" or the headings "0=Unclear 1=Clear" may influence how I coded responses. Even if these effects are small, consistency in the subjective experience of the coder is at least worth keeping in mind when assessing intercoder reliability. Note that headings and formatting for the data was changed at several points during coding, as I developed a more streamlined format. Thus, the format presented in OSF does not necessarily represent the format present when initially coding, or in subsequent recoding. This may have had some influence on how I coded. While I doubt, if it did, that it was significant, it is still worth noting that changes were made throughout the process that could have had some minor influence on results.

Prior to analyzing a given dataset, I had some *a priori* conception of what would constitute a clear instance of an intended or unintended interpretation. For instance, when assessing Goodwin and Darley's (2008) questions, which asked participants what they thought the source of their moral disagreements with other participants could be. In order for a participant's response to indicate that they clearly interpreted the source of disagreement as intended, their response would have to indicate that they attributed the disagreement to a genuine difference in moral beliefs, rather than some other cause (e.g. the other person misunderstanding the question, or thinking of a different situation than the participant). Thus, I set the criteria for a clear interpretation to be one in which the participant attributed the cause of the disagreement to a *fundamental* moral disagreement, rather than some other cause (see e.g., Bush, 2016; Bush & Moss, 2021). For each dataset, I included (i) context about how the question was asked (where appropriate), (ii) the original item that the participant was presented with, and (iii) a question directed at the coder (usually myself) intended to guide the coding process.

I employed a similar process for other datasets. For instance, if the participant was asked what the term "objective" means in the context of moral claims, an intended interpretation would require

alluding in some way to stance-independence, while if the prompt asked the participant to explain why they agreed or disagreed with a statement on a metaethics scale, an intended interpretation would require any articulation of any stance related to realism or antirealism.[178]

All information relevant to coding was presented in bold at the top of the file directly above the items. The row this information was presented in was "frozen" so that it would be visible as the coder scrolled down, which allowed it to remain visible throughout the coding process.[179] All datasets were included in a single Google Sheet file with a table of contents and an alphanumeric code used to abbreviate datasets, which are each presented separately in their own tabs accessible at the bottom of the screen. Moving from left to right, columns typically included:

(1) A number representing the order of the item as it initially appeared in datasets downloaded from Qualtrics. Order of items was not changed for most datasets, though where there were exceptions an additional column was included immediately to the right of the order of items as coded that represents the order of the items as it appeared in the original dataset.

(2) Condition. Some datasets include multiple conditions. These are featured in a second column, often with some abbreviation, e.g., "o1" may refer to objectivism condition #1.

(3) The participant's response. These responses are presented in unedited form.

---

[178] For some datasets, I recruited additional coders. Both coders have at least some background in empirical research on folk metaethics and have collaborated with me in the past. Thus, they both had at least some prior experience with the topic. One coder, David Moss, also specializes in research on the psychology of metaethics, while the other, Tyler Millhouse, is familiar with metaethics and with my views in particular. Unfortunately, the cost of this familiarity and experience is that both coders were not blind to my hypotheses. In addition, David Moss shares my skepticism about the proportion of people who interpret metaethical stimuli as intended, while Tyler Millhouse shares my skepticism about moral realism. Both hold distinct stances towards metaethics and their own distinct perspective about folk metaethics. Neither coder may claim to be naive about the data, but are instead susceptible to similar biases in coding as I am. Finally, in both cases, disagreements between coders were resolved via flagging points of disagreement and discussing them. In some cases, we were able to reach consensus, but when we could not, I left those differences unchanged. Thus, for each other coder, there are two sets of code: their coding prior to discussing and assessing the data to reach consensus, and after doing so. The coding prior reflects a more unadulterated perspective on the data, while the latter reflects a more careful assessment of the data on both our parts. While I consider the post-discussion coding a more accurate reflection of the data, it also resulted in much higher interrater reliability, which could misleadingly give the impression that coders independently assessing the data.

[179] "Freeze" is a term that appears in Google Sheets options. It can be located by going to "View" in the toolbar, then selecting "Freeze." You can then select how many rows or columns you want to freeze from a dropdown menu.

(4) Whether the interpretation was intended or unintended.

(5) Whether the interpretation was clear or unclear.

(6) The final code, formed by concatenating the input in the relevant row for the previous two columns. For instance, if an item was coded as "unintended," this would result in a "1" in the interpretation column, and if it were coded as "unclear," it would result in a "0" in the clarity column. Concatenation simply takes these puts the contents of whatever is in these columns together in order without spaces. In this case, the result would be "10." This would indicate an interpretation of 1|0, an unclear intended interpretation.

(7) The next column is "notes." Notes includes proposals for how to code the item and any other remarks deemed relevant.

(8) The final column is "Themes," which represents the final set of themes for a particular item.

(9) I included a summary of the percentage and total count for the quaternary coding scheme at the bottom of each dataset.

Study 1A was coded by David Moss as well. I discuss these details in the sections addressing those specific studies.

### S4.2.3 A note on the second reanalysis of Goodwin and Darley (2008)

Preliminary results of my reanalysis of the data presented in Goodwin and Darley (2008) were initially reported in Bush and Moss (2020). In that article, I reported 41% clear intended responses, 44% clear unintended responses, and 15% unclear responses. This data represents my initial coding, prior to recruiting David Moss to code the data as well, and prior to any attempt to resolve disagreements with David. There is a recognizable drop in clearly intended interpretations (and a comparative increase in clear intended interpretations for David, after discussion and attempts to resolve disagreement). However, I believe this drop is justified, and that my initial coding was excessively charitable.

In the interests of transparency with respect to my initial reanalysis of the data and reanalysis following coding, I have made both the initial coding results available in OSF. I have also made David's initial coding available, prior to our discussion. I am very receptive to criticism of my initial coding, my current coding, and any other coding I have conducted. In addition, disagreements were tracked

and resolved in Google Sheets. This process involved scanning and identifying disagreements, after which David and I met online. During this meeting, we discussed our reasoning for coding items the way that we did, and attempted to reach an agreement about the proper way to code items that we disagreed about. During this process, David pointed out numerous ambiguities and alternative ways of interpreting responses that undermined my confidence that a clear intended interpretation was most plausible, which caused many items to shift towards 1|0 or 0|0. On the other hand, I explained the rationale behind coding some items as clear intended (1|1), and was on occasion able to make a compelling case for a preponderance of reasons in favor of interpreting an item as clear intended interpretation. Thus, neither David nor I were uniformly moved towards the other's perspective, but instead converged on an overall picture of the data somewhat intermediate between his less charitable and my more charitable interpretation of responses.

Nevertheless, while we converged on what I am confident is a more accurate analysis of the data, I do not claim to have coded "perfectly," nor do I think such a goal is achievable. There really is an eliminable degree of subjectivity in the way any particular researcher codes. Nevertheless, this does *not* indicate that such coding is irrelevant, worthless, or somehow epistemically inferior to quantitative analysis.

For comparison, it would be absurd to conclude that nobody is capable of judging the culinary skills of chefs *merely* because there is an ineliminable degree of subjectivity in assessing the quality of a dish. I encourage anyone who disagrees to seriously suggest that we can draw no meaningful conclusions about the comparative quality of Mexican restaurants with Michelin stars and a Nacho Cheese Doritos® Locos Taco from Taco Bell. Quantitative analysis may permit greater *precision* for well-specified questions, but quantitative methods are not feasible for some questions, and in those cases where they are used, they are often little more than a proxy or downstream measure of some subjective mode of evaluation, anyway. After all, we could give food critics Likert scales and ask them

to score food from different restaurants. This could provide insights and some level of precision in certain quantifiable measures that wouldn't be obtained by reading a Yelp review; yet it would be absurd to suggest that a variety of important insights wouldn't be lost in this process.

Any attempt to transmogrify data or observations that have a substantive qualitative character into something quantitative violates a cousin (if a distant one) of the second law of thermodynamics, what we might call the "second law of social scientific informational dynamics": at least some information will be lost in this process; it's not a matter of *if*, but *how much*. In many cases, such losses will be worth the gains in increased precision and the capacity to subject data to statistical analysis. But in some cases, the loss will be significant enough that such methods are simply not appropriate. Psychologists ought to appreciate this fact and (at least some) should stop operating under the dogmatic pretense that quantitative psychological findings are strictly superior to qualitative methods.

Nevertheless, merely demonstrating that participants attributed the source of the disagreement in a pair of studies does not, by itself, indicate that there is a widespread tendency for participants to interpret questions about metaethics in unintended ways. More importantly, even if the disagreement paradigm was not a valid measure of folk metaethical stances or commitments, this would not demonstrate that metaethical indeterminacy is the best account of the way ordinary people think about moral issues (i.e., that however people think about moral issues, they don't typically have any determinate stances or commitments about whether any particular moral issue, or morality in general, is realist or antirealist). One way of assessing whether ordinary have determinate metaethical standards would be to more directly assess whether they interpret expressions of realism and antirealism as intended. In the next set of studies, I assess whether people interpret *responses* to questions about realism and antirealism as intended.

**S4.3 Additional commentary on general predictions**

To provide an example of why making predictions is not only difficult, but actively harmful, consider the implications if a coder knows that the predicted rate of intended interpretations is expected to be at a particular rate, such as 40%. Halfway through coding a set of data they notice that intended interpretation rates seem to be much higher than this. If so, they may become more critical in how they code subsequent items in an unconscious attempt to adjust the response rate to fit the desired proportion. Their ability to reevaluate items or otherwise move between the phases in accordance with the iterative nature of reflexive thematic analysis would also be compromised, since there would be incentive to recode previous items to reach a desired proportion or to fit desired proportions of particular themes. This need not be conscious malfeasance; a coder could sincerely believe that their adjustments are justifiable course-corrections, and they may even be able to provide good reasons for biased coding, convincing themselves and even others that their decisions are appropriate. It may therefore be much better for coders to actively avoid making any precise predictions, to avoid a self-fulfilling prophecy driven by unconscious bias.

Furthermore, any discussions between multiple coders to account for differences in how individual items were coded would be threatened by the shared incentive of both coders to resolve disagreements in a way favorable towards the predicted outcome. Taken together, then, it may be best to not commit oneself to particular expectations about how to code data when one is not blind to their own hypotheses. However, it would be preferable to preregister details related to data collection and data analysis to ensure that clear and consistent standards are employed prior to analysis.[180]

Future studies could also attempt to offer more precise interpretation rates (and attempt to corroborate those reported here) by training coders blind to hypotheses, or recruiting adversarial

---

[180] In accordance with preregistration guidelines Hartman, Kern, and Mellor adapted from Kern and Gleditsch (2017) and made available at: https://osf.io/j7ghv/.

coders (e.g., people with opposing expectations about what the data should yield). In such cases, precise predictions would be more appropriate. There are limitations with both of these approaches, however. Coders blind to hypotheses may lack the competence to properly evaluate items, while adversarial coders would likely have sufficient competence, but could simply be biased in conflicting ways. Two or more biased coders won't necessarily result in unbiased estimates. Unfortunately, there may also be no theoretically neutral means of resolving such disagreements were they to arise. If, for instance, my coding of a dataset resulted in a clear intended interpretation rate of 20%, while a moral realist critical of my work arrived at 80%, there may be no way to definitively resolve such a dispute. Hopefully, discussion over our reasons and justifications for individual interpretations would allow for us to mutually converge on the most agreeable outcome, but it may be difficult to arrive at such a point in practice, and if we did arrive there, it would require extensive discussion, some of which may be intensely philosophical, and not merely a clash of data. Adequate assessment of folk metaethics may simply be unable to escape a dynamic interplay between philosophy and psychology.

**S4.4 Additional studies**

*S4.4.1 Study 1C: New test of the disagreement paradigm*

It is possible that idiosyncratic features of the sample population that participated in Goodwin and Darley's (2008) studies led to unusually low rates of intended interpretations of the source of disagreement. It is also possible that stimuli or other features of the original study prompted a higher rate of unintended interpretations. Study 1C was an attempt to assess whether clear intended interpretation rates would be higher if the disagreement paradigm were presented in an abstract, simplified form with minimal instructions in a different sample. However, one major difference between this study and the others is that I did not ask about the source of disagreement, but instead asked the participant why they chose the particular response that they did. Thus, the primary measure

was not whether they attributed the source of disagreement to a fundamental difference in moral values, but whether their response reflected a metaethical rationale for their response.

While I could have asked participants why they judged that the two people disagreed, so little contextual information was available that this would have been an unreasonable task. Participants were told that two hypothetical, unnamed strangers disagreed about an unspecified moral issue. While I *could* have asked why they thought people would typically disagree in such circumstances, such a response would not be especially indicative of any particular capacity for thinking in metaethical terms, nor would it be especially informative with respect to assessing the validity of the disagreement paradigm.

By employing a *third person* disagreement between two other people, rather than a *first person* disagreement between the participant and what was described (untruthfully) as another real, previous participant, this minimized the risk that participants would reasonably suspect that the other participant misunderstood the question; this was also minimized by using an abstract moral disagreement rather than a concrete one. In addition, by judging a disagreement between two other people, the participant was removed from the situation, which minimizes the risk that their normative concerns could bias their judgment or prompt the participant to interpret the question to concern their first-order moral judgments, thereby minimizing some of the primary inadequacies with traditional first-person, concrete versions of the disagreement paradigm.

**Methods**

***Participants.*** 106 participants participated in the survey, but six participants did not complete the primary measure and were thus excluded from analysis. As a result, participants consisted of 100 adult US residents on Amazon's Mechanical Turk.

***Procedure.*** To assess interpretation rates, I asked participants to answer a question about a moral disagreement. This was a moral disagreement between two unnamed people about an unspecified moral issue, and therefore represents an *abstract* moral disagreement rather than a *concrete* moral issue. The degree to which a moral disagreement concerns an abstract from concrete moral considerations is of course a matter of degree, but I operationalize *abstract* moral considerations as those concerning the moral domain as a whole, while concrete *moral* considerations specify a particular moral issue, such as murder, theft, or abortion. This study employed an abstract moral disagreement, whereas Goodwin and Darley (2008) made exclusive use of concrete moral disagreements.

However, the response options were similar to Goodwin and Darley. Participants were given the choice to judge whether, if two people disagree about a moral issue, that *both* can be correct, or that *neither* can be correct. The judgment that both can be correct could be interpreted as an expression of *relativism* or at least *non-objectivism*, while judging that at least one of them must be incorrect could be interpreted as *realism*. Second, participants were asked to explain why they chose to respond as they did with an open response question.

***Measures.*** For all measures, I asked participants to judge whether two people who disagreed about a moral issue could both be correct, or whether at least one of them must be incorrect, and then asked them to explain why they chose this response. All participants were presented with the following multiple choice question:

*When two people disagree about a moral issue, do you think they can both be correct, or must at least one of them be incorrect?*
- *They can both be correct*
- *At least one of them must be incorrect*

Participants were then presented with a text box and asked to:

*Please briefly explain why you chose this response.*

No further instructions were provided.

**Results**

As expected, most participants did not provide a clear intended metaethical rationale for their response.[181] The proportion of participants coded as *clear intended* was 23% ($n = 23$), which was significantly less than 0.5, $\chi2(1, N = 100) = 29.16$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 30.58%. These results indicate that most participants did not offer a clear metaethical rationale for why they chose the particular response that they did. However, 40% of participants were coded as *clear unintended* ($n = 40$), indicating that a substantial proportion of participants offered an explanation that was clearly *unrelated* to the kind of metaethical rationale that would indicate that they responded based on their metaethical stances or commitments. The remaining 37% of responses were *unclear* ($n = 37$). Results may be seen in

**Figure S4.2**.

---

[181] Although it was not an important part of the analysis, 85% of participants chose the *relativist* response, while 15% chose the realist response. The proportion of participants coded as clearly interpreted was significantly different from 0.5, $\chi2(1, N = 100) = 49$, $p < 0.001$, 95% CI [0.77, 0.91], indicating that significantly more people chose the *relativist* response over the *realist* response.

**Figure S4.2**.

*Interpretation rates for Study 1C*
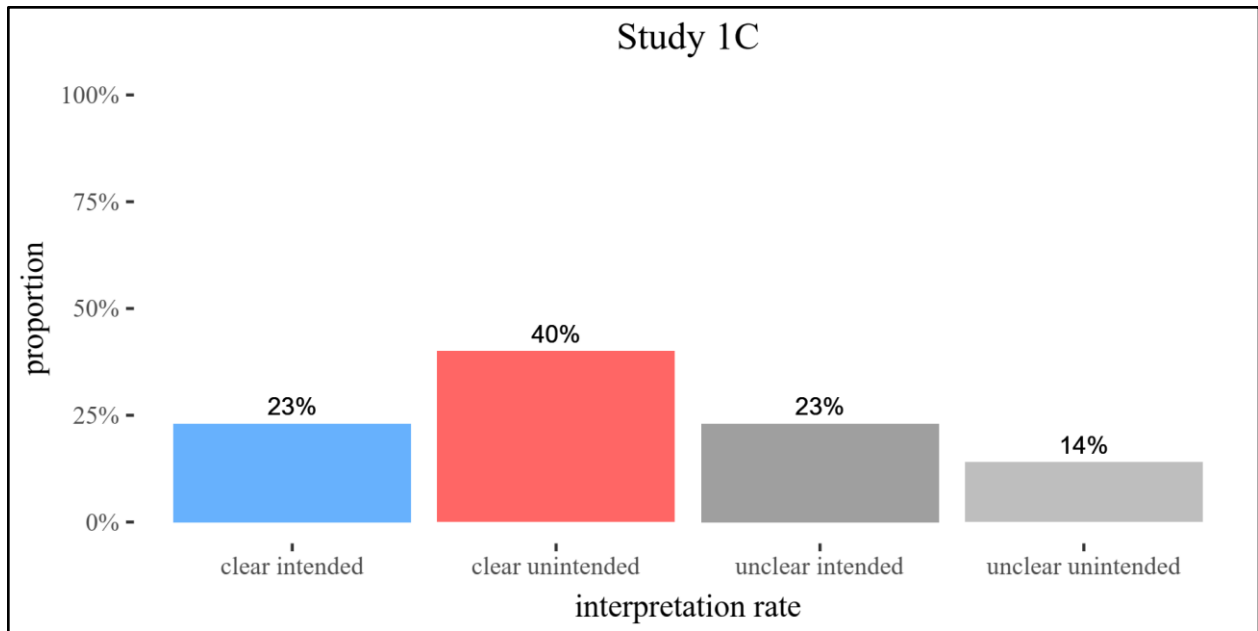


**Discussion**

Overall, these findings support the conclusion that people do not interpret questions about moral disagreements as intended. This is consistent with both the conclusion that the disagreement paradigm is not a valid measure of folk metaethical belief, and with metaethical indeterminacy. It may be that people rarely offer a metaethical interpretation for questions about metaethics because they don't have metaethical stances or commitments, and that, as a result of not thinking in metaethical terms, they tend to interpret questions intended to represent metaethics in some other way.

**S4.4.2 Study 6: The New Metaethics Questionnaire (NMQ)**

Yilmaz and Bahçekapili (2015a; 2015b; 2018; 2020) introduced a handful of scale items to assess folk metaethical beliefs. In at least one instance they employed a set of three items (YB3), but other versions present eight items in total, consisting of two four-item subscales: a "subjectivism" subscale, and an "objectivism" subscale, which they dub the New Meta-ethics Questionnaire (NMQ). As argued in Bush and Moss (2020), these items suffer extremely poor face validity. Nevertheless, Yilmaz and

Bahçekapili (2020) have continued to use versions of these items and claim that they exhibit various positive indications of validity:

Moral subjectivism scale [...] comprising four items on a 7-point Likert-type scale (e.g. "Since moral rules are not right or wrong in an absolute sense, moral arguments are always destined to remain futile"; Cronbach's $\alpha_{\text{Turkish}} = .78$; Cronbach's $\alpha_{\text{American}} = .75$) was used to measure moral subjectivism. This scale was also used in further research and showed good predictive and convergent validities [...]. For example, higher endorsement of subjective morality predicts left-wing political orientation and lower endorsement of objective morality and belief that morality is founded on divine authority [...]. (pp. 235-236)

This is unfortunate, given that the very year this was published, I argued in Bush and Moss (2020) that these measures are *not* valid. Unfortunately, these items suffer such poor face validity that there is little chance they offer appropriate measures of folk metaethical stances or commitments. Take the very example they use:

*Since moral rules are not right or wrong in an absolute sense, moral arguments are always destined to remain futile*

This is supposed to be a measure of subjectivism, yet it does not clearly convey the meaning of subjectivism. Subjectivism is the view that moral claims are true or false relative to the standards of individuals. Why should we expect ordinary people to interpret the example item they give to convey this claim? Firstly, it says that "since moral rules are not right or wrong *in an absolute sense*." This would appear to more accurately reflect a rejection of absolutism about moral rules; that is, a rejection of the claim that there are moral rules that do not admit of exceptions, e.g., moral rules such as "abortion is *always* wrong," rather than a more flexible rule that says abortion is permissible in some cases but not others. In other words, part of the statement seems to express *situationism* or the rejection of *absolutism*; what it does *not* do is express anything about moral truth being subjective.

The second part of the statement states that "moral arguments are always destined to remain futile." This is unclear. What does it mean to say moral arguments are "destined to remain futile"? To say that something is futile typically means something like "pointless." Yet even if moral arguments were pointless, why would agreeing with that indicate subjectivism? A moral realist could think moral arguments are futile for a variety of reasons that have nothing to do with morality being subjective. One straightforward reading of this, for instance, is that such arguments are pointless because it's difficult to know what the correct answer is. That is, this remark could be readily interpreted as an *epistemic* statement about our ability to resolve moral disputes. Yet it could turn on social or practical facts, such as human stubbornness, e.g., one might think that although there are stance-independent moral facts, some people are unwilling to change their minds. Simply put, one's stance towards the futility of moral arguments has no obvious connection to moral subjectivism. More generally, both realists and antirealists could believe moral disputes are futile or not futile; such a judgment is simply orthogonal to the dispute between realism and antirealism.

Finally, note that this is an especially complex statement, because it doesn't simply assert that "*moral rules are not right or wrong in an absolute sense*" nor that "*moral arguments are always destined to remain futile*" nor even the conjunction of these two statements. Rather, it asserts that *because* the former statement is true, that *therefore* the latter statement follows. That is, it isn't merely a double-barreled question, which would be problematic enough, i.e., a statement that asserts "X and Y." Such questions are not appropriate, because there is no way to agree with X but not Y, or Y but not X. Instead, this item asserts not only that X and Y are true, but that Y is true *because* X is true. That is, it asserts something that more closely approximates: X, X→Y, and Y. This is an incredibly complex item, and it may be too cognitively demanding to expert participants to interpret it as intended. For instance, people may not interpret it as making the claim that Y is true because X is true, and may instead express agreement in some crude way that amounts to averaging their agreement with X and Y

individually. Worst of all, however, it's unclear why agreeing with the item as a whole would entail subjectivism. Suppose you don't believe moral rules are true or false "in an absolute sense" (whatever that means), and that *because of this* moral arguments will *always* be futile. Even if you interpreted the first part to entail subjectivism, and you agree with that, you may not agree that moral arguments "will always be futile." It is not part of moral subjectivism that moral arguments are futile!

There is no charitable way to put it: this item is not a valid measure of subjectivism. I doubt it's a valid measure of anything. It seems to represent little more than a convoluted parroting of the kinds of terms and phrases moral philosophers use, without any apparent appreciation for what those terms mean and how they relate to one another. And that's the one they chose as an example! Yilmaz and Bahçekapili also claim that their scale exhibits high predictive and convergent validity. While it may provide *some* evidence for convergent validity that their scale correlates with other paradigms used to measure folk metaethics, if those scales are also invalid, they could both pick up on the same patterns in how people respond without those patterns necessarily reflecting genuine folk metaethical stances and commitments. That is, if people systematically interpret questions about metaethics in unintended ways that are similar across studies, which is precisely what I am claiming and is precisely what the data I report here suggests, one could observe similar patterns of results across different paradigms. For example, if many people conflate statements intended to reflect *relativism* or *subjectivism* with *situationism* or *descriptive relativism*, and if many people conflate *realism* with *absolutism*, one could observe similar patterns in participant response using a variety of paradigms. Such paradigms would correlate with each other, *even though they are all invalid*. It is *not enough* to show that two or more measures correlate with one another; you *must* provide direct evidence that they are measuring the construct of interest.

Predictive validity is also inadequate to establish the validity of their measures. They find that higher subjectivism scores predict left-wing political orientation, lower objectivism (i.e., *realism*), and

less endorsement that morality is based on divine authority. Yet this is completely consistent with the patterns of unintended interpretations indicated by my analysis of open response questions. If someone interprets subjectivism/relativism as situationism and realism as absolutism, their responses to both types of questions will tend to correlate because situationism and absolutism conflict with one another, *not* their views on relativism and realism. Indeed, since realism does not in any way entail intolerance, conservative values, or insensitivity to context, and antirealism doesn't entail the contrary, the correlation between "metaethics" scores and these measures is equally good if not better evidence that their measures aren't valid.

More generally, there may be a variety of reasons why someone who is on the political left may tend to favor subjectivist/relativist responses by crudely associating such remarks with a more tolerant and inclusive attitude towards people with different beliefs and backgrounds, and items reflecting realism as expressing a rigid, "black and white," dogmatic, and intolerant attitude. That is, items ostensibly intended to purely convey metaethical stances may be associated with the *normative* content of left and right political ideology. If so, the tendency for responses to subjectivism/relativism and realism/objectivism items to correlate with one another, and for the former to be associated with left-wing political ideology would be better explained by these items simply reflecting the non-metaethical values of their respective political ideologies (Collier-Spruel et al., 2019; cf. Goodwin & Darley, 2008).[182]

In light of these considerations, there is little reason to believe Y&B's scale items are a valid measure of folk metaethics. However, we should not be content with armchair observations about the poor face validity of these items, nor should we be so confident that we can explain away the predictive

---

[182] Collier-Spruel et al.'s (2019) MRS demonstrated a significant association with a variety of measures associated with political ideology, including a negative correlation with right-wing authoritarianism, conservative political orientation, and the three moral foundations typically associated with American conservatism: loyalty, respect, and purity (Graham, Haidt, & Nosek, 2009).

and convergent validity of their items on the hypothesis that people do not interpret these scale items as intended. Instead, I once again gathered open response data in order to assess interpretation rates and identify recurring themes. This consisted of three sets of data. Two analyses assessed interpretation rates for the YB3. The third and fourth studies assess interpretation rates for the NMQ, which consists of a 4-item subjectivism subscale and the 4-item objectivism subscale. Like Study 5, I asked participants both to explain *why* (studies 6A and 6C) they answered the way that they did, and to *explain* what the items mean (studies 6B and 6D).

*S4.4.2.1 Study 6A: YB3 - Why*

**Methods**

**Participants.** Participants were drawn from a larger sample of 2010 participants who each responded to a variety of questions about metaethics. Only those participants who were assigned to an item from the 3-item Yilmaz (YB3; Yilmaz & Bahçekapili, 2015b) scale or one of the eight items from the NMQ will be analyzed here. These conditions accounted for 142 in the YB3 conditions and 281 participants in the NMQ conditions, resulting in a total of 423 participants across all conditions. No demographic data was collected.

**Procedure.** All participants were asked to rate how much they agree or disagree with a given statement, to explain why they chose the level of agreement that they did, and were asked to explain in their own words what they believe the item they were given means. Order did not vary. The study had a between-subjects design with all participants assigned to respond to these three questions for one item. Wording was follows for the three questions:

    (1) Please rate how much you agree or disagree with the following statement.

**[statement]**

*[1 = Strongly disagree, 7 = Strongly agree]*

    (2) Please briefly explain why you chose this response.

(3) In your own words, please briefly explain what this statement means:

**[statement]**

**Measures.** Level of agreement was measured via a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree). The request for participants to explain why they expressed the level of agreement that they did and the request to explain what the item means were both presented as open response questions. For both responses, interpretation rates and themes were evaluated using the method outlined in the introduction.

For *why* conditions, any response that clearly conveyed a metaethical position (of any kind) was coded as a clear intended interpretation. For *explain* conditions, any response that matched the metaethical position that the item was intended to convey was coded as a clear intended interpretation. For the YB3, all three items were intended to express *relativism*. For the NMQ, items #1-#4 reflected *relativism*, while items #5-8 reflected *realism*. This is a paradigmatic instance of a clear intended interpretation for one of a *why* question (YB3 item #3, response 107):

*Ethics, to me, is a standard of behavior that assures honest, fair and unbiased actions in work and/or personal activities and is not up to an individual. Neither are moral standards up to an individual, if they were murder would be moral to the person commiting the crime.*

This item was coded as a clear intended interpretation because the participant conveyed that whether an action is morally right or wrong is "not up to an individual." This suggests that this participant believes that moral standards are stance-independent. With respect to the *explain* conditions, the following response reflects a paradigmatic instance of a clear intended interpretation to an item reflecting relativism (NMQ, item #2, response 67):

*Morality is subjective thus debates can't be settled concrete as they are a matter of preference.*

This is a clear intended interpretation both because the participant explicitly describes morality as subjective (note that the original item did not include the term "subjective" so they were not merely repeating terms provided in the stimuli), and because they make clear that the reason moral disputes cannot be resolved (a notion conveyed in the item) is that because morality is a *matter of preference*. This is a surprisingly succinct expression of subjectivism. While further discussion with this participant may reveal confusion, inconsistency, or uncertainty, for so short a response this is about as good as one could reasonably expect a response to be.

**Results**

As expected, most participants did not clearly interpret the items as intended. Across all items, the clear intended interpretation rate was 16.9% ($n = 24$). The clear intended interpretation rate for item #1 was 2.2% ($n = 1$), item #2 20.0% ($n = 9$), and item #3 27.5% ($n = 14$). Aggregating across all three items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi 2(1, N = 142) = 62.23$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 22.68%. The clear unintended interpretation across all items was 24.6% ($n = 35$), and for item #1 was 21.7% ($n = 10$), item #2 22.2% ($n = 10$), and item #3 29.4% ($n = 15$). Full details of interpretation rates are featured in **Figure S4.3.**

**Figure S4.3**

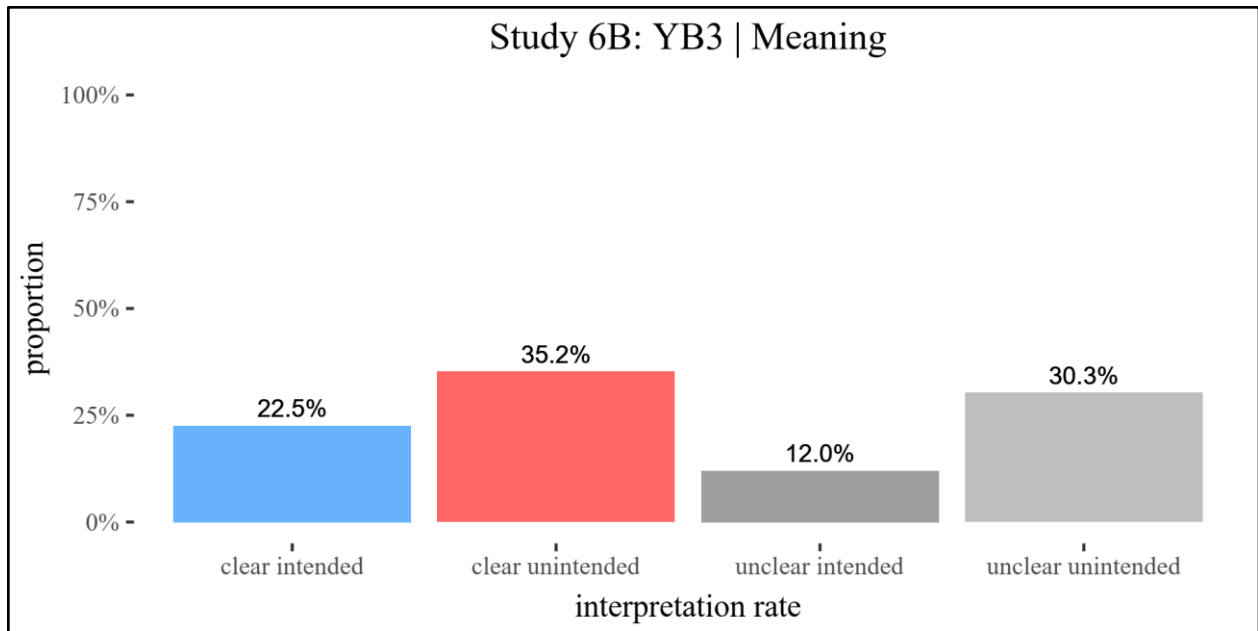*Interpretation rates for Study 6A*
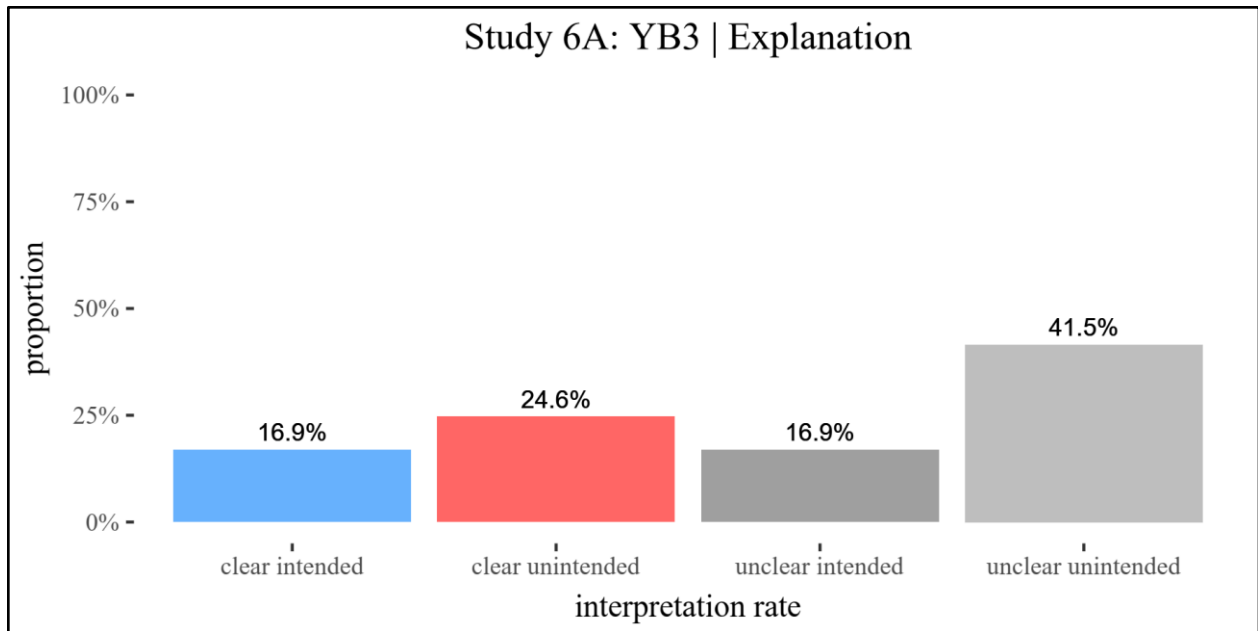


*S4.4.2.2 Study 6B: YB3 - Explain*

As expected, when aggregated across conditions most participants did not provide a clear intended interpretation. The total clear intended interpretation rate was 22.5% ($n = 32$). The clear intended interpretation rate for item #1 was 26.1% ($n = 12$), 20.0% for item #2 ($n = 9$), and 21.6% for item #3 ($n = 11$). Aggregating across all three items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 142) = 42.85, p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 28.78%. The total clear unintended interpretation rate was high at 35.2% ($n = 35.2\%$). The clear unintended rate for item #1 was 34.8% ($n = 16$), for item #2 it was 24.4% ($n = 11$), and for item #3 it was 45.1% ($n = 23$). Full details of interpretation rates are in **Figure S4.4.**

**Figure S4.4**

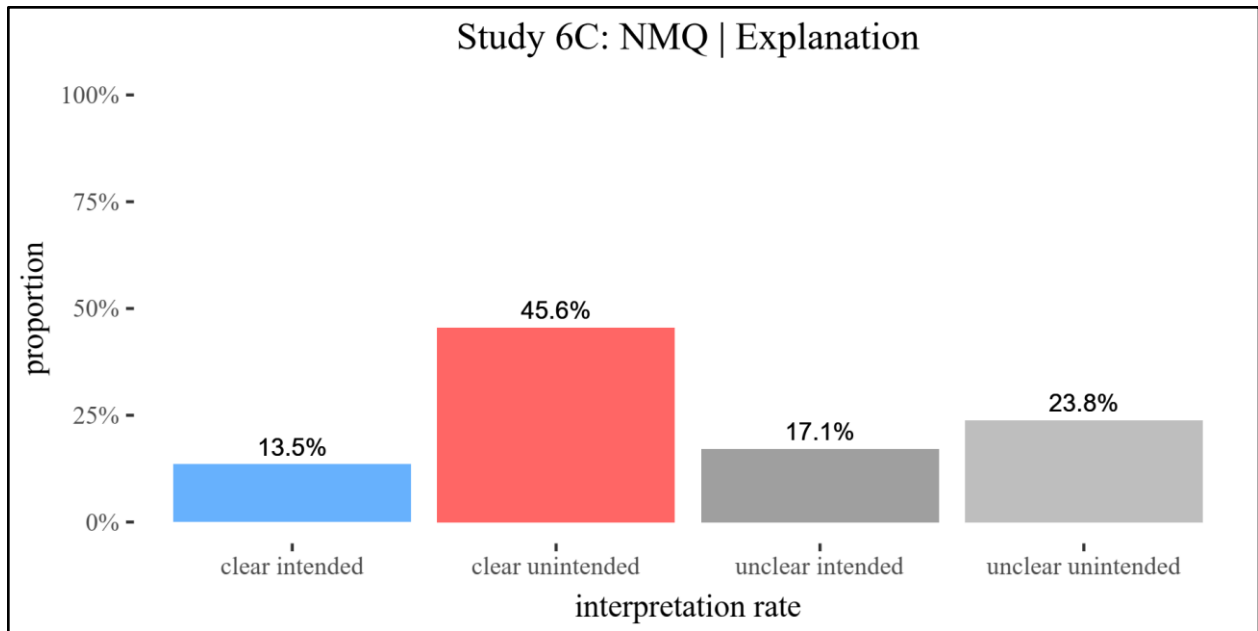*Interpretation rates for study 6B*



*S4.4.2.3 Study 6C: NMQ - Why*

As expected, less than half of participants provided a clear intended interpretation. Across all items, the total clear intended interpretation rate was 13.5% ($n = 38$). The clear intended interpretation rate for item #1 was 34.1% ($n = 15$), item #2 was 22.0% ($n = 11$), item #3 was 9.3% ($n = 4$), item #4 was 0.0% ($n = 0$), item #5 was 6.4% ($n = 3$), and item #6 was 10.6% ($n = 5$). Aggregating across all three items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 281) = 149.56, p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 17.23%. The total clear unintended interpretation rate was very high at 45.6% ($n = 128$). The clear unintended interpretation rate for item #1 was 15.9% ($n = 8$), item #2 36.0% ($n = 18$), item #3 46.5% ($n = 20$), item #4 80.0% ($n = 40$), item #5 36.2% ($n = 17$), and item #6 53.2% ($n = 25$). All interpretation rates are in **Figure S4.5**.

**Figure 4.5**

*Interpretation rates for Study 6C*



*S4.4.2.4 Study 6D: NMQ - Explain*

As expected, less than half of participants provided a clear intended interpretation. Across all items, the total clear intended interpretation rate was 11.4% ($n = 32$). The clear intended interpretation rate for item #1 was 22.7% ($n = 10$), item #2 was 20.0% ($n = 10$), item #3 was 2.3% ($n = 1$), item #4 was 0.0% ($n = 0$), item #5 was 12.8 ($n = 6$), and item #6 was 10.6% ($n = 5$). Aggregating across all three items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 281) = 167.58$, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 14.88%. The total clear unintended interpretation rate was very high at 67.6% ($n = 190$). The clear unintended interpretation rate for item #1 was 56.8%% ($n = 8$), item #2 42.0% ($n = 18$), item #3 69.8% ($n = 20$), item #4 90.0% ($n = 40$), item #5 68.1% ($n = 17$), and item #6 78.7% ($n = 25$). All interpretation rates are on **Figure S4.6**.

**Figure S4.6**

*Interpretation rates for Study 6D*
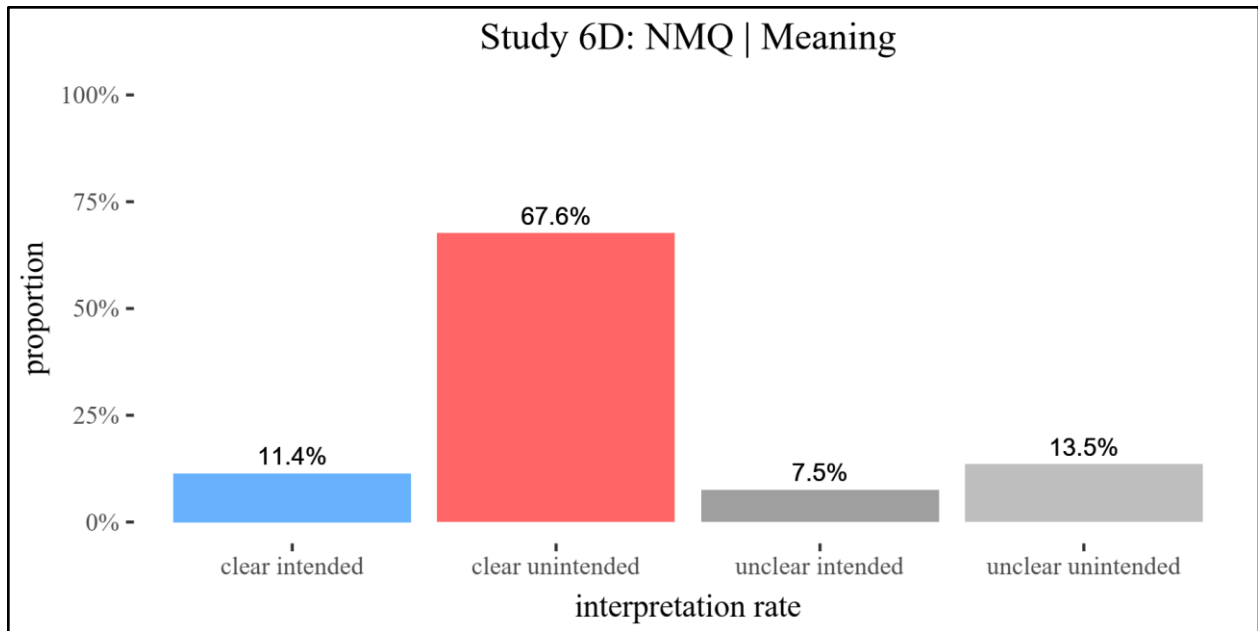


**Discussion**

Overall, few participants demonstrated that they interpreted questions on the YB3 or NMQ as intended. Conversely, a considerable portion of participants offered clear unintended interpretations. Thematic analysis was consistent with the ways participants are expected to conflate items reflecting relativism with other considerations, with the *descriptive*, *normative*, and *universal* themes making frequent appearances, and the remaining themes indicative of participants struggling to interpret questions as intended in ways consistent with other studies evaluating open response questions. Of course, these studies suffer the same limitations as other open response questions: the high rate of unclear responses makes it hard to provide a precise estimate of the number of people who interpreted questions as intended, and it is possible that participants interpreted questions as intended even if their responses do not reflect this fact. Like other studies, the number of participants per item was not especially large, so the exact proportions of clear intended, unintended, and unclear responses is likely to be a noisy and imprecise indication of the truth per-item interpretation rates.

Nevertheless, these findings support my central hypotheses. First, the interpretation rates reported here are consistent with the more general claim that studies on folk metaethics are invalid because many participants do not interpret stimuli as intended. Second, these findings provide further support for metaethical indeterminacy. Every instance and iteration of ordinary people struggling to understand and articulate metaethical notions hints at the possibility that they struggle because they don't possess such concepts. While it remains possible that people have an implicit competence with metaethical concepts that they struggle to articulate, the onus is on those who believe this is the case to provide concrete evidence for such claims. At present, I am aware of little evidence, much less decisive evidence, that metaethical commitments are implicit in the way ordinary people speak and think.

### S4.4.3 Study 7: Folk Moral Objectivism scale (FMO)

Zijlstra (2019) has recently developed the Folk Moral Objectivism (FMO) scale. The scale's name is deceptively understated: Zijlstra conceives of folk metaethics as a multidimensional cluster of constructs, and devised the FMO in an ambitious effort to capture five subdimensions: *no truth*, *relativism*, *universalism*, *absolutism*, and *divine command theory (DCT)*, each of which appears as a 4-item subscale of the FMO. There are serious problems with the face validity of these items, and several of the subscales are not directly related to realism and antirealism. Universalism and absolutism concern the scope and generality of normative moral rules, respectively. It is puzzling they were included in a folk moral *objectivism* scale, given that neither is directly related to objectivism: one could be a realist or an antirealist and endorse or reject universalism or absolutism. These distinctions are orthogonal to realism, so they are not actually subdimensions of folk moral *objectivism* but just loosely associated notions.

Divine command theory is also hard to categorize. Whether moral facts are facts about God's commands *is* a metaethical claim. However, it is difficult to neatly fit DCT within the realist vs.

antirealist dichotomy, since DCT could be construed in both stance-dependent and stance-independent terms.[183] As such, while there is considerable value in assessing whether ordinary people think moral facts depend on God, or consist in God's commands, it is unclear whether such views would reliably indicate whether someone who does ground morality in God endorses moral realism or not. As such, I opted not to report coding for these items for three reasons. First, DCT is not directly related to the central objective of assessing interpretation rates for items related to realism and antirealism. Second, while DCT items do fall within the scope of metaethics, these items were very different from other items, so coding them would not only not only be a digression, but a laborious one. Third, and finally, I critique the validity of these items in chapter **Supplement 3**. As I argue there, three of these items aren't face valid, while the fourth may fail to be as well. Yet the most serious problem with DCT measures is that even if we took agreement with DCT items to reflect a commitment to moral realism, we cannot take disagreement to reflect antirealism, because someone could endorse moral realism for reasons other than via DCT. In other words, even if DCT represented a distinctive form of moral realism, rejecting it would simply mean that you disagree with a specific form of moral realism, not that you reject moral realism. As such, even if these items were valid measures of DCT, they couldn't serve as valid measures of realism or antirealism because level of agreement with these items cannot be interpreted as a reliable indicator of realism or antirealism.

I decided instead to focus exclusively on the four items on the *relativism* subscale, since this was the only dimension that represented a substantive metaethical position consistent with the

---

[183] One could conceive of DCT as the claim that moral facts depend on God's stance, and that since moral facts are stance-dependent, it is a form of antirealism, albeit a nonrelative one. This is what Joyce (2015) calls a relation-designating account. While moral facts are stance-dependent, a proper moral claim would not contain an implicit indexical element. Since all moral claims refer to God's commands, the truth value of moral claims could not vary from speaker to speaker. Such accounts are merely a form of nonrelativistic antirealism, along with ideal observer theories and any other metaethical positions which hold that moral facts are stance-dependent, but that there is only one stance on which they depend. On the other hand, one could by some other means maintain that the moral facts with which God furnishes us are, for whatever reason, not reducible to or mere expressions of God's subjective standards, or otherwise merely facts about God's stance, but are in fact reflections of some stance-independent moral truths that (again, for whatever reason) require God.

purpose of assessing folk metaethical stances and commitments.[184] My goal in this study was to focus strictly on assessing interpretation rates for the four-item relativism subscale of the FMO. While I have collected data on absolutism, universalism, and DCT items, assessing interpretation rates for these items will be reserved for future projects. Like studies 5 and 6, participants were presented with the original question used on these scales and asked to express their level of agreement, were then asked why they answered in this way (the *why* condition), and finally were asked to explain what they thought the item meant in their own words (the *explain* condition). Like previous studies, I expected fewer than half of participants to provide a clear intended interpretation for both the *explain* and *why* conditions.

---

[184] *No truth* is also a viable candidate for items that could reflect folk metaethical stances or commitments, but it's not clear what would constitute an intended interpretation, so I opted not to assess these. This is because "no truth" doesn't represent a clear metaethical position. While noncognitivists and error theorists may believe there are no moral truths, none of the items unambiguously represent any particular metaethical position, so it is unclear how to interpret what it would mean to agree or disagree with them. Take, for instance, this item:

*What people believe to be morally right and wrong are merely social conventions that could have been different*

It's not obvious why agreeing with this would entail that you don't believe in moral truth. After all, you could believe that most people's moral beliefs are merely social conventions, and that there are moral truths. This looks more like a descriptive claim than a substantive metaethical one.

Likewise, suppose you disagree with this item. Does that entail that you do believe there are moral truths? Not necessarily. Simply because you don't think moral beliefs are merely social conventions that could have been different doesn't mean you're a moral realist. In fact, the prominent error theorist Joyce (2006) argues for a fairly strong form of moral nativism, and would probably *not* agree that moral beliefs are merely social conventions that could have been different. Thus, both agreement and disagreement with this item are consistent with both realism and antirealism, and not merely consistent in principle, but actually reflect the positions defended by central figures in contemporary metaethics. It's simply not the case that disagreeing with this item entails you're a moral realist, unless the *only* way to be a moral antirealist was to believe moral beliefs are "merely social conventions that could have been different." This is false, so this isn't a valid item for measuring realism or antirealism.

In short, it was not clear how to code these items, since it is unclear, if these items represented a particular psychological construct, what metaethical position that construct would correspond to, or what it would mean to agree or disagree with these items, or why a reason given for one's agreement or disagreement would represent an "intended" interpretation. Simply put: I don't know what the intended interpretation of these items is, so there's no way to code them with that purpose in mind. You can't reliably judge if someone is hitting a target if you don't know what the target is.

**Methods**

**Participants.** Participants were drawn from a larger study that consisted of 2010 participants. Only the results of participants assigned to one of the four items on the relativism subscale of the FMO were analyzed here, resulting in a total of 217 participants. No demographic data was collected.

**Procedure.** The procedure was identical to studies 5 and 6. All participants were randomly assigned to receive one item from the relativism subscale of the FMO or an item from other subscales or scales. Each participant was then asked to express their level of agreement on a seven-point Likert scale (1 = Strongly disagree, 7 = strongly agree). They were then asked to "Please briefly explain why you chose this response." Finally, they were asked: "In your own words, please briefly explain what this statement means: **[statement.]**" These two questions were open response questions. For both questions participants were presented with a text box where participants could write a response. I used a somewhat small text box for both questions, and emphasized that their responses should be *brief*, both of which were intended to encourage short responses.

**Measures.** All conditions shared the same five measures: level of agreement, coded interpretation rates for each of the two open response questions in according with the coding scheme outlined in the introduction (i.e., 1|1 clear intended, 0|1 clear unintended, 1|0 unclear intended, 0|0 unclear unintended), and thematic analysis for each of the two open response questions.
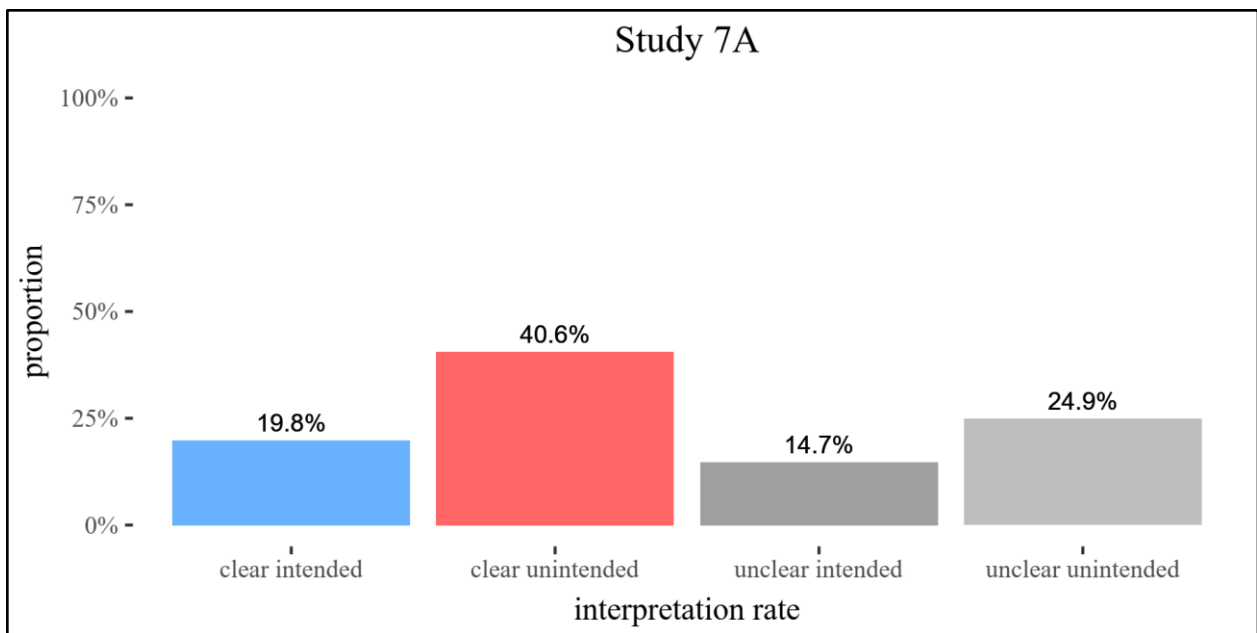
**Results**

*S4.4.3.1 Study 7A: Why*

As expected, most participants did not provide a clear interpretation for items on the relativism subscale of the FMO. Across all items, the clear intended interpretation rate was 19.8% ($n = 43$). The clear intended interpretation rate for item #1 was 26.9% ($n = 14$), for item #2 it was 17.0% ($n = 9$), for item #3 it was 18.2% ($n = 10$), and for item #4 it was 17.5% ($n = 10$). Aggregating across all items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N =$

217) = 79.08, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 24.63%. The overall clear unintended interpretation rate was 40.6% ($n = 88$). The clear unintended interpretation rate for item #1 was 21.2% ($n = 11$), while it was 35.8% ($n = 19$) for item #2, 56.5% ($n = 31$) for item #3, and 47.4% ($n = 27$) for item #4. All results are featured in **Figure S4.7.**

**Figure S4.7**

*Interpretation rates for Study 7A*



*S4.4.3.2 Study 7B: Explain condition*

As expected, most participants did not provide a clear interpretation for items on the relativism subscale of the FMO. Across all items, the clear intended interpretation rate was 12.4% ($n = 27$). The clear intended interpretation rate for item #1 was 7.7% ($n = 4$), for item #2 it was 11.3% ($n = 6$), for item #3 it was 10.9% ($n = 6$), and for item #4 it was 19.3% ($n = 11$). Aggregating across all items, the proportion of participants coded as clearly interpreted was significantly less than 0.5, $\chi2(1, N = 217)$ = 122.44, $p < 0.001$. With 95% confidence, the percentage of responses coded as clearly interpreted was less than 16.60%. The overall clear unintended interpretation rate was 48.4% ($n = 105$). The clear

unintended interpretation rate for item #1 was 17.3% ($n = 9$), while it was 41.5% ($n = 22$) for item #2, 76.4% ($n = 42$) for item #3, and 56.1% ($n = 32$) for item #4. All results are featured in **Figure S4.8**.

**Figure S4.8**

*Interpretation rates for Study 7B*
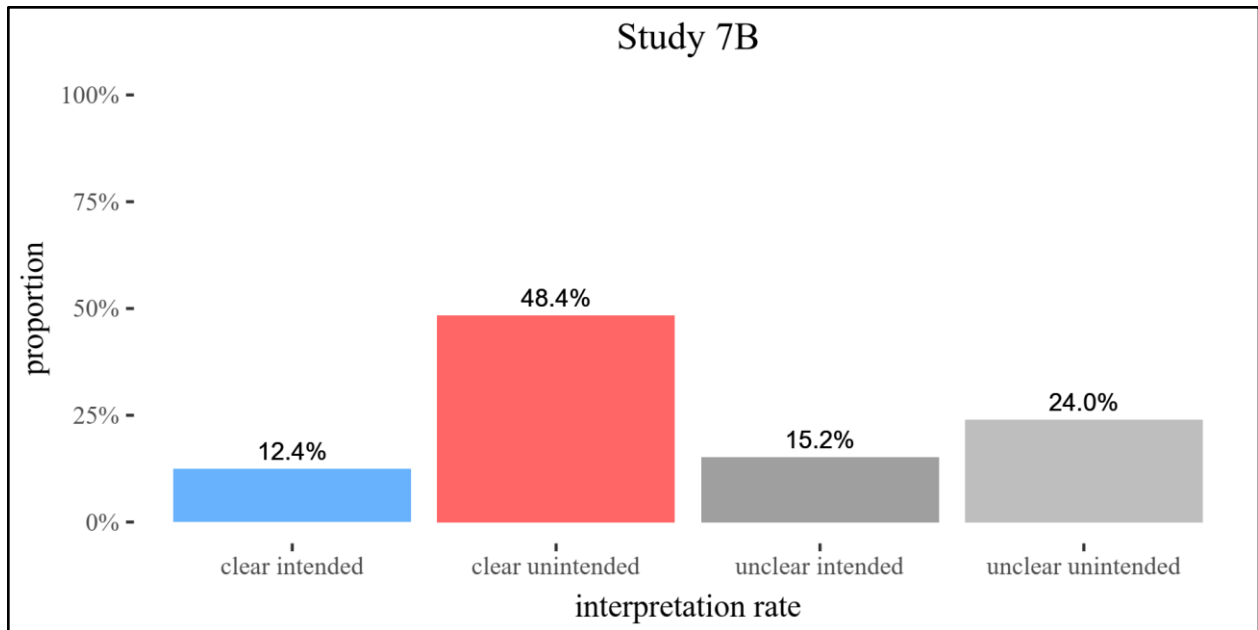


**Discussion**

Most people failed to provide clear intended interpretation for every item in both conditions, while a much higher proportion of participants provided clear unintended responses. Once again, the results of analyzing open response questions intended to assess how people interpret questions about metaethics reveals that people struggle to interpret questions as intended. In this case, the relativism subscale of the FMO reveals that people consistently fail to interpret questions about relativism as intended, and typically do not even interpret them in metaethical terms at all.

In the *why* condition, participants consistently explained why they answered as they did by appealing to a variety of considerations irrelevant to relativism, and that hint at unintended interpretations, such as the fact that different people have different moral beliefs, claims about how

people acquire their moral beliefs, the notion that whether an action is right or wrong depends on the context in which it occurs, or rejecting the notion that morality is "black and white." These claims seem to conflate metaethical considerations with descriptive and normative considerations that have no direct relation to relativism. A moral realist can affirm or deny descriptive claims about the etiology and diversity of moral values, and they could affirm or deny the degree to which moral claims are sensitive to contextual considerations. Nothing about such claims provides any direct basis for accepting or rejecting relativism.

The *explain* condition reveals a similar pattern of seemingly unintended interpretations. When asked to explain what the relativism items on the FMO mean, participants frequently described items as descriptive claims about the diversity or origins of moral beliefs, or the observation that people regard their own positions as correct. Notably, a handful of participants even interpreted these items as reflecting the notion that there are no moral truths (i.e., the *nihilism* theme). In an interesting twist, these participants appear to have interpreted the items in metaethical terms, but understood the relevant metaethical position in ways inconsistent with their academic counterparts: relativism *does not* deny that there are moral truths, it merely holds that such truths are relative. This may seem like an unimportant distinction, but an inaccuracy is still an inaccuracy. Notably, the FMO does have a four item "no truth" subscale as well, which reveals that Zijlstra himself sees "no truth" and "relativism" as conceptually distinct. If participants cannot draw a distinction between there being no moral truth, and moral truth being relative, this is yet another challenge to the notion that people have determinate metaethical standards. For comparison, if someone could not distinguish purple from blue, this would raise doubts that this person considered either blue or purple, in particular, to be their favorite color. Just the same, it's questionable whether we can attribute a determinate metaethical stance to people who cannot distinguish different metaethical stances from one another even when they're explicitly presented alongside one another via ostensibly face valid items.

Of course, the alternative to metaethical indeterminacy is simply that the relativism subscale of the FMO failed to present metaethical distinctions adequately. This is a limitation that plagues *all* of the studies reported here, and represents one of the most significant shortcomings with open response analysis as a tool for assessing validity. Nevertheless, even if such problems do limit the degree to which findings support metaethical indeterminacy, they present researchers with an awkward tension: the more confident researchers are that a given measure is valid, the more difficult it becomes to explain away widespread evidence that people reliably fail to interpret stimuli as intended. If an item really is valid, and most people don't interpret as intended anyway, indeterminacy becomes an increasingly attractive explanation for why there would be such poor rates of intended interpretation.

### S4.5 Limitations

#### S4.5.1 General limitations

There are a variety of limitations that apply to the method employed in **Chapter 4**. It is possible, for instance, that participants have implicit metaethical commitments and that such stances drove their response to the question about moral disagreement, but that they lack introspective access to these stances and were thus unable to accurately report this when asked why they answered as they did. It is even possible people do have an explicit metaethical stance but lack the terms or concepts to adequately express their metaethical position. Finally, it is possible that people could have responded to the original question based on their metaethical stance or commitment, but interpreted the question about why they chose this response in a way that prompted them to offer some explanation *other than* the metaethical rationale for their choice. This is quite plausible in other contexts, and could readily explain the low rate of intended interpretations. For instance, suppose I surveyed people about why they went to the movies, and my goal was to determine whether they went to the particular movie because they *thought the movie would be good*. If I asked people leaving the theater "Why did you go to see this movie?", I might get responses like the following:

*"Because it's Friday night."*

*"Because today is my day off of work."*

*"Because this is when my friends wanted to go."*

These are perfectly sensible responses to the question of why they went, but they are all responses that address why they went to movies *now* rather than some other time. It would be absurd to conclude that if many people responded this way that therefore people did not go to the movies because they wanted to see the movie in question. If asked "Did you go to the movie because you wanted to see the movie?" you may get a puzzled "...yes…" from most people, and a few people who say "no, my boyfriend insisted we see it but I really wanted to see a different movie…." This might give us a more accurate account of the proportion who went to the movie in question because they wanted to see it, but there would be shortcomings and limitations to this approach, too.

For instance, suppose someone did go to the movie because their boyfriend dragged them to it. One person might respond by saying that "No, I didn't want to see this movie, I wanted to see that other movie. But I went because I wanted to have a good time with my boyfriend and figured it'd be easier to see this movie now, and the other one next week." Another person might respond differently, by saying, "Yes, I did want to see this movie. While I wanted to see that other movie more, I figured it'd be easier to see this movie now and the other one next week." If these participants had a conversation with one another might agree that they both felt *exactly the same way*, and yet they gave categorically different responses. Why? Because the first participant's interpretation of the question more closely approximated, "Was this the movie you wanted to see most of all?" while the second participant's response more closely reflected, "Was this trip to the movie theater something you wanted to do?" *Both* participants would respond to the first with "No" and the second with "Yes." Background assumptions about what is being asked can influence how people interpret seemingly straightforward stimuli, resulting in interpretative variation even for very simple questions. And direct questions can

suffer from other difficulties. If you were asking someone who didn't want to see a movie if they wanted to see it in front of friends or family or on a date, they might say "Yes" anyway.

While questions about whether you wanted to go to the movies may be simple and straightforward, questions about metaethics are far from simple or straightforward. Direct questions may be inappropriate or even more difficult to interpret than indirect ones. This is, after all, one of the rationales for employing the disagreement paradigm in the first place. We don't expect ordinary people to respond to questions such as "What is your metaethical position?" in an especially informative way.

Nevertheless, *even if this is what we should do*, that's not a fault with *my* study, it's a fault with the disagreement paradigm itself. This is, after all, a metastudy, and if there are methodological shortcomings with the open response question not telling us much about whether the participant's response reflects a metaethical position or not, these are recapitulated by the original study itself. That is, even if we cannot infer much from analyzing written explanations for why participants responded as they did, we are *also* far from being able to confidently infer that the disagreement paradigm itself provides a valid measure of metaethical stances or commitments in the absence of such evidence. Perhaps people's explanations for their responses are not especially diagnostic of how they interpreted the question. But this still leaves us with the question of how they did interpret the disagreement paradigm. The onus is on those who do think people interpret such questions as intended to demonstrate that this is the case.

### S4.5.2 Generalizability

Almost all participants were US residents on Amazon's Mechanical Turk, with the exception of the reanalysis of Goodwin and Darley's data (2008), which consisted of Princeton students. For many

studies, I collected little or no demographic data[185], which limits additional analysis of demographic differences among participants. To the extent that MTurk participants reflect the US population as a whole, we may generalize to people in the US. However, the US is the epitome of a WEIRD population (i.e., a nation that is, at least compared to other nations, more **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic, Henrich, Heine, & Norenzayan, 2010). Since WEIRD populations tend not to be representative of humanity as a whole, findings reported here may not tell us much about humanity in general.

Furthermore, since no efforts have been made to assess distinct subpopulations, we also have little insight into differences based on religion, gender, ethnicity, native language, cultural background, membership in particular communities or subcultures, or neuroatypicality. It may be that people from distinct populations may be more or less adept at providing clear intended interpretations, or that patterns of interpretation may differ among different populations. This could be due to some populations having a higher proportion of people with determinate metaethical stances or commitments. For instance, members of an Amish community may have a well-developed and articulable account of their view of morality which could be readily expressed in surveys that confound typical US residents. Alternatively, certain concepts may be easier to assess or express in other languages; as such, it is possible that something about English renders items used to express metaethics especially ambiguous or difficult to interpret, which would result in low interpretation rates *even if flue*nt English speakers have determinate metaethical stances or commitments. Of course, this is all

---

[185] This was largely due to resource constraints: I was simply unable or uncomfortable spending more money to collect demographic data for what amounted to *thousands* of responses. Given that most samples were drawn from MTurk, there are unlikely to have been many surprises I failed to identify. Nevertheless, I believe demographic data is important for, among other things, assessing the generalizability of results, and I do not intend to avoid collecting it in the future. In addition, identifying differences within subpopulations may be of interest for exploratory purposes, and to generate hypotheses about differences along demographic lines. However, those hypotheses are not central to the analyses discussed here. Future studies can, and should, seek to identify differences in interpretation rates and themes for different subpopulations. However, targeting distinct populations may be a more viable approach. One limitation of such efforts is that this would require collecting prohibitively large quantities of open response data, which would be extraordinarily time-consuming to code.

speculative. There may be few significant differences across populations. I suspect this is generally the case, and that cross-cultural research and focused efforts to assess interpretation rates among different populations would reveal a similar pattern of low intended interpretation rates, and that the same would hold for assessing interpretation rates among demographic subpopulations. There are no obvious *a priori* reasons to suspect that, e.g., Italian speakers, the elderly, or members of the cosplay community would be especially likely to provide lucid interpretations of metaethical stimuli, though perhaps *some* communities would perform notably better, e.g., perhaps members of a devout religious community with substantive knowledge of religious doctrine may reliably respond as realists. It remains to be seen.

At present, generalizations towards populations outside the United States should be made with caution. At best, it would appear that most adult US residents on Amazon's Mechanical Turk do not clearly interpret questions about metaethics as intended. While these findings suggest that most people in the United States do not interpret such questions as intended, our confidence that such findings would emerge in other populations should decline for societies as they are increasingly dissimilar to people living in the United States, and to WEIRD populations in general, e.g., it's plausible that people in Canada or Australia would exhibit similar interpretation rates, but it is less clear how the Pirahã or Hadza would respond. The possibility that such findings may be an idiosyncratic feature of WEIRD populations is not without merit. One reason why this may be the case is that the very notion that there is a distinct category of moral norms is *itself* culturally idiosyncratic, and that it originated in the precursors to contemporary WEIRD societies. As Stich (2018) argues, there may be no principled distinction between moral and nonmoral norms precisely *because* thinking in moral terms is a culturally idiosyncratic phenomenon that originated in particular cultures and, to the extent that it is present in others, is present only via cultural diffusion, and *not* because humanity as a whole is innately predisposed to think in distinctively moral terms. According to Stich, efforts by both philosophers

and psychologists to offer a unifying account of morality have failed because "There *is* no correct definition of morality. *There is no moral domain*" (p. 554). Stich concludes that: "…the conviction that there must be a natural or well-motivated way of dividing normative judgments into those that are *moral* and those that are *nonmoral* is, I think, an illusion fostered by Christian theology and Western moral philosophy" (p. 554). Machery (2018) builds on this thesis by presenting the *historicist view* of morality, which "proposes that morality is culturally specific—morality is only found in some cultures—and instead of being a product of evolution, it is a product of particular, still ill-understood, historical circumstances" (p. 259). Much of Machery's discussions focuses on the role distinctively moral thought plays in Western populations:

> This body of evidence suggests that Westerners' distinction between moral and nonmoral norms is more than just verbal: Rather, it marks distinct psychological constructs. But do other cultures draw a similar distinction? And does it have the same psychological significance? (p. 262)

Machery goes on to discuss findings which hint at the possibility that the answer is "no." While *normative* thought and language may be universal and a product of evolution, distinctively *moral* thought and language may not be. Machery notes that "In line with the proposal that normative cognition is a fundamental building block of cognition, deontic modals—that is, words translating *ought*—and translations of the normative predicates *good* and *bad* are apparently found in every language" (p. 262). However, Machery continues:

> […] expressions related to the moral domain in the United States are not found in all languages. Whereas judgments about whether something is "right" and "wrong" in the United States are tightly connected to whether the action belongs to the moral domain […] translations of *right* and *wrong* are not found in every language. (p. 262)

Recent evidence corroborates these findings. For instance, Berniūnas (2019) has recently addressed how the English term "moral" is translated in Mongolian. Typically, "moral" is translated as *yos surtakhuun*. However, Mongolians typically appeal to a different normative concept, *khündlekh*—which

more closely captures something like "respect"—when describing paradigmatic normative considerations such as harm and fairness. According to Berniūnas,

> The lack of convergence between moral and yos surtakhuun suggests that the term "moral" does not refer to universal "moral" cognition that specifically deals with harm and/or fairness. On the contrary, I would argue that the term "moral" brings to mind exclusively WEIRD associations, and yos surtakhuun brings to mind specifically Mongolian associations. (p. 59)

Similar dissociations may be present in other languages as well e.g., English translations of "moral" may not perfectly map onto conventional Chinese translations of *daode* 道德 (Dranseika, Berniūnas, & Silius, 2018). In short, it may be that there are subtle, or not-so-subtle differences in the way people from different cultures carve up normative concepts, and these ways may differ to varying degrees from WEIRD conceptions of morality. If such variation is present, it would suggest that normative categories are culturally constructed and historically contingent, and that, consequently, the metanormative properties associated with various normative domains could themselves exhibit cultural variability. In other words, if moral terms and concepts are culturally specific, then whatever metanormative characteristics are exhibited by moral terms and concepts may be culturally specific as well. If other cultures don't think in accordance with WEIRD conceptions of morality, it makes little sense to wonder whether they are *moral* realists or antirealists in particular. After all, you can't fall on one or another side of a dispute about a given concept if you don't have the concept! Of course, we are in the early days of such research, and it remains to be seen how different cultures conceptualize normative domains. Early evidence suggests that such results are likely to raise serious challenges about the degree to which we could generalize from WEIRD populations. Machery and collaborators suggest, for instance, that while people from the United States draw a clear distinction between moral and nonmoral norms, that "Indian participants do not seem to draw the distinction between moral

and nonmoral norms, suggesting that the moral domain may not be a universal" (Machery, 2018, p. 263).

Even if we were to conduct cross-cultural research, we would face enormous difficulties: how are we to translate items which employ terms that may carry culturally idiosyncratic and specific meanings that cannot be perfectly translated? Any efforts to ensure adequate translation may themselves only be evaluable by people with an education and training in philosophical traditions that are distinctively WEIRD and may prompt biases into the very means by which we might ordinarily judge the adequacy of a translation, or the validity of an item, or the meaning of a response. That is, anyone trained in contemporary analytic metaethics will necessarily require training in a distinctively WEIRD mode of thinking. There may be no way to engage with and evaluate cross-cultural research on the psychology of folk metaethics that does not involve understanding WEIRD notions of morality, which may be incredibly difficult to acquire without *internalizing* the relevant terms and concepts, and, in virtue of doing so, rendering oneself susceptible to whatever biases or mistakes that might accompany such understanding, e.g., the curse of knowledge (Birch & Bloom, 2007; Hinds, 1999; Ryskin & Brown-Schmidt, 2014).

Note, however, that the threats posed by the poor generalizability of these findings do not threaten either of the core hypotheses. If the goal of a particular psychological hypothesis were to identify general features of our shared human psychology, then the poor generalizability of replicable findings within one population to other populations would undermine such efforts, or at least result in those phenomena being constrained cultural or other factors distinct to some but not other populations. However, my goal is to argue that (a) people do not interpret questions about metaethics as intended and that (b) this is because most people don't have determinate metaethical stances or commitments. While we cannot be sure that a more representative sample of US residents, or samples from culturally diverse participants that better reflect humanity as a whole would not reveal

subpopulations that *do* have determinate stances or commitments, the primary reason why the present findings lack generalizability is in large part due to the possibility that the very terms and concepts on which metaethical distinctions are predicated, *themselves*, don't generalize to other populations. If this is the case, there would be little reason to think untested populations would be competent with the relevant concepts and distinctions (i.e., moral realism and antirealism, or their derivatives, e.g., naturalism and non-naturalism, relativism, noncognitivism, and so on). For comparison, consider a sport like cricket. Cricket is often perceived to have baffling and arcane rules. Outside the UK and former British territories, few people are likely to know the rules. If we found that people *within* these nations reliably failed to understand questions about the rules of cricket, we would hardly expect people *outside* these nations to do any better. Quite the opposite. Just the same, societies that lack moral concepts, or think of moral concepts in very divergent ways, are, if anything, *less* likely to interpret questions about metaethics in accordance with categories and distinctions devised by members of WEIRD populations and predominantly tested on and validated in WEIRD samples. Even so, it remains an open possibility that there are cultures, languages, and ways of thinking that encourage more determinate conceptions of normative moral standards that approximate moral standards well enough that such populations would exhibit higher intended interpretation rates, and more determinate metaethical stances and commitments.

We also have little insight into how well academic philosophers would perform on these measures. I suspect those who specialize in metaethics or ethics in general will perform very well, while philosophers who do not specialize in these areas may do surprisingly poorly (though still better than laypeople). That remains to be seen. If it turns out that those with varying levels of relevant domain expertise still perform poorly on these measures, this could be explained in at least two ways. Most likely, this would indicate that there are significant methodological problems with the methods employed in this chapter. We could think of experts as a proper control, for whom we presume that

if the methods employed in this chapter are appropriate, they are appropriate only if those with relevant expertise would consistently respond in ways judged to be clear intended interpretations. *If* they do not, this could mean that the method itself is inadequate. However, it could also indicate that many philosophers are far less competent with the relevant metaethical distinctions than might otherwise be supposed. This is at least somewhat plausible for philosophers who do not specialize in ethics or metaethics. Such a discovery, while surprising, is at least somewhat plausible.

However, if philosophers who specialize in metaethics provide low intended interpretation rates, this would most likely be due to problems with the studies themselves, since it is implausible that specialists are unfamiliar with the relevant concepts or have no determinate metaethical stances or commitments. Unfortunately, it would be difficult to gather data from specialists. There aren't that many, and those who opt to respond may not represent those who don't for two reasons. First, those who opt to respond may be especially likely to respond appropriately to the questions, and second, they may be people especially likely to be members of similar social and academic networks as researchers who study the psychology of folk metaethics. As a result, they may be especially likely to be familiar with empirical research on metaethics, and would thus already be at least somewhat aware of the stimuli. More generally, many metaethicists are likely *already* aware of this type of research, and are likely to become more familiar with it over time. As such, it may become increasingly difficult to solicit responses from naive expert populations. Notably, it might only be possible to reach such a population once before they would no longer be naive. Nevertheless, such concerns may not be fatal, and it would still be worthwhile to investigate how specialists in metaethics would respond to the items and questions presented in this chapter.

### S4.5.3 Sample quality

Most studies were conducted on Amazon's Mechanical Turk. One concern with focusing exclusively on this population is that they may have been less attentive to stimuli, which could have reduced

intended interpretation rates. Pölzler (2021) has recently made a compelling case that *insufficient effort responding* (IER), i.e., any factors that cause participants to put insufficient cognitive effort into a study, regardless of the cause, can undermine the validity of research in experimental philosophy (Huang et al., 2015). A handful of participants complained about inadequate compensation, suggesting that I may have provided inadequate compensation to participants (especially in early studies, where I was less experienced with timing them). If so, participants may have put less effort into their responses than they would have if they had been adequately compensated. Such responses were not common, and these participants may have been under the impression that I was seeking longer and more detailed answers than I was. Later studies requested that participants provide *brief* explanations and provided smaller text boxes, to encourage shorter responses. Yet several studies exhibited high rates of participants answering other stimuli, but failing to respond to the open response portions of a survey, which hints that participants found such tasks undesirable and opted not to respond to open response questions. In some cases unanswered questions contribute to the low proportion of intended interpretations, but a more troubling possibility is that those participants who did answer did so with less effort than they might otherwise have exhibited. However, this may have been an issue primarily for early studies, which offered a worse pay rate and did not clearly indicate that I was seeking short responses.

Another reason that such concerns are unlikely to account for poor interpretation rates is the presence of open response data from several other sources, which were not gathered on MTurk: data from Goodwin and Darley (2008), Wright, Grandjean, and McWhite (2013).[186] While I do not report the analysis of the latter here, Goodwin and Darley's results showed comparably low clear intended

---

[186] I also obtained and coded data from Sousa et al. (2021). However, their data was collected on MTurk as well, and they report t paying $0.40 for 5 minutes of time, which isn't appreciably greater than the compensation offered in the studies reported here (and may be less compensation). As such, their findings may suffer the same issues as my own. Nevertheless, partial analysis of one of their data sets strongly suggested that most participants did not appear to interpret questions as intended, either.

interpretation rates, while preliminary analysis of Wright, Grandjean, and McWhite's findings likewise reveal extremely low rates of clear intended interpretations. In the case of Wright, Grandjean, and McWhite, this may be due in part to the sheer volume of open response questions asked, which prompted large numbers of participants to not respond. Regardless of the sample, clear intended interpretation rates remained consistently low. Nevertheless, it is possible that open response questions about metaethics are especially unappealing, and that participants are consistently less motivated to invest effort in responding. Questions about metaethics may seem especially obscure, complex, or unusual, which could increase cognitive demand or otherwise result in such questions being especially onerous. If so, providing additional incentive may increase clear intended interpretation rates.

Since I only drew on MTurk participants, any problems associated with sampling from MTurk will be present across all studies. If there are significant differences in interpretation rates when sampling from other sources, even if they consist exclusively of US residents, this would never show up in my findings. Thus, if there are systematic methodological worries with sampling MTurk participants, this would be largely undetectable. This is unfortunate, and future studies should employ other survey platforms and methods of soliciting responses, e.g., employing polling agents to collect data from a representative sampling of US participants. Although I do have data from Goodwin and Darley and Wright and colleagues, supplementing these findings with undergraduates may not be especially helpful, since such populations may be unrepresentative of broader populations, especially given data suggesting that people in their late teens and early twenties are especially likely to endorse antirealist responses (Beebe & Sackris, 2016).[187]

---

[187] Unfortunately only one study explores the question of age differences in metaethical standards, though one could likely assess age differences in studies that collected adequate demographic data. Such findings are also consistent with the widely reported, if anecdotal phenomenon of "student relativism": the tendency for college students to be especially disposed to report relativism or deny moral realism. This is consistent with my own experience, teaching two semesters of a course on moral psychology: in one class (of about 25), *all* students endorsed relativism, while the majority (with the exception of

Another concern with drawing on MTurk participants is that they may differ from other participants in terms of the time and attention they invested in responding to questions about metaethics. MTurk participants may have had little incentive to spend a great deal of time on open response questions, and may have lacked adequate incentive, to put much effort into responding. It is possible that under conditions in which participants are under less of a time constraint, and have greater incentive to think carefully that interpretation rates would substantially improve.

A more general problem with the kinds of open response questions employed in this study is that even if participants are motivated to respond to the best of their ability, the questions may be underspecified or ambiguous. Many responses reflect reasonable interpretations of what they took the open response questions they were given to be asking, but were not the intended interpretation. Ironically, the very means by which I seek to estimate the proportion of participants who interpreted a question as intended is itself subject to the same concern that people may not have interpreted *my* questions as intended. I see no way around this without the diminishing marginal returns of an infinite regress of questions about how people interpreted the preceding question. If so, I may be underestimating the amount of people who did (or would) interpret metaethical questions as intended.

For example, perhaps asking someone why they answered a question in a particular way encourages a variety of interpretations and reactions that are coded as unintended interpretations not because people lack competence with the relevant metaethical concepts, but because there are many

---

one or two) endorsed relativism in the other class. Such commitment to relativism was especially remarkable, given that these students were explicitly given detailed descriptions of the metaethical positions available. I opted to press one student with the standard "What about Hitler?" style of questioning, where I asked if they maintained their relativism even in the face of atrocities. They did not bend under such questioning. Granted, I may have lacked the rhetorical flourishes or authentic line of condemnation that may be expressed by a realist, so my efforts may have been insufficient to embarrass a student into conciliatory concessions to realism. Nevertheless, my experiences add to the veritable mountain of anecdotes attesting to student relativism. Student relativism is a unique and fascinating phenomenon all on its own, and should be the subject of targeted inquiry. I suspect student relativism largely serves a social, performative, and signaling role: students, placed in a novel environment, with their identity and values challenged and up for grabs, may be motivated by the unique nature of their environment to express highly tolerant and agreeable attitudes towards other students and other people and cultures more generally for *social* reasons, rather than philosophically defensible reasons.

ways of interpreting a question about why you performed a particular task that don't result in revealing how one interpreted the question. Take the question posed in Study 5, from the MRS. Participants were first asked to express their level of agreement with a moral statement, such as item #1:

| **MRS #1 Relativism** | Different people can have opposing views on what is moral and immoral without anyone being wrong. |
|---|---|

After expressing their level of agreement, participants are then asked why they chose the response that they did. Yet it is not obvious that explaining why you responded will necessarily reveal that you were expressing a metaethical position. You could interpret this to be a question about what motivated you to answer the question this way, or how you came to hold the beliefs that you do. That is, you might not interpret this to be a question that is soliciting your *rationale* or *justification* for answering, but instead interpret it as a question about what background experiences or psychological factors caused the response. They might even be inclined to think that the question is absurd: why else would you choose the response you did other than because *that's what you think*. This does appear to be how some participants interpreted the question:

*It's what I believe.*

*because that is my opinion*

Participants who respond this way tend to receive a 0|0 code, an "unclear unintended" interpretation, effectively removing them from the pool of participants whose responses are clear enough to judge whether they interpreted the question as intended or not. Yet far from serving as evidence for metaethical indeterminacy, or even that the question about metaethics they were being asked about is invalid, such responses seem most plausibly to simply reflect an unintended interpretation of the open response question itself, or, perhaps, a lack of motivation to put the effort into responding. Either way, such responses cannot tell us one way or the other whether the initial question was valid, or

whether the person in question has a determinate metaethical stance or commitment. The question is how many such responses are present in any given set of data. If they are very common, this weakens the degree to which my conclusions follow from my results. I'm not sure how to judge how frequent such responses are, and I'm not sure how easy it would be to test how frequent they are. It *might* be worth going up one more meta-level and asking people how they interpreted the kinds of questions I've used to ask people how they interpreted questions about metaethics. But whatever the results of this inquiry might be, we could still ask whether people understood the questions I asked about the questions I asked about the questions I asked, and propose we go up to a fourth level of abstraction and ask a question about *that*. As much as I might enjoy a foray into such dizzying heights of recursion, I don't think this would make for especially publishable, or even comprehensible, results.

Such concerns may be less applicable to other forms of open response question. In particular, asking people what a question means, or what a response to a question means, is less subject to the kind of unintended interpretations that may characterize questions about why a person answered a question the way they did. This was, in fact, one of the reasons I asked such questions. As I've shown, interpretation rates remain low in such cases, suggesting that unintended interpretations of the questions posed in this chapter may not be so great a hindrance that they threaten my conclusions.

### S4.5.4 Coding problems

One of the more obvious limitations with the results presented here is that they depend on how well I've coded the data. Coder bias and coder competence both represent serious limitations in how much stock to put in my results. There is only so much I can do to mitigate these concerns. David Moss coded some of the responses reported in Study 1. However, David has the same philosophical inclinations as I do (e.g., a love of Wittgenstein), we've spoken extensively on the topic, and we've worked together on numerous projects. It is safe to say we are far from independent coders, and David may exhibit many of the same biases I do. Ideally, independent coders who don't know me could be

enticed to code some of the data sets themselves, so that results could be compared. This would be especially ideal if some of those coders were adversarial, in that they explicitly held different metaethical views from my own (e.g., endorsing moral realism or some determinate antirealist stance such as error theory), or more importantly were critical of my views about the determinacy and of folk metaethics. Until such coding is complete, it remains an open possibility that my analysis does not accurately reflect how people interpreted questions about metaethics.

It's even possible I've *overestimated* intended interpretations. I may, for instance, have been so afraid of coder bias that I overcorrected and exhibited excessive levels of charitability towards responses, coding some as clear intended interpretations when I wouldn't have if I was less anxious. It's hard to know for sure without directly comparing my results with other people's. To enhance the likelihood that others take up the call to code the data, I've endeavored to make all of the data publicly available, and I encourage anyone interested in coding the data to do so on their own, without contacting me, or to get in touch. Either way, I do not think my eyes should be the only ones on the data.

### S4.5.5 Introspective access

Think of the first animal that comes to mind. What is it? A dog? A pygmy marmoset? A longisquama[188]? Whatever it is, consider *why* that particular animal came to mind. Whatever response you provide, there is a good chance it reflects little more than a post hoc rationale or confabulated account, i.e., a theory about what you think *may* have driven your response. This is because people may lack introspective access to the psychological processes prompting their judgments or behavior, or at least struggle to verbally report such processes (Block, 2011).

In their classic paper, Nisbett and Wilson (1977) argue that people lack access to the cognitive processes prompting their judgments and behavior by reviewing a variety of research which suggests

---

[188] You probably didn't think of longisquama, but you should!

people are frequently oblivious to stimuli that appeared to casually influence their responses or the causal impact stimuli or specific processes have on their judgments and behavior (see also Wilson & Nisbett, 1978). Indeed, even when informed of the potential influence of unconscious processes, people often insist such influences did not influence their judgments, even when available data suggests otherwise (e.g., McPherson & Frantz, 2006; Pronin, 2007; Pronin, Lin, & Ross, 2002). While critics maintain that such severe pessimism is unwarranted[189] (see e.g., Newell & Shanks, 2014; Smith & Miller, 1978; White, 1980), we need not adjudicate the degree to which people lack introspective access to the psychological processes prompting judgments here. Instead, we may frame this concern as a conditional: *if* people lack access to their psychological processes influencing their responses, or at least have difficulty verbalizing those processes, this could represent a devastating criticism of the findings reported in this chapter.

Consider one of the standard questions presented in this chapter. Participants were asked to explain why they chose whatever response they chose. It is possible that ordinary people's moral judgments are implicitly committed to particular metaethical presuppositions, but that they lack introspective access to the metaethical commitments driving their moral judgments and behavior. If so, people may be competent at engaging in moral judgment in a way that does conform to e.g., realism or antirealism, but are unable to explicitly report on or verbalize these commitments. Present findings cannot directly or decisively exclude this possibility. As such, I must simply concede that it is possible that ordinary moral judgments are governed by an unconscious commitment to various metaethical positions. Nevertheless, there are several reasons to doubt this possibility or to question its relevance.

---

[189] For instance, Rich (1979) criticized Nisbett and Wilson's arguments as ambiguous and underspecified, while Smith and Miller (1978) argued that Nisbett and Wilson's initial formulation of their anti-introspectivist view was unfalsifiable, that their conception of mental processes is ill-defined, and that the studies they use to support their claims employed inappropriate analyses. Likewise, in a retrospective analysis ten years after the initial publication of Nisbett and Wilson's article, White (1980) reiterates many of these concerns, and adds that it's not clear whether verbal reports are an appropriate test of introspective access.

At the very least, the onus is on those who do believe people exhibit an implicit competence with metaethical concepts to demonstrate that this is the case; one cannot simply *presume* that realist and antirealist commitments play a causal role in ordinary moral judgment and behavior without compelling arguments or evidence to justify such a presumption.

First, skepticism about introspective access largely concerns whether people are able to verbally report the psychological processes that caused their judgments or behavior. Yet, for the most part, this is *not* what the present set of studies is asking people to report. Even in the case of asking people why they answered as they did, the goal is not to ask people about the potential influence of an obscure and unfamiliar psychological state, or to report all relevant causal influences on their judgments. Studies purporting to show that people lack introspective access to unconscious psychological processes often involve exposing participants to novel stimuli, then assessing whether people were aware of the stimuli or its influence on their judgments. Even if people are unaware of implicit heuristics or biases, such as anchoring or the availability heuristic, the present set of studies are not seeking to assess whether people are aware of influences like these, but to ask them to provide the reasons why they offered a judgment in a rather direct way. That is, the goal of asking people why they answered in the way that they did is to assess whether people have introspective access to some psychological process a person may not be explicitly aware of, but to ask them why they answered a question the meaning of which is *supposed* to be transparent, and which they are *supposed* to have interpreted in a way that *is* accessible to conscious awareness. For comparison, if a person is ordering a pizza, and is asked which toppings they would like, the question:

*Which toppings do you like on your pizza?*

This isn't a trick question, and such a person hasn't been placed in some experimental condition where some manipulation was intended to induce some psychological state that may be introspectively inaccessible. The point of such a question is to prompt that person to draw on their *explicit* knowledge

of their preferences, and to simply report them. The transparency of such a task is so blatant that asking the person why they chose those particular toppings would border on redundant, and might prompt a strange reaction: "Why do you mean, *why*? Because those are the toppings I prefer…" If our goal were to merely assess whether this person understood the question as researchers intended, the purpose of a *why* question isn't to draw out some hidden process or cause, but an ostensibly fairly straightforward one. Nevertheless, we may grant that people lack introspective access to some implicit metaethical commitment prompting their response, and are thus unable to report on it. However, *even if* this is the case, their response may still provide some indication of whether they interpreted the question as intended. If their response suggests an unintended interpretation inconsistent with interpreting the question to relate to metaethics, this shifts at least some of the explanatory burden onto anyone assuming people are interpreting questions as intended.

More generally, *why* questions do not involve subtle manipulations, nor are they intended to assess whether people recognize subtle psychological processes. *Ex hypothesi*, if people had explicit metaethical stances, one might expect them to appeal to or describe such standards, or at least allude to them, when responding to questions intended to assess their metaethical views. That is, *if* the purpose of studies intended to assess folk metaethical views is to assess their explicit metaethical stances, such results at the very least challenge such claims. Yet even if our goal is to assess whether people have implicit metaethical commitments, we might still expect people's explanations for their responses to at least be *relevant*. If they consistently were, this would support the conclusion that people's responses did reflect their metaethical commitments. Given that they typically are not, this counts against the presumption that people have implicit competence with and commitment to metaethical positions. After all, if you ask someone a true or false question, and they say "true," but when you then ask them to explain why they chose "true" and their response has *nothing to do with the question you asked*, but *does* appear to be a reasonable response if you were asking a slightly different

question, this at the very least increases the probability that they thought you were asking some other question than the one you intended. And this kind of understandable irrelevance that suggests participants were responding to a different question than what was asked is precisely what we find when considering the most prominent themes to emerge via thematic analysis. It is puzzling that many people's explanations for their responses reliably reflect certain thematic patterns that recur across different studies, such as the conflation between realism and absolutism or metaethical relativism with descriptive relativism. Such explanations suggest that people did not interpret the question in metaethical terms, or at least not the intended metaethical terms.

Even so, it is still possible that participants have implicit metaethical commitments, and that asking them to explain why they answer moral questions in the way they did is simply incapable of revealing their commitment to or competence with those concepts. Yet concerns about lack of introspective access cannot readily extend to the other methods employed in this chapter. In study 1, participants were asked to explain why they thought the other person disagreed. In this case, there was no fact of the matter (since the other person and their response were fake) about why the other person disagreed. More importantly, a question about why someone else disagreed with you isn't even attempting to solicit a report about one's own psychological processes, so concerns about introspective access make little sense: such concerns just aren't what these questions are about. In fact, participants don't even need to be *correct*. So long as they point to one or more reasons why someone may have disagreed with them other than having a fundamental moral disagreement, it remains possible that such possibilities were salient and played an active role in how they responded to the disagreement paradigm. While it may be difficult to determine whether their model of the disagreement between themselves and a previous participant actually did incorporate unintended beliefs about the cause of the disagreement, the fact remains that such presuppositions are plausible, and the fact that participants readily reported such possibilities when asked prohibits any confident

presumption that the intended interpretation of the disagreement was the only salient interpretation when participants were responding to questions in the original study. After all, if your study *requires* people to attribute the source of disagreement to *x*, and when someone is asked why they think the person disagreed with them, they say "perhaps because of reasons *y* or *z*" but they *don't* mention *x*, it would be absurd to just assume, without any evidence, that *x* was the *only* factor relevant to their initial response when they did respond. Sure, perhaps participants only considered possible causes of moral disagreement after the fact that were totally unrelated to the cause they assumed was the case when they did respond, but again, are researchers going to simply *assume* that the intended interpretation was the dominant one even when participants are readily capable of offering a variety of other interpretations?

Questions about the source of a disagreement are not the only type of open response question that cannot plausibly be dismissed on the grounds that participants lack introspective access. I have also asked participants to *explain* what various terms or items mean. Reliably, in *every* study, fewer than half of participants offered an interpretation of an item or term (e.g. "objective") consistent with the intended interpretation. Such questions don't concern access to psychological states, but to the meaning of words and sentences. While people can be competent with terms or concepts even if they struggle to verbalize their competence, the fact that they consistently provide unintended interpretations is evidence that they don't. While items that appear on scales such as the FMO or MRS may appeal to efforts to validate the items on these scales, traditional methods of survey validation, such as a high Cronbach's alpha or decent factor loading, may be insufficient: thematic analysis suggests that there systematic, recurring patterns in the ways people interpret questions in unintended ways, which could cause *consistent* unintended responses. After all, if someone interpreted a dozen questions about "the bank" in an unintended way (e.g., a riverbank, rather than a financial institution) their responses would be consistent with one another, even if they weren't consistent with what the

researcher sought to measure. Consistency in responses *is not enough*, a point demonstrating in an amusing and startling way by Maul (2017), who found that substituting conventional scale items for sentences composed of uninterpretable nonsense (lorem ipsum) or even completely blank items still resulted in high internal consistency and acceptable factor loadings. Likewise, agreement among experts that a given set of items is face valid (as in Collier-Spruel et al., 2019) is also not sufficient: simply because researchers understand a given set of items doesn't mean laypeople will.

Another reason to doubt implicit competence with the relevant terms or concepts is that the recurring themes that emerge in thematic analysis don't suggest that people are simply ignorant of the relevant terms or concepts. Instead, people appeal to a variety of *plausible* and *meaningful* alternative interpretations that, if anything, reveal their linguistic *competence*. The problem with many questions that appear on metaethics scales, and with terms like "objective" and "relative" isn't that they are so obscure and hard to interpret that participants are reliably baffled. Rather, it's that they are underspecified, ambiguous, and capable of being interpreted in a variety of ways unrelated to the intended interpretation. Recurring patterns of interpretation reliably emerge for particular items, revealing that participants tend to be picking up on the same patterns and meanings as one another; they aren't simply responding randomly because they have no idea what the items mean. This suggests that interpretations often reflect meaningful ways of construing terms and phrases that are available to competent English speakers; they just aren't the meaningful interpretations researchers expected. The consistency across participants, and the intelligibility or even reasonableness of their interpretations suggests both that participants are not interpreting terms and items as intended *and* that they are competent with the terms and phrases used in studies on metaethics.

In short, the problem is not that participants lack competence with the relevant terms or concepts, but that the intended meaning of the terms and concepts parasitizes words and phrases that *already have a variety of plausible meanings and interpretations*. There is a reason, after all, while academics

develop specialized jargon in their respective fields of inquiry: to cut down on the ambiguity, underspecificity, and confusion that would result from employing everyday terms and concepts, which have a wide variety of possible meanings and are typically interpreted in ways that draw heavily on the relevant context of utterance. This is why I must reiterate that, once again, the *context* in which a term or phrase occurs isn't critical, but *essential* to fixing its meaning. The exact same sentence can mean something completely different in different contexts. Statements about metaethics, and the central terms used in those statements, are no exception. The term "objective" can mean "unbiased," or "subject to a publicly evaluable and quantifiable standard of evaluation," e.g., a thermometer provides an "objective" measure of the temperature (as opposed to consulting an individual's subjective report on what they think the temperature is based on how it feels to them). Think about the following sentences:

> *Alex did not believe her boss could provide an* objective *assessment of her work performance, given that her boss was also her father.*

> *Alex and Sam could not agree on how cold it was outside, so they agreed to use a thermometer to get an* objective *measure.*

The term "objective" is not a novel piece of jargon invented specifically for conveying claims about metaethics. It is a polysemous term the meaning of which varies depending on the context in which it is used. Few people would struggle to interpret either of the statements above, yet "objective" means something different from what "objective" means in metaethics. To make matters worse, the various colloquial uses of "objective" are conceptually related. "Objective" *isn't* a word whose meanings are *merely* discrete, distinct differences in meaning, such as the term *bank* referring to a financial institution or a riverbank. Rather, there is so much conceptual overlap between various colloquial uses of the term and its philosophical conception that its bizarre researchers would expect participants to exclusively interpret questions about moral "objectivism" in a narrow and highly specific way. Furthermore, when people are asked about financial institutions and riverbanks, there's typically

enough context to dissolve the ambiguity. Not so with questions about metaethics. Participants are typically given little or no context. Thus, not only are they given ambiguous sentences, they are given in the sterile, low-context environment of a survey, where one's ability to resolve the ambiguity in the intended way is severely impaired.

Worse still, researchers expect people to interpret terms and phrases intended to convey objectivism in a way that does *not* reflect standard colloquial usage of the relevant terms and concepts. Suppose you asked participants how much they agreed with claims like the following:

*All first-order normative claims are false.*

*There are no stance-independent moral facts.*

*We have categorical epistemic reasons, but we do not have categorical moral reasons.*

These claims are *jargony*. Ordinary people would not understand them almost by definition: to understand these sentences *just means* you're no longer an ordinary person with respect to them. Since researchers are wise enough to recognize that they can't use items like these, they've opted for attempting to convey claims about metaethics using ordinary language.

This is the heart of the problem: metaethicists developed novel terms like "categorical reasons" and "first-order normative claims" precisely *because* ordinary language is too ambiguous and imprecise to convey what they wish to convey. Ordinary language is so inadequate for the task of conveying metaethical claims that, if anything, it serves as an active impediment to clearly conveying metaethical concepts. Philosophers spend as much or more of their time merely trying to clarify what a term like "stance-independent" means as they do arguing for whatever position they take up. Even then, they still consistently fail to convey what they mean to their colleagues, given how often the published literature is populated with accusations of misinterpretations. It is *incredibly difficult* to adequately convey a particular metaethical position. The whole point of abandoning colloquial terms is to circumvent the ambiguity, imprecision, and underspecificity that invariably accompanies ordinary

language. Researchers may find it necessary to employ such language, but it may not be possible to prompt the intended interpretation with ambiguous, underspecified terms and phrases that lack adequate context. In short, researchers leave participants without the standard resources to figure out what they're being asked.

This provides a simple and straightforward explanation for why participants appear to interpret metaethical stimuli in unintended ways: *because they do interpret the stimuli in unintended ways*. The alternative explanation requires us to imagine that participants are competent with the concepts and respond appropriately to questions about metaethics, but are usually unable to clearly verbalize their understanding. While possible, this strikes me as a strained and implausible position to take given that there is little data or compelling theoretical rationale to presume people would be competent with metaethical concepts, much less that they'd have a determinate metaethical stance that conforms to traditional philosophical theories.

I grant that, absent evidence or arguments to the contrary, researchers who carefully design a study for the purpose of measuring a particular psychological phenomenon may presume that their methods do so, provided they at least superficially appear to do so by passing some minimal bar of face validity. However, my goal has been to shift the burden of proof on researchers who presume people do interpret questions about metaethics as intended. After all, why should the burden be on me to show that people *don't* interpret questions as intended, rather than on the researchers asking the questions to show that they *do*? I've offered a massive body of evidence which offers at least some *prima facie* evidence that people do not interpret questions about metaethics as intended, and I have supported this with an avalanche of supporting theoretical reasons to expect low rates of intended interpretation *because* (a) items often exhibit poor face validity and (b) people don't interpret what they're being asked as intended even when an item could reasonably convey the intended metaethical

position. If I'm mistaken, those who believe most participants do interpret questions about metaethics as intended should be able to provide evidence that they do.

### S4.5.6 Tension between interpretation rates and indeterminacy

Even if most people do not interpret questions about metaethics as intended, this does not directly demonstrate that people have no determinate metaethical stances or commitments. *Why* people do not interpret metaethical stimuli as intended is just as important as establishing *that* they don't interpret metaethical stimuli as intended in the first place. People could still have determinate metaethical standards even if the measures used to assess folk metaethical beliefs are not valid, for the same reason people could still believe in God even if you gave them a survey they didn't understand because it was presented in a language they didn't understand. If they were asked in their native language, then their responses would readily reflect their beliefs about God. Likewise, it's possible that the reason people do not interpret questions about metaethics as intended is because they don't understand what they're being asked, but that *if* they understood what they were asked, *then* they'd be able to give answers that would reveal determinate metaethical stances or commitments.

I have gone out of my way to argue that many of the measures used to assess folk metaethical stances and commitments have poor face validity, employ confusing instructions or ambiguous items, or in a variety of other ways present participants with stimuli that would be hard to interpret even if they had determinate metaethical standards. I have determinate metaethical views myself, but even I would struggle to interpret many of the questions presented in these studies. The low intended interpretation rate could be largely attributable to the methodological shortcomings of existing research. As a result, there is considerable tension between two of my central claims:

(1) Most studies used to assess folk metaethical stances are not valid

(2) Most ordinary people have no determinate metaethical stances or commitments

While these claims are not *incompatible*, evidence of (1) limits the degree to which we can infer (2). If our measures were far better, and people *still* consistently failed to interpret them as intended, this would be far stronger evidence for indeterminacy.

Although I will raise some points to soften the blow, I cannot completely overcome this concern. I simply concede that it's possible the reason why so few participants have interpreted questions about metaethics as intended is simply due to the shortcomings of previous studies, and that future studies could reveal ordinary people's metaethical stances and commitments. The best solution to this problem is to simply devise better measures, then assess interpretation rates under these improved conditions.

There are a few problems with this, however. First, this has, to some extent, already been done: the MRS was carefully developed with an eye towards greater face validity, and was at least somewhat successful. And while interpretation rates went up, they didn't go up enough to lend much confidence to the notion that *most* people would interpret questions about metaethics as intended. Furthermore, Collier Spruel et al.'s (2019) findings are limited by only evaluating views towards relativism. Even if people had some understanding of relativism, it would not follow that they had an understanding of stance-dependence and stance-independence, error theory, noncognitivism, and so on. Furthermore, the "clear intended" interpretation rates may *overestimate* interpretation rates for reasons discussed in the main text.

I will outline a handful of instances from Study 5 to illustrate the point. Consider the following responses, all coded as clear intended interpretations. Here are two examples from the *Why* condition:

> *Moral values often depend on the individual.*

> *there is only one true moral answer in any situation*

Here are two examples for the *Explain* condition:

> *morality is subjective.*

> *This means that morality is subjective. This would suggest that there are no absolute morals - that nothing is absolutely wrong.*

These are decent responses. The notion that there is only "one true moral answer" *seems* like an expression of realism, while the claim that moral values "depend on the individual" *seems* like what a relativist might say; after all, almost all accounts of relativism hold that moral claims *depend on* the values of individuals or groups; this kind of remark is very similar to what a professional philosopher might say. But there is imprecision, slippage, and ambiguity even in these "best of" cases. That is, these responses are of sufficiently high clarity that they were coded as *clear* intended interpretations, for the simple reason that they *appear to* convey the intended kind of metaethical stance.

Yet it is possible even these participants did not interpret questions about metaethics as intended. In fact, I suspect this is *likely* to be the case, but suspicions are no substitute for data, and deep skepticism is no justification for coding every response as unclear or unintended. In this analysis, more so than in quantitative research, one's suspicions must be kept on a tight leash, or one will simply see in the data whatever they wish to see. Nevertheless, readers may not appreciate *why* I am so suspicious of these sorts of responses. So let's consider one such item:

> *Moral values often depend on the individual.*

While this is the sort of thing a moral relativist might say, it would be more apt for a relativist to say that *moral truth* depends on the beliefs or values of individuals. This item does not explicitly state that moral *truths* depend on the individual. Even if they did, we could still reasonably wonder what they mean by "depend": depend *in what way*, exactly? A moral relativist may maintain that the standards of individuals or groups "serve as truth-makers" or "make true" the moral standards held by those individuals or groups, or they might offer some other technical language or philosophical argot. Ordinary people lack this vocabulary, even if they possess the appropriate concepts, so it would be a mistake to judge their responses by the standards we'd have for professional philosophers. More

generally, this statement could have a variety of meanings that would not amount to moral realism. It could be a descriptive claim; that is, the participant could simply mean that different individuals have different moral values. For instance, we might say "food preference depends on the individual." This requires no stance about the nature of truth, or the meaning of food claims (e.g., "gastronomic cognitivism"), but may instead simply reflect something like:

*"There is no general fact about what foods people prefer; rather, different individuals prefer different foods."*
Such a remark would be consistent with all metanormative stances about food preferences. Just the same, such a remark, when directed towards moral values, may not indicate an actual metaethical stance.

As I demonstrated in Study 4, most people do not appear to interpret direct statements about morality being "objective" or "relative" as intended. In light of this, what should we make of the two remarks from the explain condition above? Here they are again:

*morality is subjective.*

*This means that morality is subjective. This would suggest that there are no absolute morals - that nothing is absolutely wrong.*

It seems unlikely that participants would almost never interpret "objective" and "relative" as intended, but interpret "subjective" in just the way philosophers use the term in contemporary metaethics. When a person says that morality is "subjective," this simply mirrors the language philosophers use. A coder wary of their personal biases, who has little knowledge of the actual base rates of comprehension for terms like "subjective," no further context to off of given the participants' lack of elaboration, and who is mindful of the optics of claiming that someone saying "morality is subjective" probably *doesn't* mean that they think *morality is subjective* (to convey the meaning of this phrase in contemporary metaethics) is in a tough spot: what am I supposed to do with this remark? I've opted to give such people the benefit of the doubt, and to follow the interpretative principle that if it looks like a duck…

Nevertheless, someone who claims that morality "is subjective," may conceive of the notion that morality is subjective in a variety of ways that either fail to adequately reflect subjectivism as a metaethical position (i.e., moral claims have an implicit indexical element such that they are true or false relative to the standards of agents, appraisers, or both). For instance, I have considerable experience engaging in philosophical discussions with laypeople. More times than I can count, someone has used the term "subjective" to refer to the notion that any given instance of a claim that a given proposition is true or false is only held according to a particular person's point of view, that is, some claim "*p* is *true*" is "subjective" in the sense that e.g., Alex believes "*p* is *true*," and, in virtue of this being Alex's *personal belief* on the truth status of *p*, this given instance of the claim "*p* is *true*" is a "subjective" claim, i.e., it is merely Alex's subjective position that p is true. Some people will even go so far as to say that all truth claims are "subjective" in this sense. What they appear to have in mind is that any actual assertion about what is true or false is always expressed by a person with a particular point of view, and thus all beliefs are necessarily "subjective", where "subjective" *just means* that they are held according to some point of view. However, this has *nothing to do* with whether "*p* is true" is true or false in the respect meant by philosophers who endorse *subjectivism*. Take two claims:

Alex: *"Carbon atoms have six protons."*

Sam: *"Carbon atoms have nine protons."*

A scientific realist who believes such claims are not subjective in the sense the term is used in philosophy would consider it trivial to acknowledge that both such claims are held by individuals, and that in some trivial sense both claims are "subjective" in that they reflect the perspectives of particular subjects, subjects who are potentially fallible, biased, and making such claims against a background set of beliefs and commitments.

These considerations highlight how *low* a bar I have to set to interpret any non-negligible number of responses as intended interpretations. Were the bar any higher, virtually no participants in

any study would be coded as a clear intended interpretation. *Some* concession must be made for the sake of charitability, and I had to draw the line somewhere. Participants were not asked to write an essay, were not trained on the subject, and typically offered very brief expressions of their views. Even in cases where we know that a person is highly competent with a concept, a brief one-sentence description offered on the fly may be far from perfect and unambiguous. Unfortunately, the only way to establish a greater base rate for the proportion of participants who understand any particular statement about metaethics would be to run far more intensive and time-consuming research, e.g., dozens or even hundreds of interviews, and such a task may need to be repeated on every new population and for each formulation or item pertaining to metaethics. Such tedium is likely to scare off even the most dedicated researcher. Simply put, the open response analysis I offer is at best a noisy signal that offers glimmers of how people interpret these questions.

Second, interpretation rates remained low even when participants were given responses deliberately constructed to accurately reflect the expression of a metaethical stance. In such cases, they weren't tasked with interpreting a question about metaethics, but understanding a straightforward, albeit short, expression of the relevant kind of metaethical stance, and interpretation rates were still low. In short, participants have already received reasonably clear expressions of metaethical standards and still exhibit extremely low rates of clear intended interpretations. While interpretation rates would likely improve with better items, and such studies should be conducted, it's unlikely they'd move from 5-30% clear intended interpretations to 90% or more with a few tweaks in the wording of items.[190]

---

[190] One proposal might be to assess interpretation rates following more robust stimuli or after some type of training exercises to familiar participants with the relevant philosophical concepts, but this simply trades one problem for another: in such cases, we'd no longer be soliciting the responses of nonphilosophers, we'd merely be soliciting the responses of novice philosophers. In such cases, there is no viable way to know whether such responses reflect pretheoretical stances or commitments, or whether spontaneous theorizing has prompted participants to respond in ways that don't reflect how they would respond prior to exposure to the instructions or training stimuli. Of course, this doesn't mean people don't have determinate metaethical views, but it may mean that conventional scientific methods are not adequate for assessing what those views are.

Yet the most important defense against the objection that evidence of poor validity is in tension with indeterminacy is that the reason why studies suffer poor validity can be in part explained by metaethical indeterminacy. That is, what I propose is that part of the reason it is so difficult for people to interpret questions about metaethics as intended is *because* they have no determinate metaethical standards. Yes, many questions have very poor face validity, e.g. items on the NMQ or FMO scales simply fail to accurately describe metaethical positions, while the disagreement paradigm is far too ambiguous and hard to interpret for results to be meaningful. Perhaps future studies will circumvent these difficulties and present questions about metaethics that people do readily interpret as intended. I predict that this will never happen. The problem is that, even if we expunge every error made by researchers, e.g., asking descriptive questions instead of metaethical questions, or conflating universalism with stance-independence, we would still be left with items that are far too ambiguous and underspecified for ordinary people to reliably interpret them as intended.

That is, even under ideal circumstances, we simply cannot present questions or items that clearly and unambiguously reflect the relevant metaethical distinctions in such a way that ordinary people would reliably interpret them as intended using one-sentence items or even short sets of instructions because these concepts are simply too unfamiliar, technical, and subtle to be conveyed to ordinary people in truncated descriptions via ordinary language. Just as we cannot hope to convey the complexities of string theory or quantum mechanics to ordinary people in simple sentences, without any training or instruction or clarification, so too can we simply not present questions about moral realism and antirealism to ordinary people in a way we can confidently presume they understand. We'd either have to gather empirical data that suggested they did reliably interpret these questions as intended, or we'd have to devise the proper stimuli that would prompt intended interpretations. *Inducing* such understanding is simply not going to work, since doing so removes participants from the pool of "ordinary people."

By comparison, researchers do not face insurmountable difficulties with soliciting responses to questions about food preferences, or religious beliefs, or personality traits. Researchers don't have to bend over backwards and give long explanations or training exercises to prompt people to understand what they mean by "tasty," or "outgoing." While there may be slippage and the occasional unintended interpretation, people understand what they're being asked in these studies *because* they possess the relevant concepts. I propose that the reason why it's so difficult to ask people questions about metaethics in the first place is *because* they don't have the relevant terms and concepts to respond appropriately, and that this isn't merely due to lack of training, but because they don't have determinate stances or commitments in the first place. If so, low interpretation rates aren't a problem researchers can overcome, but an insurmountable difficulty that results from the fact that people simply lack the terms, education, and knowledge to consider the questions in the first place.

In other words, we could have imagined a scenario where researchers simply failed to adequately present items intended to reflect different metaethical positions. This resulted in a particular pattern of responses. Yet once someone came along and employed qualitative methods to assess how people were interpreting these questions, they discovered that they weren't interpreting them as intended. Researchers could then go back to the drawing board, write up a new set of questions, and then give these questions to people, who would then interpret them as intended. If so, then the poor interpretation rates would be due to correctable misoperationalization. Yet this isn't the picture that we actually observe when we examine how researchers have conducted studies on folk metaethics. While some studies do exhibit correctable flaws, even when these flaws are minimized there is still little theoretical grounds for supposing that people consistently interpret questions about metaethics as intended, nor is there any convincing empirical evidence that they do so. Despite my best efforts, so far, clear intended interpretation rates are consistently and astonishingly low in almost every case. More importantly, even efforts to correct for previous errors and to minimize ambiguity

are inadequate. At present, there is no clear way to construct a pool of survey items that convey metaethical positions in a single sentence that are sufficiently clear and unambiguous that we can reasonably expect people to interpret them as intended. Short of extensive instructions, training paradigms, and comprehension checks we have little reason to be confident people are interpreting questions as intended. For instance, I have done my best to carefully construct items that represent different metaethical positions. Nevertheless, I have little confidence participants would perform especially well with these items, either. Here is one example:

| **Realism** | Moral truth is independent of cultural standards and personal beliefs. |

Each of these items is arguably a better representation of the respective metaethical position than the items we've typically seen in metaethics scales and paradigms. Yet there is little reason to believe people would do appreciably better at interpreting these than there is for the items I've already coded. Each is saddled with problems of its own that can be readily anticipated in light of existing findings, previous research, and familiarity with the difficulties of conveying the relevant philosophical notions. For instance, the realism item presupposes that there is such a thing as "moral truth," yet antirealists may not believe there is any such thing as moral truth. Furthermore, we have little knowledge of how participants interpret the term "moral truth," or even what notions of "truth" they'd have in mind e.g., whether they adopt a truth-correspondence notion of truth, and whether this notion is salient when responding to questions of this kind. Pölzler and Wright (2020b) list this among their catalog of shortcomings and limitations with existing research on folk metaethics, noting that "So far no evidence has been provided for the claim that participants understand studies' underlying concepts of truth, rightness or correctness in a correspondence theoretic sense" (p. 58).

This is a problem, because if ordinary people vary in their conception of truth, then we have yet another form of interpretative variation that could result in responses meaning one thing for one

set of participants, but something else for another set of participants, and, critically, some of these responses may involve conceptions of truth that are *inconsistent with an intended interpretation*. Yet Pölzler and Wright acknowledge in their own work that they "assume a correspondence theory of moral truth" when referencing the notion of "moral truth," even though we have not established that this conception of truth is, in fact, operative in the way participants respond. Elsewhere, Pölzler and Wright (2020b) attempted to assess folk conceptions of truth precisely for this reason with mixed results. As they observed, their attempts "sparked lots of confusion, as evidenced by participants' verbal explanations" (p. 19, footnote 15).[191]

This points to a more general problem with questions about metaethics: a philosopher's stance on realism or antirealism does not typically exist independently of the rest of their philosophical stances and commitments, but is instead a piece of a broader, holistic philosophical puzzle; we can see one's stance on each specific philosophical question as a cog in a grand philosophical machine whose parts are constantly being swapped out, modified, and greased in response to the pressures of reflection and argumentation. That is, insofar as philosophers come to reflect on and adopt particular philosophical positions, they do so *holistically*, against the backdrop of other philosophical stances and commitments. The positions philosophers adopt typically hinge on their more fundamental philosophical commitments, e.g., whether they are rationalists or empiricists or pragmatists, whether they are naturalists or non-naturalists, their views on epistemology, and on their metaphilosophical views, e.g., whether they endorse mainstream views on the legitimacy of conceptual analysis, or ordinary language philosophy, or adopt quietistic or deflationary approaches towards one or more philosophical issues. In short, philosophical views don't typically exist in isolation, but are dependent

---

[191] Subsequent attempts were not especially successful. Although Pölzler and Wright appear more optimistic than I am that participant responses provided some *prima facie* evidence that correspondence theoretic notions of truth may be common among ordinary people, evidence is at best thin, indecisive, and hardly establishing anything approximating an overwhelming consensus among the folk.

on a web of background beliefs and assumptions. The holistic and interconnected way in which philosophers adopt particular philosophical stances and commitments influences how those philosophers interpret questions. A response of "yes" or "no" from two different philosophers to the very same question might mean something different because what those philosophers take their response to that question to mean can vary.

I have made a point of emphasizing that ordinary people are likely to react to philosophical questions by exhibiting disproportionately high levels of interpretative variation when compared to conventional social scientific questions, e.g., food preferences or emotions. Such interpretative variation is mirrored and even amplified in the way philosophers would respond to questions about "morality" and "truth." How someone interprets one question about philosophy, and what their response to that question will mean, will depend on and vary as a function of the rest of their philosophical stances and commitments. For instance, suppose we ask a moral realist and a moral antirealist whether they think "murder is wrong" and they both say "yes." This means something *completely different* for the realist and the antirealist. The moral realist may mean "I believe there is a stance-independent moral fact that furnishes us with a decisive reason to not murder," while the antirealist might mean "I disapprove of murder" or "Murder? Boo!" Assent to the first-order normative question, "Is murder wrong?" is underwritten by a different set of metaethical presuppositions, and thus a "yes" masks underlying differences in meaning between philosophers.

Any studies designed to assess philosophical positions about *one* philosophical issue that try to hold the rest of a person's philosophical stances and commitments constant, or that simply ignores them, may be doomed from the outset. Researchers are, in effect, attempting to solicit piecemeal philosophical stances using the same questions and stimuli, without appreciating that this just isn't how philosophy itself works: the same terms and phrases mean different things when expressed by different philosophers. This isn't surprising to anyone who understands the methods of contemporary

analytic philosophy. Take something like conceptual analysis. We might wish to analyze the kinds of claims people make in everyday discourse, e.g.:

*"I know the sun will rise tomorrow."*

Philosophers have engaged in a relentless, protracted battle over the meaning of a term like "know." What they don't typically dispute is that the sentence "I know the sun will rise tomorrow" is the kind of sentence we might expect ordinary people to employ in ordinary contexts. And that's just the point: even if we hold every term constant, and ask nonphilosophers how strongly they agree with some statement, or ask them to select an option from a list of multiple choice options, researchers presume that a "6" on a Likert scale or a "yes" or a "they could both be correct" *means the same thing for all participants*, when *the whole point of philosophical inquiry is to figure out which competing account of what such statements mean is the correct one*. Such disputes would only persist if philosophers reacted differently to such questions, that is, whether they were inclined towards different accounts of the meaning of various terms, words, phrases, and so on. And if the meaning of such terms, words, or phrases, in doubt, with multiple, competing accounts on the table, and with, from all appearances, little consensus among experts who think about these issues for a living, why should we assume that *nonphilosophers* are going to interpret questions, phrased in precisely the same way, in such a way that they all interpret them in the same way, or at least a close enough way, as one another, for any particular measure we employ to be a measure of the same thing across participants? That is, if the very issue philosophers face is that it's not obvious how to interpret some term, phrase, or sentence in accordance with one or another of competing philosophical accounts, and philosophers reach wildly different conclusions, why would we expect this to be any different for ordinary people? Why not expect it to be *far worse*? And if ordinary people's interpretations are wildly different from one another, how can we take what is intended to be the same measure to be interpreted the same way across participants merely because the wording is the same? That a phrase with the same wording could mean different things is the very

issue at stake to begin with! In short, social scientific methods employed in folk metaethics research relies on the presumption that there is detectable individual differences in folk philosophical stances and commitments, that this variation can be measured using a particular set of stimuli and questions, that such variation can piggyback on the same terms and phrases, such that the same terms and phrases can mean different things to different people, but at the same time variation in how people interpret or understand all the terms and concepts employed in the stimuli that *isn't* the subject of the measures *is* being interpreted in the same way, such that there is little or no variation in how people interpret all aspects of the stimuli and questions that aren't being measured.

For instance, when researchers ask ordinary people questions about philosophical issues, they are presupposing that a question with a particular set of terms, arranged in a particular way, e.g., "I know the sun will rise tomorrow" …will be interpreted in *exactly the same way* across all participants, with the *only* exception being the distinction between the relevant philosophical concepts the researchers are interested in, with every aspect of the meaning of the sentence other than the philosophical issue of interest held constant across participants, even though participants have no idea that researchers intend to hold the meaning, and thus the interpretation, of every other aspect of the semantic content, pragmatic implication of the content of the sentence, and so on constant, with the sole exception of whatever it is they are asking about, which is, by design either (a) concealed from participants, rendering any hope of holding everything else constant dubious at best or (b) made clear to participants, which exposes the study to demand effects, spontaneous theorizing, etc. For instance, when participants are asked whether "two people who disagree about a moral issue can both be correct", participants are expected to:

(a) *Uniformly share a truth-correspondence notion of truth*

(b) *Understand "disagreement" in the same way as other participants*

(c) *Understand "moral" in the same way as other participants*

(d) *Presume both sides of a disagreement are sincere, competent, not confused, not lying, etc.*

(e) *Recognize that the moral disagreement is due to a fundamental difference in moral values and not due to differences in nonmoral beliefs*

(f) *Recognize that both people are not talking past one another, misunderstanding the situation, or imagining different moral issues from one another, but are in fact imagining precisely the same situation*

(g) *Understand the disagreement to concern a discrete matter of the truth or falsify of a single shared proposition under consideration; it's not the case that someone could be correct about part of the issue, and the other correct about some other part; no, what is at stake is a simple, atomic proposition of the form "X is morally right/wrong" where X is sufficiently well-specified*

(h) *Exclude epistemic considerations, e.g., the question is about whether both people are correct at the same time and in the same respect, not whether both people could be justified in their beliefs, differ in their access to the evidence, and so on*

(i) *Reliably resolve modal operator scope ambiguity in the same way as one another, and recognize that they are being asked to judge whether both people can be correct at the same time and in the same respect, and not whether it could be that one* or *the other could be correct but not at the same time*

This list is probably not even complete or exhaustive. This may strike readers as an excessively pessimistic and demanding take on what is needed to ensure participant interpretation. But note that social scientific research typically sets aside concerns about the meaning of terms and variation in people's philosophical commitments. Researchers pursuing traditional philosophical questions, on the other hand, are asking the very sorts of questions that have proven recalcitrant in the face of relentless philosophical dispute, much of it centered on the meaning of terms as they are used in ordinary language. If we're to don our philosophical hats, and adopt the view that the meaning of everyday terms and phrases is nonobvious and could turn out to differ in philosophically interesting and substantive ways, and at the same time acknowledge that philosophers virtually never agree or even reach a strong consensus on many of these matters, and instead divide into camps that stubbornly insist on what often amount to diametrically opposed views on the meaning of a given term, concept, or phrase, can we then present questions to nonphilosophers and expect such variation to be absent?

Isn't the potential for such variation the very thing we'd be presuming in the very act of inquiring into folk philosophical views?

In short: The presumption when studying folk philosophical views is that there is measurable individual variation in people's philosophical stance regarding some phenomenon, X. However, *if* this is true, it will typically be presumed to be one example among many of the ways in which ordinary people's philosophical views could vary. And since variation in folk views could influence how they interpret a variety of the terms and phrased used in the stimuli used to ask them about the philosophical issue of interest, their responses may vary not as a function of differences in their philosophical position towards the issue in question, but as a function of differences in their philosophical position with respect to elements of the question that aren't the subject of inquiry, e.g., their notion of "truth" or something being "correct." That is, if people vary in one way, then *ex hypothesi*, they plausibly vary in other ways. As a result, if you ask a question that superficially appears identical in virtue of the fact that you employ the exact same wording, responses may in practice reflect answers to different questions that only superficially appear to be answers to the same question because two participants both selected 7 = Strongly agree" or "(a) yes" or whatever. Just as the realist and the antirealist can mean completely different things when responding to the same question with a "yes," so too could participant responses to the "same" question actually result in different answers. And since you can't hold these constant when asking a question, without clarifying precisely what you are asking in a way that effectively suspends or holds all these background philosophical views constant across participants, when asking the same, simple, one sentence question, then what you effectively get is a set of answers to *different* questions.

In addition to the potential for individual differences in other philosophical stance and commitments resulting in interpretative variation for stimuli and questions that are intended to be interpreted the same way, which could undermine the validity of studies, there is also recalcitrant

ambiguity that may even be ineliminable due to the limitations of ordinary language and the truncated form of questions social scientific surveys typically take, since we cannot, after all, expect participants to read voluminous treatises that clarify every potential point of concern before answering a questionnaire. For instance, for the realist item above, people might even struggle to understand what it would mean for moral truth to be "independent" of cultural standards and personal beliefs: independent in what respect? What does this mean, exactly? The respect in which moral realists think stance-independent moral facts are "independent" of the standards of individuals and cultures would typically require a degree of explication, and nonspecialists often struggle to understand precisely what respect moral facts are supposed to be "independent" even when given such explanations. But how *else* are we to ask participants? Versions of the disagreement paradigm don't appear to work, nor do other attempts to convey the notion of a stance-independence moral fact that is, e.g., irreducibly normative or that entails the existence of categorical reasons to act in accordance with whatever the putative facts may be easy to convey without jargon, in ordinary language, without a mountain of text to elaborate and clarify just what the claim is. A critic might suggest that this is overly pessimistic, and that *if* ordinary people have some preexisting notions of realist and antirealist conceptions of morality, then simple wording may be sufficient to convey what is being asked. This chapter will, I suspect, have gone a long way in dispelling such aspirations. Yet suppose people do have some implicit competence with these concepts: if so, the onus is on those making this claim to demonstrate as much; philosophers and psychologists are not entitled to presume sophisticated metaphysical theses are jangling around in humanity's collective unconscious without provide a *very* good account of why we should suppose this is the case.

Let's briefly consider a couple more examples of items I've constructed that are more face valid than most published scale items:

| **Cultural relativism** | Moral principles can only be true or false according to the moral standards of different cultures. |
|---|---|

| **Error theory** | All claims about actions being morally right or wrong are false. |
|---|---|

Once again, I have no confidence people would interpret these items as intended. People would likely interpret the item about cultural relativism in descriptive terms more often than they would in metaethical terms. I have no idea how people would interpret the item about error theory, but I find it incredibly implausible they'd understand error theory from an 11-word description, when it's difficult in practice to clearly convey what error theory's central claim is when you're given a few paragraphs to do so and don't have to be circumspect with what you're trying to convey for fear of demand effects. To be fair, I have not gathered data on any of these questions, but given the amount of data I've collected and analyzed, I can merely plead that I had to reach a stopping point eventually, and collecting nearly 6,000 responses is, I would hope, more than enough to make my case.

Nevertheless, the concerns raised here highlight an important tradeoff: the worse an item's face validity, the more low intended interpretation are best explained by faults with the item rather than the absence of determinate metaethical stances or commitments; conversely, the more face valid items are, the better low interpretation rates are explained as the result of metaethical indeterminacy. Thus, we are confronted with two methodological problems that are in at least some tension with one another: some stimuli ostensibly intended to measure folk metaethics are so poor that their failure to be interpreted as intended provides little or no information about whether ordinary people have metaethical views. Instead, such findings simply reveal that researchers failed to operationalize realism and antirealism appropriately. I have gone out of my way to argue that many studies have poor face validity or constitute inadequate operationalizations of folk metaethical views. Yet the worse these

studies are at operationalizing metaethical concepts, the less low interpretation rates of such items can be used as evidence for metaethical indeterminacy.

### S4.5.7 Untested paradigms

In addition, note that we might try to move away from short items or simple instructions, and employ more richly detailed instructions or stimuli, explaining metaethical concepts or asking an extended series of follow-up questions to assess comprehension, i.e., the methodological equivalent of "are you *sure*?" over and over until we're more confident their responses are telling us what we want to know. So long as this could be done in a way where the very act of investigating how people interpret questions doesn't alter how participants respond, such efforts could bear fruit. Yet I suspect such efforts will reveal the opposite, and serve only to further cement the far simpler explanation: that ordinary people are largely clueless about metaethical concepts. To the extent that clarification, elaboration, and training precede measures, such measures won't be able to tell us what we want to know: what *ordinary people's* metaethical stances and commitments are, since as I've argued elsewhere, we won't be able to tell whether such responses reflect the genuine metaethical stances and commitments of ordinary people, or are instead the product of spontaneous theorizing. If, on the other hand, elaboration, clarification, and training follow measures, and simply serve as detailed comprehension checks, my expectation is that we'll simply recapitulate the low rate of intended interpretation reported throughout this chapter. For instance, we could explain metaethical concepts such as cognitivism or realism to participants *after* they've been given the measures presented in a study, then ask them if these distinctions were reflected in their responses.

Pölzler and Wright (2020b) and Wright (2018) have recently adopted the former route, opting to explain metaethical concepts to people, train them in the relevant distinctions, then solicit responses. One shortcoming with the results reported here is that I have yet to collect data on and assess interpretation rates following more elaborate instructions or after engaging in training exercises.

If people readily take to such training, and their reactions are the equivalent of "oh, *that's* what you were asking. Yes, I endorsed realism all along…" this would serve as some evidence against my views. Conversely, if efforts to evaluate interpretation rates after training or more elaborate instruction revealed that people still struggled to interpret questions about metaethics as intended this would provide further support for indeterminacy (and, incidentally, further support that studies that don't employ training or more elaborate instructions are even more hopeless than I already propose that they are). Either way, such findings would not speak too much one way or the other towards metaethical indeterminacy, since the central problem with their methods is spontaneous theorizing; even if people did report having had such views all along, it would be difficult to distinguish genuine reports from confabulations. As such, training paradigms are unlikely to represent a serious threat to my hypotheses.

However, I have yet to assess interpretation rates for the paradigms Pölzler and Wright introduce, even in the absence of elaborate instructions or training exercises. This includes:

(a) *The theory task*

(b) *The metaphor task*

(c) *The comparison task*

(d) *The truth-aptness task*

Whatever concerns I have with the likely validity of these questions, I have yet to empirically assess how people interpret these tasks. Such tasks could, in principle, be presented without training exercises, extensive instructions, or excessive elaboration, after which questions about how participants interpreted these tasks could be used to assess interpretation rates. If so, results could reveal that most people interpret one or more of these paradigms as intended. I would take a substantial majority interpreting any one of these tasks as a potentially lethal blow to my case for metaethical indeterminacy. As such, a critical next step would be to evaluate interpretation rates for

these paradigms, along with any novel paradigms that sufficiently diverge from paradigms that have already been assessed.

Alternatively, we could conduct interviews or extended debriefings or follow-up questionnaires that go beyond merely asking for a short, written response. Perhaps such efforts will bear fruit. Yet the one instance I'm aware of where something along these lines was conducted supports my conclusion: David Moss (personal communication) conducted a set of semi-structured interviews on British residents designed to assess their metaethical views. The pool of interviewees was eclectic. Although most were students, one was a French student studying in the UK, and one was a PhD student in philosophy specializing in political theory (Moss, personal communication). David reports that participants struggled to interpret questions as intended and that they "all seemed confused by the questions" with the exception of the philosopher. David added that they:

> [...] displayed near constant levels of contradiction and backtracking. Failure to understand what was being asked was ubiquitous. But beyond that people's views were (by the standards of a philosopher) substantively exceptionally confused. It is definitely very tempting to say that most failed to express a clear and determinate metaethical stance. But definitely there were some respondents who seemed to lean very heavily towards some species of subjectivism/relativism, even if they weren't fully consistent or coherent. (David Moss, personal communication)

This is a bit more optimistic about determinacy than what I would expect, though it isn't insurmountable to the case for metaethical indeterminacy. First, recall that my position is that metaethical indeterminacy is the dominant account of folk metaethics, accompanied by a smattering of metaethical pluralism. What David reports here is largely consistent with this: *most* people may have indeterminate metaethical standards, while a handful may have determinate but pluralistic and potentially incoherent standards, consistent with e.g., Loeb (2008) and Colebrook (2021).

We may also question whether, and to what extent, the seemingly determinate relativistic standards are genuine and adequate reflections of a substantive *metaethical* position. "Student relativism" is a frequently-reported phenomenon (Paden, 1994; Pfister, 2019; Satris, 1986), and my

own students almost always expressed such views. Further discussion revealed considerable, consistent, and persistent misunderstandings about relativism and its implications: typically students conflated relativism with a variety of normative stances about how we can or should treat people from other cultures, even though this does not follow from a strictly metaethical conception of relativism (see Bush, 2016, Gowans, 2021).

It may be that most people lack determinate metaethical standards, but that something about the age, cultural background, or social context of being a college student heightens the degree to which people purport to endorse some form of relativism. More data would be needed to assess whether the handful of seemingly determinate relativists really did plausibly have determinate metaethical standards, or if their views could be best construed as a kind of vulgar relativism that was so intermingled with descriptive and normative claims that it would be hard to tell whether it adequately reflected relativism as a distinct metaethical position. That is, "relativism," in common parlance, may be understood to convey attitudes of tolerance or respect towards people and cultures with different moral standards, a sophisticated, cosmopolitan recognition that we're part of a global community with different histories and experiences, and so on.

Finally, at least some instances of interviewees expressing seemingly-determinate metaethical stances are often followed by remarks that at least hint to the contrary. Take this exchange:

> *David: If someone disagrees with you about whether that [donating to charity...] statement is true, is it possible for both of you to be correct or must at least one of you be mistaken?*
>
> *Participant: I think it's possible for both of us to be correct.*
>
> *David: Can you say why?*
>
> *Participant: No. I don't know why hypothetically they would hold a different view. As I said, I think that there's probably a lot of different definitions involved in different arguments and I don't know them all.*
>
> *David: Given that you say that both could be correct, what do you mean when you say that you agree with the statement?*

*Participant: It could be the case that it's sometimes true, sometimes morally good to give to charity but other times not. I guess some people also might argue that giving to charity for bad reasons or giving to poor charity is ... I mean, if the charity is to pay for rich people's mansions or something like that, then it might not in that case be moral for someone to give to those charities. I mean, for instance, it's definitely not moral for private schools to have charitable status. I think it's not so much to do with the abstract moral ambiguity but to deal with the statement being insufficiently precise.*

In this case, the participant's response appears to reveal the same tendency to interpret questions about metaethics in an unintended way: they say, "It could be the case that it's sometimes true, sometimes morally good to give to charity but other times not." Yet this captures the notion that the moral status of some action type, e.g., "giving to charity" depends on situational factors that can vary from one circumstance to another, so no categorical claim about the moral status of giving to charity can be made about the action type as a whole, e.g., that all instances of "giving to charity" are uniformly good or bad. What initially appears to be a "relativist" response, on examination, appears to be more consistent with an unintended interpretation. This participant went on to vacillate about their position, eventually offering what appeared to be a qualified realist response, changing their mind and judging that under the relevant degree of specification about the moral issue in question, stating "[...] I'm just going to say it's probably not the case that two people with different views could be right," but "with the caveat that I don't feel confident that I have knowledge of every possible moral case anyone has made about charitable giving." While this participant does not appear to have a clear perspective on moral realism or antirealism, they do appear to recognize their considerable confusion about the discussion, and convey this to the interviewer. This is useful, in that it highlights an instance of a person's initial response turning out, on examination, to conceal considerable confusion and uncertainty. The forced choice nature of metaethics paradigms once again rears its head: people can superficially appear to have a metaethical view when in fact they have little understanding of what they're responding to. Note, as well, that we shouldn't expect everyone to be so candid: people may wish to conceal their confusion or ignorance by *pretending* to understand, or even dupe themselves into

thinking they do understand, further obscuring attempts to study how people think about philosophical issues. In short, people *may* have determinate metaethical views, but then again, they might not: what appear to be instances of expressing a determinate view might, once we dig a little deeper, turn out to be little more than ephemeral glimmers of apparent comprehension, a kind of pseudo-philosophical pantomiming of genuine understanding because the language the participant mirrors the language used in metaethics; this is not surprising, either, given that many of the terms used in metaethics are simply repurposed versions of familiar terms, e.g., "objective," and it is even less surprising when the very act of speaking to participants about these topics often involves furnishing participants with enough terms and concepts they can echo them back to the researcher. Any of us who have lectured or graded papers will be familiar with a student attempting to cobble together the semblance of a coherent response by borrowing the terms and concepts used in the question or discussed in the class, stringing them together in a barely-concealed attempt to earn a point or two, a gamble that often pays off due to inattentive grading. But on those occasions when we've downed enough coffee to recall past lives, most of us have likely spotted a student trying to slip this kind of nonsense past us; it should come as no surprise that someone would do the same when asked an open response question in a survey, or when they feel a moment of embarrassment in an interview. Countless responses coded as "clear intended" interpretations consisted of little more than the participant echoing back some portion of the items or stimuli, or just saying "it means morality is subjective" without elaboration. Do these people *truly* have understand what researchers intend to ask? And when they use such terms, are they referring to the positions those terms refer to in academic philosophy? I doubt it, but decisively demonstrating the superficiality of such responses would require precisely the kind of painstaking efforts David Moss went through to personally interview people, a task that requires time, dedication, the relevant training and experience, philosophical expertise, social deftness, and the intellectual deftness to combine all of this to effectively navigate an unstructured or

semi-structured interviews well enough to extract informative responses from participants. This simply isn't a task that can be handed off to a research assistant, which unfortunately puts rather severe constraints on how much of this kind of data we can collect, without major institutional redesign that involves teaching philosophers to adopt some of the methods employed by psychologists, anthropologists, and so on.

### S4.5.8 Additional discussion and limitations for specific studies

*S4.5.8.1 Study 3: Additional discussion and limitations*

There are a handful of shortcomings with this Study 3 that I didn't mention in the main text. First, the particular wording used in these studies was not used in any published research. Thus, low rates of intended interpretations and high rates of unintended interpretations do not *directly* challenge the validity of any particular set of measures. Rather, such findings only indirectly support the hypothesis that participants *generally* struggle to interpret questions about metaethics.

Second, the noncognitivism condition had poor face validity, and thus the low rate of clear intended interpretations (just 0.5%, $n = 2$) is not a strong indication of a widespread failure to interpret expressions of noncognitivism as intended. Consider the wording: "there is no fact of the matter." Noncognitivism holds that moral claims are not propositional, and thus cannot be true or false. It is not obvious that this is clearly conveyed by saying that, with respect to which charities do the "most good," that this should be understood to mean that claims about what charities do the most good cannot be true or false. While it may have seemed to myself and my colleagues that to say that there's "no fact of the matter" about some truth claim, that this means that considerations about which charities do more good cannot be true or false, in practice, this *could* be interpreted as the claim that there is no *single* or *stance-independent* fact of the matter, which would be consistent with relativism. For instance, suppose someone said:

> *"There is no fact of the matter about which flavor of ice cream is best."*

Does this entail that claims about which flavor of ice cream is best cannot be true or false? Not necessarily. "Fact of the matter," could *just mean* a single, universal, or stance-independent fact, or perhaps even a non-relative fact. If so, someone could think that someone who said such a thing could still believe there are facts about what is good or bad relative to the standards or preferences of people or cultures. In other words, while I might think there is no "fact of the matter" about whether chocolate or vanilla is the best flavor, I could *also* think that when you say that "chocolate is best" and I say "vanilla is best," that we are both correct relative to our preferences; it's just that the facts we are correct about aren't subsumed by the phrase "fact of the matter." However, strictly speaking this is *not* what it means for there to be no fact of the matter. Yet discovering that many people do not interpret this phrase in line with noncognitivism, but that at least some do interpret it in line with relativism/subjectivism, this may demonstrate little more than that participants fail to reliably interpret a particular turn of phrase as intended. By itself, this arguably does not count for much.

Finally, all conditions suffer from a shared limitation. Previous research suggests that when participants are asked to classify issues as moral or nonmoral, most participants do not judge donating to charity to be a moral issue. For instance, Wright, Grandjean, and McWhite (2013) found then when participants were asked to classify charitable giving as a moral, social, or personal issue, that only 11% classified donating to charity as a moral issue (the remaining 89% all classified donating to charity as a personal issue; p. 5). Findings like this suggest that people may not view donating to charity to be a moral issue. The risk that participants may not have interpreted these questions to concern morality is further compounded by the questions not making any explicit reference to morality. That is, all three items simply referred to charities doing "more good," or the "most good," without explicitly stating more *morally* good or the most *morally* good. It's not entirely clear *why* people don't regard donating to charity as a moral issue, or what they have in mind when they classify it as a "personal" rather than moral issue (after all, it's not obvious that personal issues are mutually exclusive with moral issues).

Regardless of how participants are thinking about donating to charity, it is possible that moral considerations in particular were not salient, and that as a result their reactions to these questions failed to prompt thinking in metaethical terms, not because ordinary people don't understand questions about metaethics, but because they didn't interpret this particular set of questions to be about metaethics because they didn't view it to be a question about morality at all. Future studies should focus on using explicit moral language and by carefully reflecting on stimuli to check which issues participants regard as moral, compare such responses to issues participants regard as nonmoral. Researchers could also use pretested pools of items previous participants considered moral issues, and to focus on using a variety of different stimuli to capture a broader spectrum of issues in and outside the moral domain.

*S4.5.8.2 Study 4: Additional discussion and limitations*

In the case of concrete moral issues, participants were likely to express normative attitudes about the moral issue in question. This is consistent with concerns that questions about metaethics may often fail to prompt metaethical thinking because normative moral considerations are more salient, and motivate participants to respond to such normative considerations instead. Put yourself in the shoes of such a person. Suppose someone asked you:

>*"Do you think Hitler's holocaust was objectively morally bad?"*

You may find that you have to suppress the disgust, horror, and moral outrage you feel at any mention of Hitler or the holocaust, which can interfere with the ability to respond to what this question is actually asking about: whether there are stance-independent moral facts, or at least, whether there is a stance-independent moral fact in this particular case. It is difficult to say "no," to this, because it can *feel like* you're not merely denying that moral facts can be stance-independent, but that you are also claiming that you don't have a strong *normative* opposition to what Hitler did. Such questions, when asked in the context of debate, are a kind of rhetorical trick: normative considerations are often

embedded inside questions about metaethics. An invitation to respond to the metaethical aspects of the question is presented alongside some awful moral atrocity. If you respond to the metaethical aspect of the question without responding to the normative elements, this can conversationally imply that you have a lax, ambivalent, or even positive attitude towards the atrocity in question. This is *normative entanglement* in action.

Ordinary people are not philosophical robots. They have reputations, and it's important to maintain those reputations. Even if one asks a dry, boring intellectual question that has a dry, boring answer, if you happen to toss in some remarks about Hitler and genocide, a reasonable respondent will realize that their reputation is on the line, and respond by condemning Hitler and the holocaust. This isn't a mistake. It's a completely reasonable move to make, given the social goals people typically have. It would be reputational suicide for anyone to seriously suspect sympathies with Hitler or the Nazis, so any situation in which Hitler and Nazis are mentioned, it's important to go out of one's way to distance oneself from these things by condemning them. This is because people are not simply attempting to state facts. They are managing background assumptions about who they are and what they stand for. And people can make inferences about who you are and what you stand for based as much on what you *don't* say as on what you *do* say. It is no great stretch to imagine that people are competent at navigating social spaces in such a way that they internalize a reflexive tendency to react to situations where social and reputational concerns are at threat by responding in ways that effectively signal positive character traits (or the absence of bad traits). In short, what we may be seeing in many studies about folk metaethics is the same *normative entanglement*, but simply on a smaller scale than in the case of Hitler and Nazis. It's *obvious* why, if someone asks you if what Hitler did is "objectively bad," that it's useful to make a point of condemning Hitler. It's less obvious when it comes to murder or abortion, but doing so still serves an important function.

The upshot is that normative entanglement may induce participants to become distracted by more salient reputational considerations and consequently focus more on ensuring anyone who might see their responses that they are against murder, or pro-choice, or pro-life, or whatever, while the metaethical considerations such questions are actually about take a back seat. If this is occurring, it does not mean ordinary people don't have metaethical stances or commitments. They could have them, but emotionally charged questions are inappropriate means of probing metaethical views. On the other hand, if people did have metaethical stances and commitments, we'd expect to be able to elicit them. In any instance in which straightforward attempts to do so appear to catastrophically fail, this is at least *some* evidence in support of metaethical indeterminacy.

Overall, the findings in these suggest that asking participants directly about whether morality is "objective" or "relative," or whether the truth of specific moral claims is "objective" or "relative," isn't a valid way to measure metaethical beliefs. This casts serious doubt on the claims made by Fisher et al. (2017). Aside from the significant theoretical errors that caused them to inappropriately claim that the measures in their first two studies were valid because a third study using a different measure yielded similar results, we now have direct evidence that the measures used in their third study were probably *not* valid. If we may employ the same flawed reasoning they did, this would, by implication, suggest that their first two studies weren't valid, either. Fortunately for Fisher et al. (2017), I wouldn't make such a claim because I don't think it would follow. I don't know if their first two studies were valid or not. But neither do they.

There are some limitations with these findings. First, the precise wording used in these questions does not mirror the wording used in any prior research. It is possible that the low rate of intended interpretations is due to the specific features of the questions I asked, and not because people are especially unlikely to interpret the terms "objective" and "relative" in moral contexts in unintended ways. This is less plausible given the variety of mutually corroborating analyses of interpretation rates

presented here, however. Taken in isolation, it might seem plausible that perhaps people have sufficiently well-developed metaethical standards that they can respond appropriately to questions that directly ask whether morality is "objective" or "relative." But given that across a variety of paradigms and prompts participants reliably underperform, the most likely explanation is simply that people do not interpret "objective" and "relative" as intended even when they are presented in a clear and appropriate way (from the perspective of philosophers and researchers).

Another, more subtle limitation is that people may be familiar with moral realism (or moral "objectivism," understood to mean the same thing), and moral relativism even if they fail to respond to these questions appropriately. These findings suggest that people tend not to interpret "objective" and "relative" in the way these terms are typically used in academic metaethics. Yet this does not show that they lack the *concepts* of objectivism (or realism) and relativism; only that such conceptual competence was not prompted by these particular studies. While the best explanation for this is that they are not competent with these concepts and may lack them, it is possible they do have them, and are competent with them. It is outrageously difficult to show, one way or the other, what is actually going on here. For instance, we might think people are competent with the concepts of *realism* (or "objectivism") and *relativism* precisely *because* they use the terms "objective" and "relative" in moral contexts. Yet this does not follow. People who employ these terms often express additional comments that indicate that they do not understand these terms the way philosophers do (see **Chapter 3**). Thus, even if a person responds to a question about relativism by saying "it means morality is relative," this does not entail that they mean the same thing that philosophers do, or that questions intended to measure their views about *relativism* are valid. Simply using the same *terms* as researchers does not entail that they are employing the same *concepts*.[192]

---

[192] One omission worth noting is that I only asked about morality being "objective" and "relative," but not "subjective." This is an oversight that future studies should correct for, and I mention this here as much as a personal reminder as I do

However, there are a number of limitations with these results. The most obvious limitation is that the MRS only assesses folk relativism. As such, it tells us nothing about interpretation rates for realism or other metaethical concepts. Another shortcoming with this study is that there are too few responses per item. There were only 14-17 responses per item, which is too small to make accurate estimates about the interpretation rates for specific items. This is the most likely explanation for high variation in clear intended interpretation rates across items, which varied from as low as 5.9% (Item #1, condition 5B), to 42.9% (item #3, Condition 5A, item #6, condition 5B). With so few responses per condition it's possible that a majority of participants would reliably interpret some items in a clear intended way. Some items had a high clear intended interpretation rate (item #3 and item #6). Note, however, that if the true proportion of participants who interpret these items is equal to or somewhat greater than half, this would still mean that a substantial number of participants do not interpret these items as intended, and it would still suggest that these are extremely poor measures of folk metaethics.

Nevertheless, it would suggest that many people *do* have determinate metaethical stances, or are at least capable of understanding questions about metaethics as intended. And at first glance, this does appear to be the case for at least *some* people. However, metaethical indeterminacy with respect to folk metaethics is simply the view that the folk conception of morality does not entail any determinate metaethical commitments with respect to realism or antirealism. It does not entail that *no* people who participate in studies have determinate metaethical commitments. Some do, but these will typically be lay people who have engaged with a sufficient amount of philosophical reflection or education to have a view on the matter. In much the same way some people who are not theoretical physicists have views on physics, or people who are not evolutionary biologists read popular science

---

to document limitations of the study. Future studies could also explore mind-independence and other concepts more directly.

and develop views on e.g., arguments about the unit of selection, without this entailing that the way ordinary people speak and think commits them to views on quantum physics and units of selection. Thus, metaethical indeterminacy does not entail that *no* participants will interpret questions as intended, or be able to clearly articulate a metaethical stance towards the issue, just that this number will be low. 50% or more is not low, however. While it is consistent with the claim that measures of folk metaethics have poor validity that there are a few items that a slim majority of participants interpret as intended, it is not consistent with metaethical indeterminacy that most participants actually have determinate metaethical stances towards these issues. So, how can metaethical indeterminacy be reconciled with a greater-than-half rate of clear intended interpretations?

First, there's a little wiggle room due to metaethical indeterminacy not requiring that the clear intended interpretation rate be zero. Although this must be quite rare if metaethical indeterminacy accounts for poor interpretation rates, it is plausible that at least *some* have a reasonable idea of what they are being asked. However, this doesn't provide that much wiggle room, so it won't be adequate.

Second, *spontaneous theorizing* may account for some proportion of clear intended interpretations. This would involve participants coming to a fixed position in virtue of participation in the study. In such instances, we will not know whether those participants held such views prior to participation. However, given the sparsity of context and details, this is not a plausible explanation for the high rate of clear intended interpretations for scale items when no instructions or training are provided.

These two considerations do little to account for the potential of a high rate of clear intended interpretation. For that matter, even a modest rate below 50% would be troubling for metaethical indeterminacy. There are two ways of addressing this response that I outline briefly here, but address in greater detail in the general discussion. First, participants may exhibit transient stances that are

expressed in the context of the study, and may even express such stances in other contexts, without such expressions reflecting a stable and substantive stance or commitment in most general contexts.

Third, the proportion of clear intended interpretations may be a substantial *overestimate* of the proportion of people who have an actual determinate stance or commitment about the metaethical issue in question. On the one hand, it could be that the method used here dramatically *underestimates* the true proportion of people who interpret any given question about metaethics as intended. After all, a substantial proportion of participants offer *unclear* interpretations. Many participants may interpret questions about metaethics as intended, but are unable or unwilling to convey this when presented with open response questions. On the other hand, the standard used to assess instances of clear intended interpretations is, I believe, so low that I have counted as clear instances numerous responses that, were I to follow up with the person expressing these views, would reveal that their understanding of the relevant metaethical concepts is at best rudimentary, and in many cases outright confused or mistaken. I address these last two concerns more thoroughly in the general discussion and the conclusion, where I take up the challenge of accounting for why metaethical indeterminacy is not only plausible, but the most likely account of the way ordinary people think (or, more accurately, *don't* think) about metaethical issues. Such a claim may seem to be on shaky footing if we can identify instances in which a substantial number of participants appear to interpret questions about metaethics as intended. However, I believe such apparent comprehension of metaethical concepts is superficial and rudimentary, that there are considerable theoretical grounds for suspecting this to be the case, and that future studies will reveal this to be the case.

## S4.6 Future directions

I have described a variety of limitations with the present findings, and have alluded to a few ways future studies could address these limitations. Here, I will summarize some of those suggestions and add a few more thoughts on the future of research on folk metaethics that employs the methods used

here, the future of folk metaethical research in general, and the use of qualitative research to assess the validity of research as it's been employed here.

The most critical step in expanding on this research is to acquire additional coders, with an eye towards adversarial coders. If even adversarial coders report low rates of clear intended interpretations, this would be especially strong evidence for the claim that most people do not interpret questions about metaethics as intended. It would also go a long way in mitigating my personal biases, errors, and idiosyncrasies. It would not surprise me if researchers reach quite different conclusions, though how to navigate such differences will be a bridge best crossed if we ever get there.

It is also essential to assess the generalizability of my findings by drawing on new participant pools. Ideally, this would involve cross-cultural research conducted in the native languages of a culturally and linguistically diverse body of participants, as well as greater efforts to reach demographically diverse populations. Of special note is the possibility that results would vary among religious subcommunities, whose moral views may differ substantially from surrounding populations. Acquiring demographic variables would also allow us to assess whether there are systematic differences across gender, socioeconomic status, age, education level, and other characteristics. All of these endeavors will pose considerable challenges. It is incredibly difficult to locate coders with the requisite expertise to properly code items of this kind, i.e., coders who are experts in metaethics and who also have training in the social sciences. Even if experts could be found, it'd be difficult to motivate them to code vast amounts of open response data. It is a laborious task that requires considerable dedication. And collecting enough responses to identify meaningful differences *within* a sample (e.g., by age group) would require even larger samples than the ones I've collected, making the task even more onerous.

Future efforts could also be made towards improving data quality. Providing greater incentives to provide more detailed or thoughtful responses by, e.g., offering greater compensation, requiring

participants to spend more time on questions, providing more detailed instructions, having participants engage in training exercises, or changing the context of the study to one where there are fewer incentives to move through stimuli quickly may go some way in improving interpretation rates, and if so, this would suggest that at least one reason why interpretation rates were consistently low was due to low effort or attention, not the absence of a determinate metaethical stance or commitment.

Another way to expand on existing results would be to assess interpretation rates for paradigms that have yet to be investigated, e.g., the novel methods introduced by Wright (2018) and Pölzler and Wright (2020a; 2020b). In addition, constructing and assessing the clearest descriptions of metaethical positions *I* can come up with, and exploring interpretation rates for those, may prove especially insightful. If I were to present items that I found to be more face valid representations of the relevant metaethical positions, and people still didn't interpret them as intended, this would bolster the case for indeterminacy. Of course, I would be coding responses to my own items, and relying on my own judgment about the quality of those items, so there'd still be a degree of bias involved. One alternative would be to assess interpretation rates for items judged by adversarial experts to be ideal representations of a given metaethical position.

There are a handful of other avenues future research could explore. One would be to assess interpretation rates for the same questions among academic philosophers. We should expect clear intended interpretation rates to be higher among philosophers, and especially higher among those with a background in metaethics. If interpretation rates remain low, this would raise questions about the adequacy of the method employed here. If people who we should expect to interpret questions as intended still don't do so, this may be indicative of either problems with the questions, or problems with my method of analysis, rather than evidence that philosophers are confused or have no determinate metaethical stances themselves.

Given the introduction of training paradigms, there could also be some value in assessing interpretation rates following more elaborate stimuli, in response to more detailed questions, or following exercises intended to familiarize participants with the relevant metaethical concepts. The results of such findings would, however, be limited. While we might see an increase in intended interpretations, and indeed, it would be shocking if we didn't see *some* increase, this may achieve little more than demonstrating the efficacy of the training methods. Yet due to spontaneous theorizing, whatever responses participants gave following such instruction wouldn't provide reliable insights into what metaethical stances or commitments they had (if any) *prior* to engaging with the instructions or training. Nevertheless, there would be at least some value in assessing the effectiveness of these methods in instilling greater competency with metaethical concepts. My suspicion is that it would be difficult to achieve more than modest improvements in understanding metaethical concepts. There's a reason people study these topics for a living. An adequate grasp is not something one can expect to obtain overnight, much less in the minute or two one is introduced to the concepts in the midst of a survey they may have little interest in.

Yet another way to expand on the current research would be to explore nonmoral domains. Investigating how people interpret normative questions regarding epistemic and aesthetic norms, as well as social conventions and scientific claims, may prove insightful.

However, the present method of analysis isn't the only way to evaluate interpretation rates or explore the possibility of metaethical indeterminacy. Future research could employ a variety of other methods, including interviews (Andow, 2016; Moss, 2017; Nadelhoffer & Nahmias, 2007), asking people directly whether they had such views prior to being told about them (an admittedly weak and potentially uninformative line of questioning), assessing interpretation rates using conventional social scientific methods that employ quantitative measures, e.g., Likert scales, checkboxes, or multiple choice (an approach I take up in the next chapter), or adapting methods or insights from anthropology,

linguistics, and other fields, though I wouldn't be in the best position to know how to best employ such methods. Researchers interested in the way people actually talk about morality might also benefit from big data approaches that involve e.g., examining how people actually talk about morality in the real world on social media sites such as Facebook or Twitter. Such approaches have the advantages of both gathering data on how people use moral terms in the real world and gathering massive amounts of data that would be otherwise unavailable to researchers. Such data could be examined for the frequency with which metaethical terms and concepts are brought up, and the degree to which such reveals insights or interesting patterns that could shed light on whether ordinary people discuss (and thus think about) metaethics in their everyday lives.

Ideally, researchers will employ all of these methods. The most compelling case for a given theory is one for which a wide array of mutually corroborating lines of evidence can be marshaled, all pointing towards the same conclusion. As Rozin (2001) puts it:

> One can reasonably look only for evidence in single research ventures, not proof. Indeed, the best hope we may have (as worked so successfully in the validation of the theory of evolution by natural selection and in most historical and archeological studies) is to accumulate flawed (ambiguous) evidence in large amounts and from many different sources and approaches. This is probably the only practical route to understanding Homo sapiens in a social context. (p. 3)

Just the same, the best way to get a handle on whether or not, and to what extent, ordinary people hold determinate metaethical stances or commitments would involve a variety of approaches, each pointing towards the same conclusion. What I have presented here is merely one modest piece in a larger puzzle.

However, the *order* in which these methods are employed. Previously, I alluded to Moss's use of interviews for exploring how people think about metaethics. While I believe this is a promising method, exclusive dependence on qualitative methods lacks the rigor and other theoretic virtues of quantitative research, and is thus insufficient, on its own, to provide sufficient raw data to build an adequate theory of folk metaethics. Nevertheless, the kinds of data we obtain via qualitative analysis,

and more generally through the collection of descriptive data, is, I believe, the best route to take. I believe researchers studying folk metaethics fell victim to a general shortcoming in psychological research: a tendency to prematurely conduct experiments and to rely too much on top-down, theoretically top-heavy and presumptuous notions of the way people think and act, without having first gathered sufficient raw data, and spent the time pouring over it, to develop well-informed hypotheses worthy of subjecting to experimentation. This concern echoes sentiments expressed by Rozin (2001) at the turn of the century. Rozin, having the advantage of experience and having lived through much of the recent development of the field of psychology, is in a far better position to cast judgment on the state of psychological research. Rozin sets the stage with a quote from Asch that effectively summarizes his thesis:

> In their anxiety to be scientific, students of psychology have often imitated the latest forms of sciences with a long history, while ignoring the steps these sciences took when they were young. They have, for example, striven to emulate the quantitative exactness of natural sciences without asking whether their own subject matter is always ripe for such treatment, failing to realize that one does not advance time by moving the hands of the clock. Because physicists cannot speak with stars or electric currents, psychologists have often been hesitant to speak to their human participants. (Asch, as quoted in Rozin, 2001, p. 2)

Rozin concurs. Commenting on the history of social psychology, Rozin states:

> I believe that social psychology, modeling itself in the mid-20th century primarily on the natural sciences and on sensory psychology, has concentrated on the advancement of a formal, precise, and experimental science. However, unlike the successful work in the natural sciences and sensory psychology, the work in social psychology has not been preceded by an extensive examination and collection of relevant phenomena and the description of universal or contingent invariances. In the more advanced sciences that social psychology would like to emulate, there is much more emphasis on phenomena and "description" than there is in social psychology, and there is less reliance on experiment. Such sciences, particularly the life sciences, also pay less attention to models and hypotheses and more attention to evidence as opposed to proof or "definitive" studies. (p. 3)

In other words, social psychology prematurely leaped into the scientific deep end, conducting experiments before engaging in the preliminary work of cataloging and describing the phenomena of

interest by collecting massive amounts of observational data. Drawing on Rozin's insights about social psychology, I believe the same could be said, with *even more* force, about the state of research on folk philosophy: while social psychologists sought to emulate the natural sciences, much of the research on folk philosophy falls within the ambit of "experimental philosophy", which has sought to emulate the social sciences (Alexander, 2012; Horvath, 2012). In doing so, research on folk philosophy is an imitation of an imitation, compounding the methodological shortcomings of the social sciences by importing a host of theoretical assumptions from analytic philosophy.

Social psychology at least has the advantage of a discipline more adjacent to the sciences, with decades to mature and a body of researchers trained in the relevant statistical and analytic tools. Yet just as psychology was beginning to stand on its own, philosophy had taken the *opposite* route: as natural philosophy gave way to modern science, philosophers became ever more insular, ever more concerned with the *a priori*, ever more removed from the natural sciences as one after another discipline left it behind, like a dozen sisters of Athena emerging from Zeus's head and never looking back. Philosophy, bitter at its ungrateful children, turned away from them, and after nearly a century of plodding along, eyes downcast, it has only finally started to gaze, over the past two decades, at an academic landscape radically altered in its absence. To be sure, philosophy cast a few furtive glances upwards through the years. But it's taken the deliverances of Quine and Dennett and Stich a long time to rekindle the empirical spirit that spark that eventually culminated in the experimental philosophy movement, which emerged only at the dawn of the 21st century in, among other works, Weinberg, Nichols, and Stich (2001) and Knobe (2003).

Yet having turned their attention back to the empirical sciences, experimental philosophers saw the social sciences conducting experiments, and simply jumped on board, doing the same. This was a mistake. Experimental philosophy imported the paradigms, distinctions, and theories from philosophy wholesale, without considering whether the concepts and distinctions of interest to

philosophers were reflected in the psychological constructs at play in everyday thought and discourse. Little attention has been given to gathering bottom-up, descriptive work without making assumptions about the philosophical views that may be present in ordinary thought.

Rozin (2001) uses the biological sciences as a model for how work in the natural sciences differs from the social sciences. Examining the contents of a handful of prominent journals, Rozin found that, unlike social psychology

(1) Most studies do not explicitly appeal to particular model or hypothesis as the rationale for conducting the study (p. 8)

(2) Only about half of the articles carried out experiments

(3) Most don't employ statistical tests

(4) Many focus on description rather than analysis. As Rozin notes: "Many of the studies in molecular biology journals are what psychologists would characterize as descriptive: The elaboration of structures, or what happens between time one and time two, are typically illustrated by photographs." (p. 9)

(5) Studies are typically motivated by a desire to determine the relationship between two or more phenomena, rather than two test a hypothesis or the strength or direction of an effect

(6) Replication of an effect found in one species in another species is an acceptable rationale for publication in top journals, in contrast to social psychology, where Rozin observes that "replication on a different group of participants [...] might be publishable, but generally not in a premier journal" (p. 9).

In short, prestigious, high-status research in the biological sciences was primarily concerned with gathering descriptive and observational data, not merely testing theories and hypotheses. And this remained so for a far more mature science. Psychology, on the other hand, has focused far more on the testing of theories and conducting experiments. This isn't to say that such research is without value. Far from it, as Rozin acknowledges, "There is no question that the experiment is the most powerful tool available to the sciences" (p. 9). But good theory and good experimentation only emerge against the backdrop of a mature science with a well-developed body of observational and descriptive

data. Researchers need enough grist for their mill. Unfortunately, social psychologists seem to have been operating under the pretense that they could bake bread without enough flour. Experimental philosophers made the even greater mistake of thinking they can bake bread out of whatever they found in their cupboards, without bothering to check if it's edible. In other words, whereas psychologists have failed to gather adequate descriptive data, philosophers have operated under the presumption that their own ways of thinking can be neatly converted into a psychological construct without wondering what psychological constructs might already be present in ordinary thought.

These concerns, if correct, offer a bleak picture not only for the study of folk metaethics, but for the study of folk philosophy in general. My concern with virtually all research attempting to determine whether ordinary people are moral realists or moral antirealists is that the very conception of "moral realism" and "moral antirealism" may very well be the artificial constructions of philosophers. Such positions may be too esoteric and too far removed from the business of everyday moral thought to be adequately reflected in any systematic way in the words and deeds of the men and women aboard the Clapham omnibus. The same may extend to other areas of folk philosophy as well. How off track are studies on free will, or the ought-implies-can principle, or personal identity, or the nature of consciousness, or the concept of knowledge?

I don't know. Perhaps researchers that focus on these fields would have greater insights into their prospects for vindication from the concerns raised here. I'm not confident. I suspect we'll find that many of the apparent patterns identified in the past few decades were illusory. However replicable a given finding may be, what researchers have taken to be individual differences in the psychological phenomena of interest, such as distinct philosophical intuitions, will turn out to be mundane variation in how different people resolve ambiguities, how sensitive they were to the unintended pragmatic differences that introduced unintentional confounds into the different conditions of a study, and so on. Indeed, such outcomes are already beginning to emerge.

Recently, Thompson (2022) found that qualitative analysis of research on the "*ought implies can*" principle, which purportedly shows that people rejected the principle, revealed that people's "judgments are misrepresented by quantitative survey questions," and that "the majority of participants uphold or preserve 'ought implies can'" (p. 1). Thompson concludes that experimental philosophers may be able to "more accurately capture judgments by using qualitative methods," and that "studies which rely on quantitative surveys possibly misrepresent participants' judgments" (p. 1). While not demonstrating indeterminacy, Thompson's results illustrate how conventional social scientific methods can fail to capture the nuance and complexity of folk thought; by forcing what Thompson describes as the "complex and multifarious" way that people thought about the ought implies can principle through the strictures of quantitative analysis, researcher may not only have failed to reach the correct conclusion, but mistakenly took themselves to find evidence of the *opposite* of the truth.

A related problem has continued to plague the most iconic finding in the field, Knobe's eponymous *Knobe effect* (Knobe, 2003). Although previous efforts failed (e.g., Adams & Steadman, 2004a; 2004b; 2007; Nichols & Ulatowski, 2007), Lindauer and Southwood (2021) have purportedly demonstrated that the effect results from unintended pragmatic implicature that has worked its way into the design, serving as a confound that can explain away the purported phenomenon as an artifact of experimental design. By purportedly demonstrating to cancel the subtle pragmatic differences between conditions, Lindauer and Southward claim to have demonstrated that the entire effect may be attributed to differences in how participants interpret conditions. Far from illustrating a novel psychological phenomenon, the Knobe effect, on their view, represents little more than an inferential mirage.

However, their pronouncement may be premature, as Sytsma, Bishop, and Schwenkler (2022) argue that their results are *themselves* an experimental artifact. It remains to be seen how this conflict

will resolve. Amusingly, it's a win-win for the critiques outlined here. On the one hand, if the Knobe effect does turn out to be a dud, it will provide more fuel for my contention that a great idea of research on folk philosophy is inattentive to ambiguity and the subtleties in how participants interpret questions.

If, on the other hand, the Knobe effect is ultimately vindicated, we will be left with the identification of a tendency for ordinary people to think in ways that run contrary to the philosophical principles philosophers tend to endorse. Indeed, the whole reason the Knobe effect gained so much traction is *because* it purported to identify a feature of ordinary thought that deviated from philosophical ideals: The Knobe effect purportedly shows that people's tendency to ascribe intentionality to agents depends on antecedent judgments about the normative status of the agent's actions: a person might consider a person morally responsible if the action is *bad*, but not if the action is *good* or *neutral*. This violates philosophical orthodoxy, which maintains that the intentionality of an action doesn't depend on one's normative stance towards the action in question. If such an effect is real, and does characterize ordinary attributions of intentionality, it would provide strong evidence that philosophical norms deviate from ordinary thought, and that philosophers themselves, insofar as they *don't* exhibit the same psychological dispositions as ordinary people, have either self-selected or developed idiosyncratic ways of thinking. This is hardly good news for philosophers, since if one of the very first findings we stumble into in the field of folk philosophy is the discovery that philosophers don't think the way ordinary people do, philosophers are hardly in a position to presume ordinary people would think the way they do with respect to *other* philosophical issues.

My point stands either way. Either the methods philosophers use haven't been adequate, or the assumptions that the concepts and distinctions that characterize philosophical thought, and the norms that govern that thought (e.g., logical consistency and the like) may turn out *not* to be as relevant to folk philosophical thought. Most importantly, *either* resolution points to the need for more

descriptive and bottom-up work. Researchers interested in the way ordinary people think about traditionally philosophical problems ought to take a page out of the anthropologist's book, set aside their assumptions and worldview, and go out there and listen, observe, and interact with ordinary people. The recent history of philosophy and much of psychology have been married by a reluctance to get our hands dirty. Philosophers and psychologists have spent far too long cooped up in Plato's cave or crunching survey results from the other side of a computer screen, far removed from the happenings of the world outside.

Rather than relying so heavily on online surveys and polling undergraduates, the future of folk metaethics may be best served by leaving the college campus and the comforts of our laptops. The ordinary life of ordinary people - what they say, think, and do - occurs *outside* the gilded cage of the college campus. Out there, amidst the birds and trees, their thoughts and actions are there to be observed, described, and cataloged in their natural environment. Experimental philosophers are fond of calling on philosophers to get out of their chairs and go conduct studies. Ironically, experimental philosophers may have succeeded in leaving the armchair in the office, only to settle into a new armchair in the lab.

My previous recommendations have concerned the future of research on folk metaethics. However, the method I've developed and implemented here could be used to assess interpretation rates for other paradigms. I can offer little guidance as to which specific areas of research would be the most viable candidates. My own knowledge is confined primarily to moral psychology, and as such any recommendations I make would be heavily skewed merely by research I'm familiar with, regardless of whether it serves as the best candidate. I suspect that those who find merit in the approach taken here, who are also familiar with other lines of research, would be in the best position to assess the prospects of applying this method to that body of research. However, I am reasonably confident that

many of the best candidates would be drawn from folk philosophy or, though I have eschewed using the term, research in experimental philosophy.

**S4.7 Thematic analysis**

*S4.7.1 Coding for thematic analysis*

My approach to thematic analysis roughly follows the guidelines proposed by Braun and Clarke (2006; 2014; 2019; 2020; Clarke & Braun, 2013) for what is now known as *reflexive thematic analysis* (RTA). I say *roughly* because Braun and Clarke have a fairly distinct conception in mind about the methods they propose and the underlying theoretical assumptions that ground their approach (see e.g., Braun & Clarke, 2019). I reference their method only as a touchstone for how I have approached coding data; I am *not* claiming to have conducted thematic analysis in a way they would acknowledge as an instance of RTA. The primary overlap is in a rough adherence to the phases they employ, and especially their fluid approach to coding that emphasizes the iterative and recursive nature of the phases involved in the process.[193]

Rather than rigidly moving from one step to the next, I actively sought to move between the phases, reflecting on a given dataset then going back over and recoding where necessary, including generating new themes or eliminating old ones. While there is no ironclad rule for when a given dataset achieved "perfect" coding, this isn't the goal. At a certain point, diminishing returns kick in and there is little value in recoding the data further. Rather, the point is to approximate as best one can a coherent and defensible analysis of the data that takes proper stock of a given set of responses. Braun and Clarke liken this approach to editing a paper. A paper written from beginning to end could benefit from careful review and editing, and a good writer will often reflect on the paper as a whole before making substantive changes to its structure and content. Yet there is no bright line dictating an official

---

[193] As they put it, RTA involves a "recursive process, where movement is back and forth as needed, throughout the process" rather than "a linear process of simply moving from one phase to the next" (Braun & Clarke, 2006, p. 86).

stopping point, nor any point at which a paper ever achieves perfection. An author simply has to decide when to move on. This is one of the limitations that accompanies a qualitative approach; I had to rely on my own discretion to judge when further refinement would yield little value. The phases outlined by Braun and Clarke (2020; Doing Reflexive TA, n.d.) in **Table S4.2**.[194]

**Table S4.2**

*Phases of reflexive thematic analysis (RTA) (adapted from Doing Reflexive TA, n.d.)*

| Step | Description |
| --- | --- |
| 1. Familiarization with the data | This phase involves reading and re-reading the data, to become immersed and intimately familiar with its content. |
| 2. Coding | This phase involves generating succinct labels (codes!) that identify important features of the data that might be relevant to answering the research question. It involves coding the entire dataset, and after that, collating all the codes and all relevant data extracts, together for later stages of analysis. |
| 3. Generating initial themes | This phase involves examining the codes and collated data to identify significant broader patterns of meaning (potential themes). It then involves collating data relevant to each candidate theme, so that you can work with the data and review the viability of each candidate theme. |
| 4. Reviewing themes | This phase involves checking the candidate themes against the dataset, to determine that they tell a convincing story of the data, and one that answers the research question. In this phase, |

---

[194] I opted to use the descriptions provided on their website rather than published text since they report that the labels have changed over the years. Their website hopefully provides the most current description of RTA.

| | |
|---|---|
| | themes are typically refined, which sometimes involves them being split, combined, or discarded. In our TA approach, themes are defined as pattern [sic] of shared meaning underpinned by a central concept or idea. |
| 5. Defining and naming themes | This phase involves developing a detailed analysis of each theme, working out the scope and focus of each theme, determining the 'story' of each. It also involves deciding on an informative name for each theme. |
| 6. Writing up | This final phase involves weaving together the analytic narrative and data extracts and contextualizing the analysis in relation to existing literature. |

Some items were initially coded only for interpretation rates, then thematic analysis was conducted later. However, for some datasets I took notes concurrently while coding the items for interpretation rate. These notes consisted of preliminary labels for recurring patterns and other comments where appropriate.[195] Regardless of whether notes were taken either concurrently or later, the goal of this initial round of assessment was to both to assist in familiarizing myself with the dataset by taking notes I could refer to later, and to begin formulating thoughts about the potential themes that would reflect the content of the data in a given dataset.

After reflecting on the overall content of a data set and reviewing my notes for recurring patterns, I developed a set of *themes* and coded each item in accordance with them using a particular label for each theme. These labels consisted of either a single word or a few words that succinctly

---

[195] This is consistent with Braun and Clarke's guidelines. They state that during phase 1, "[...] it is a good idea to start taking notes or marking ideas for coding that you will then go back to in subsequent phases. Once you have done this, you are ready to begin, the more formal coding process. In essence, coding continues to be developed and defined throughout the entire analysis." (2006, p. 87).

captured the meaning of a particular theme. *Themes* refer to relatively well-defined categories, or types of response that share a common meaning or cluster of related meanings. Braun and Clarke define a theme as: "a common, recurring pattern across a dataset, clustered around a central organising concept," adding that "A theme tends to describe the different facets of that singular idea, demonstrating the patterning of the theme in the dataset" (FAQs, n.d.). For instance, some participants interpreted statements about *metaethical relativism* to refer to the descriptive claim that people have different moral beliefs (i.e., *descriptive relativism*, Bush, 2016; Gowans, 2021). When this occurred, such items were coded with the label "descriptive." One difference between my approach and Braun and Clarke's is that the codes I used, i.e. the labels used in the thematic analysis portion of my datasets to indicate a particular theme tended to reflect broader thematic categories. In other words, Braun and Clarke tend to see codes as more specific, and serving as elements or facets of themes, which are broader and typically encompass one or more codes. I struck a middle path, tending to employ codes that more closely resembled themes, and as such there was little need to incorporate a cluster of related codes into a single theme. I effectively collapsed this distinction, employing codes of sufficient generality that I treated them as themes, and will refer to them as such.

This process of generating, defining, reviewing, and naming themes was not linear. That is, I did not generate an initial set of themes, write up a definition for each, slap a label onto them, and then impose this in a top down fashion on each dataset with the goal of dogmatically ratcheting every response into a predetermined category. Instead, as I reviewed responses I remained open to the possibility that each response could not neatly fit with one of the categories from the current pool of themes I'd developed. When this occurred, I had to make a judgment call about whether to make a note that the item defied easy categorization, whether I had to redefine, expand, or split existing themes into multiple labels, or even collapse two labels into one another, if a response prompted an insight that there was a close connection between what I'd been treating as two distinct themes. In

other words, I adopted a *nonlinear* approach to these phases, moving back and forth between each of the phases. This also included recoding a dataset in light of insights gleaned from coding other datasets.

Like Braun and Clarke, I appreciate that the coder plays an active role in the coding process, and that a linear and rigid approach to coding can lead to an inadequate reading of the data. For instance, suppose it only becomes apparent that there are two distinct themes emerging from the data halfway through coding the data. Rigid adherence to a particular coding scheme might motivate the coder to ignore this, and awkwardly code a particular pattern of responses in accordance with the closest-fitting label. Alternatively, a new theme could be introduced midway through coding, which could lead to coding inconsistencies. It would instead be best to go back over previous responses and reexamine them in light of the new insights gleaned from a particular response, or that occur spontaneously at any point during the coding process. Although it is far more labor intensive, it is best to instead refine and recode items. Braun and Clarke describe this process and the justification for it as follows:

> The coding process requires a continual bending back on oneself – questioning and querying the assumptions we are making in interpreting and coding the data. Themes are analytic outputs developed through and from the creative labour of our coding. They reflect considerable analytic 'work,' and are actively created by the researcher at the intersection of data, analytic process and subjectivity. Themes do not passively emerge from either data or coding; they are not 'in' the data, waiting to be identified and retrieved by the researcher. Themes are creative and interpretive stories about the data, produced at the intersection of the researcher's theoretical assumptions, their analytic resources and skill, and the data themselves. Quality reflexive TA is not about following procedures 'correctly' (or about 'accurate' and 'reliable' coding, or achieving consensus between coders), but about the researcher's reflective and thoughtful engagement with their data and their reflexive and thoughtful engagement with the analytic process. (p. 594)

Unfortunately, I do not believe Braun and Clarke are as sensitive to the shortcomings and risks posed by this approach as they ought to be, or at least they do not acknowledge these shortcomings or address them to the extent that I believe they should.[196]

One significant risk is that an active effort to extract meaning and patterns from the data could result in the construction and rationalization of illusory patterns, a kind of analytic apophenia. *Apophenia* refers to the perception of meaningful patterns or relationships where no such patterns or relationships exist, i.e. *illusory pattern detection* (Conrad, 1958; Ellerby & Tunney, 2017). In fact, Buetow (2019) explicitly suggests that qualitative approaches like those developed by Braun and Clarke risk their practitioners succumbing to apophenia, resulting in the detection or construction patterns which are simply not there. Since I cannot guarantee that every theme I have identified is genuine rather than a figment of my imagination, the best I can do is draw attention to this possibility and invite others to look at the data for themselves.

Second, I am not blind to my own expectations and hypotheses. Confirmation bias (Nickerson, 1998) could result both in biasing my coding of interpretation rates in a way favorable to my conclusions (i.e., fewer intended interpretations, more unintended interpretations), and in detecting themes that are consistent with my expectations and support my overall account of the data. Once again, one of the best checks against this is to invite others, especially those skeptical of my conclusions, to evaluate the data for themselves.[197]

Finally, I should note that, while it would be ideal to provide only a single theme for all responses, this would not accurately reflect the content of some responses. Most responses had only a single theme,

---

[196]At least not in this article; perhaps they do so in other publications.

[197] These risks may go some way in vindicating skepticism about the rigor of qualitative methods, but this simply isn't the case. Replicability and conventional validation (e.g., Cronbach's alpha, factor analysis, convergent validity, convergent validity) procedures are not adequate on their own to ensure that a measure captures what it is intended to (Lilienfeld & Strother, 2020; Maul, 2017). Even if researchers took a more quantitative approach to assessing participant interpretation, subjective judgment about whether participants interpreted stimuli as intended would have to enter the assessment at some point.

but some responses fell into two or more categories. Whenever this occurred, all categories were represented in the coding for that response, separated by a comma. Order of coding for responses with multiple codes does *not* move from more central to less central themes (or vice versa), but merely reflects the order in which I judged the theme in question to be present. As such, the order of themes listed for multi-theme responses does not reflect claims about the data, it is merely an artifact of my personal method of coding. Generally, such coding reflects the order in which the theme in question became apparent in the text, but this is not always the case.

When a response could not be readily categorized in accordance with a recurring theme, it was coded as "other." When the participant did not respond, it was coded as "no response." When the response was uninterpretable, it was categorized as "uninterpretable." Some items may include additional notes when they were sufficiently distinctive to warrant additional commentary. Insights can often be gleaned from these responses, even if they do not conform to any particular theme. Many such responses are worth reflecting on and discussing individually, since many of the most interesting features of a response cannot be captured by labels alone. Coupled with my discussion of the recurring themes that accompany each dataset, I will discuss these items in a more open-ended fashion, both because doing so often serves to support my analysis and because discussion of individual items plays an important role in justifying my coding decisions.

I focused primarily on generating themes for unintended interpretations.[198] Before coding responses, I already had strong theoretical grounds for expecting particular, recurrent ways participants would be likely to interpret metaethical stimuli in unintended ways. For instance, I

---

[198] It should come as no surprise that unclear responses are of far less interest since items coded this way were often uninterpretable, confusing, or simply left blank. Little of interest can be gleaned from such responses. Likewise, there is little value in identifying themes in intended interpretations, since the central theme would consistently simply be the relevant metaethical interpretation. Nevertheless, the content of intended interpretations is often worth discussing, since it is still worth considering just how well responses coded as intended interpretations map onto the relevant metaethical distinctions, and there is also value in assessing the specific wording such responses employed (such as how often participants use explicit metaethical jargon, e.g., "objective" and "relative").

expected many participants to appeal to epistemic or normative consideration, rather than metaethical ones.

However, I did not rigidly adhere to a list of expected themes I developed in advance of analyzing any particular dataset. Instead, I remained open to generating unexpected themes when they offered the best account of a given response or pattern of responses. Thus, my approach to generating themes was guided both by top-down, theoretically motivated expectation *and* an open-minded receptivity to generating ad hoc themes where necessary. For instance, I expected many participants to conflate metaethical relativism with descriptive claims about the existence of moral diversity, and I expected many participants to conflate objectivism with absolutism (i.e., the notion that there are exceptionless moral rules), and was primed in advance to recognize instances of these interpretations. However, I was not expecting that, for some studies, many participants would interpret a statement about objectivism to mean the exact opposite, yet this is exactly what I found. That is, when explicitly asked what it means for morality to be "objective," many participants stated something like the following:

> *Moral truth is objective because everyone has different views on what is moral or not. What might be immoral to one person might be perfectly moral to another person.*

When unexpected responses like these appeared in a dataset, I created labels for them and coded the dataset accordingly. If I had relied on *a priori* classifications prior to coding the data, these responses would not fit any of the categories I'd have come up with. I was also surprised by how much variation manifested *between* datasets. Variation in wording and stimuli led to remarkably different profiles of responses, which meant there was little value in attempting to import the categories from one dataset to another, beyond a few recurring themes. In short, there were both expected and unexpected themes in each dataset. While I began coding with certain expectations in mind, these expectations would be inadequate to cover the range of interpretations that actually emerged from the data. As a result, I

supplemented these expectations with the creation of labels suitable for adequate characterization of the data.

### S4.7.2 Predictions for thematic analysis

Predictions about the themes I expected to emerge for a given dataset are best addressed on a per-dataset basis, and I discuss these predictions for each dataset below. However, I did expect several general patterns to emerge.

First, I expected many participants to report first-order normative judgments rather than metaethical judgments. That is, when confronted with questions that were intended to elicit a metaethical interpretation (a second-order interpretation), I expected many participants to report their stance about what is morally right or wrong, to state that they are correct, or to state that someone who held a moral stance contrary to their own was incorrect or a bad person. This prediction is largely attributable to Wainryb et al. (2004), which indicated that the vast majority of children, when asked to explain their response to questions ostensibly about metaethics, simply stated their first-order, normative position. I expected many adults would do the same, in large part because most everyday moral judgments are about actually making moral judgments and making moral claims, rather than thinking about abstract considerations about what one is doing when rendering a moral judgment or making a moral claim. In particular, I expected these responses to be especially likely when participants were presented with the disagreement paradigm, since the disagreement paradigm tends to describe people who express moral standards that conflict with the participant's.

### S4.7.3 Study 1A: Thematic analysis

Thematic analysis was conducted in accordance with the procedures outlined in section S4.7. To simplify the analysis, I will focus only on the five most common themes, which may be seen in **Figure S4.9** and **Table S4.3**.[199,200]

**Figure S4.9**

*Most common themes for Study 1A*

---
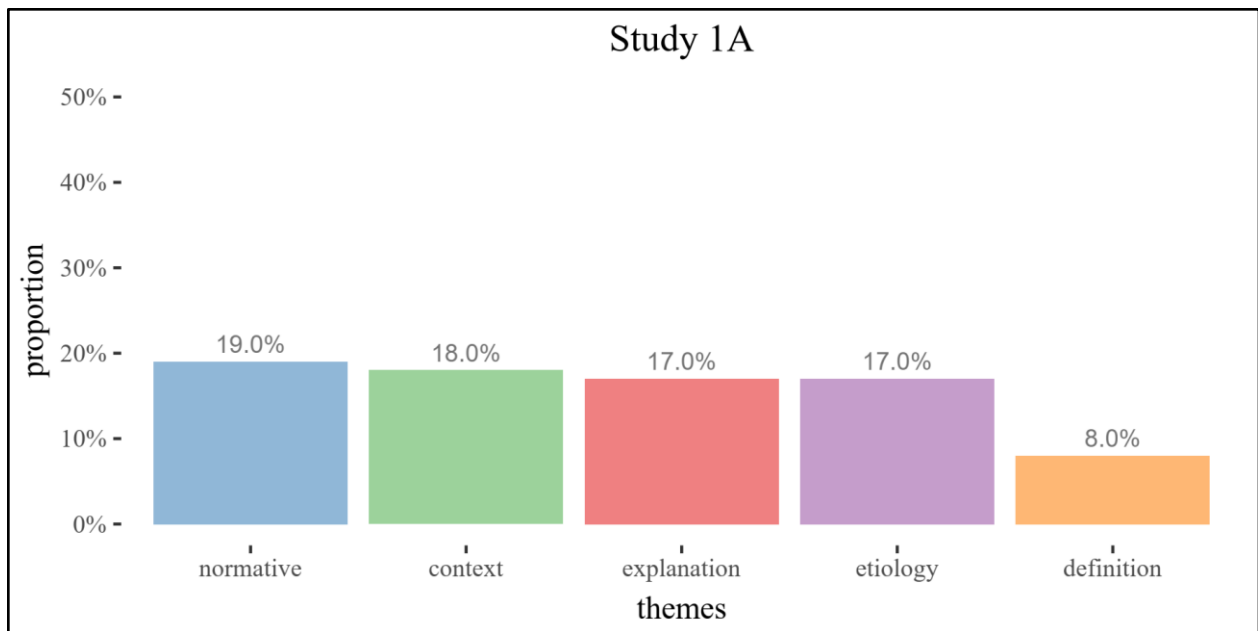
[199] Coding included indicating when an item appeared correct or unclear, but these do not represent actual themes and will not be discussed in analysis, as they are qualitative recapitulations of the quaternary coding scheme.

[200] Note that since items could have more than one theme, items coded with one of these themes may have exhibited other themes as well.

**Table S4.3**

*Most common themes for Study 1A*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *Normative* | Expressed moral judgment about the person who disagreed with them | 19.0% | 19 |
| *Context* | Disagreement attributed to different assumptions about the circumstances in which the action was performed | 18.0% | 18 |
| *Explanation* | Offered an explanation for why the person might believe what they do | 17.0% | 17 |
| *Etiology* | Offered a causal account of how the person would respond as they did | 17.0% | 17 |
| *Definition* | Attributed disagreement to different definitions of moral terms/concepts | 8.0% | 8 |

The *normative* theme refers to instances in which participants expressed a normative (first-order) moral judgment of some kind. In other words, a direct judgment about what is moral or immoral, rather than a judgment about *what it means* for a judgment to be moral or immoral. Typically, this involved expressing moral condemnation of the person who disagreed with them. For instance, one participant stated:

> *While this may be the other person's opinion, it is such a deplorable view that I feel it is completely wrong.*

Note that such a response does not attribute the source of disagreement to different moral standards, nor does it attribute it to something other than a difference in moral belief. Instead, it fails to adequately address the question at all. As such, we cannot judge this participant to have clearly interpreted the source of the disagreement as intended. Instead, such responses indicated a negative

moral attitude towards the person who disagreed with them. This cannot tell us whether the participant believes that whoever disagreed with them did so because that person had different moral standards. They could think that this person recognizes the same moral truths as they do, but simply does not care about those moral facts, or they could think that the person disagreed with them as a result of nonmoral differences in belief. This is a common point philosophers raise when discussing whether there are universal moral standards. Two people may disagree about the moral status of an action not because they have different moral standards, but because they disagree about the nonmoral facts.[201] Thus, absent additional commentary from the participant, such responses do not indicate that the participant understood the disagreement paradigm as intended.

*Context* and *etiology* are tied for the second most common theme. *Context* refers to instances in which participants attributed the source of disagreement to different conceptions of the nature of the action in question. For instance, when asked why someone disagreed with them about the moral status of firing a gun on a crowded city street, one participant stated:

> *A difference in perception of a situation in which gunfire was opened on a crowded city street. I was thinking gunfire from terrorists/ criminals; other person may have thought gunfire from police officers to catch a criminal.*

Without additional comments that attributed the source of disagreement to a difference in moral values, such responses typically indicated a clearly unintended interpretation. This is because such responses involve attributing the disagreement to something other than a genuine difference in moral beliefs.

---

[201] For instance, many people oppose beating children and capital punishment because they believe these practices are *ineffective*. Two people might both agree that *if* capital punishment were an effect deterrent, then it would be morally justified, but they simply disagree about whether it is an effective deterrent. In such circumstances, people's moral beliefs are due to different nonmoral beliefs. If so, someone might think another person is mistaken for reasons unrelated to whether there is a stance-independent fact of the matter.

*Etiology* refers to comments that offer an account of how the other person may have arrived at the particular stance they took towards the moral issue in question, such as attributing it to how the person was raised, or what they were taught, e.g.:

> *I would think this person has been brought up a different way than I have been raised.*

Such responses are consistent with attributing the source of disagreement to a difference in moral beliefs, such remarks do not *clearly* do so. Someone raised to hold different moral standards may hold those standards due to differences in nonmoral beliefs. If so, then the reason why they disagree with the participant could be for reasons other than a fundamental difference in moral belief. In some cases, participants offered a psychological explanation, or described the sorts of attitudes the other person may have, e.g.:

> *Animosity towards bureaucracy.*

In such cases, it is unclear whether the participant regarded the other person as having different moral standards or values. While it is plausible they do think this, they may have interpreted the question about the source of moral disagreement to be asking them for an ultimate, rather than proximal explanation. Simply put, explaining why a person responded to a question differently than you does not necessarily entail that they did so because they have different moral standards. Such responses suggest that participants may have interpreted the task in an unintended way for completely understandable reasons. If I ask you why someone disagreed with you about whether cheeseburgers were good, you *might* say "because they have different food preferences than me." This would be analogous to the kind of response Goodwin and Darley need to confirm that people understood moral disagreements as intended: the disagreement about whether a particular type of food was good or bad *must be attributed to different food preferences.* But you might instead give a causal-historical explanation: "perhaps where they grew up nobody served good burgers. If they tried the burgers at my uncle's barbeque, they'd change their mind." This is a *perfectly reasonable response* and it does respond to the

question *why does this person disagree with you*? However, it does not tell us whether you think this person has different taste preferences or not. Rather, it could simply point to this person having had different experiences that have led them to form a conclusion on the basis of differential access to certain kinds of information. Just the same, someone *could* think that people's moral values are fundamentally the same (at least with respect to the moral issue in question), but that in virtue of their history and life experiences they are motivated to respond to the question in a way that conflicts with the participant, *even though they do not fundamentally disagree about what is morally good or bad*. In short, while *etiology* responses are a perfectly appropriate way to respond to a question about why someone disagrees with you about a moral issue, it does not clearly indicate that they disagree *because* they are committed to a different moral standard about the issue in question.

Ironically, it may be that unintended interpretations of the question about the source of disagreement can make it difficult to tell whether the participant interpreted the disagreement paradigm itself as intended. If so, then this may indicate that the coding scheme used here underestimates the rate of clearly intended interpretations. This works against my suspicion that most participants did not interpret the source of disagreement as intended, but it indirectly serves to illustrate how readily people can interpret questions in unintended ways; and, in any case, I don't think attributing the source of the disagreement to something other than different moral standards is the only or primary reason why the disagreement paradigm is invalid, anyway; it's just *one* of the many points where a subset of participants interpreted the question in an unintended way.

Finally, some participants attributed the source of moral disagreement to different definitions or conceptions of morality. While this might seem like a straightforward case of a clearly intended interpretation, it is not. In order for people to have a genuine difference in moral standards, they would have to have a different first-order moral stance towards the same moral issue, where "moral issues" are understood in approximately the same way. To illustrate why, consider two people who

both like ice cream sandwiches, but do not like ordinary sandwiches (i.e., sandwiches that use sliced bread and include savory ingredients). However, these people have different definitions of "sandwich." One person has a narrow conception of sandwiches, which includes only savory ingredients between pieces of bread, while the other person has a broader definition that includes any ingredients, sweet or savory, and that therefore includes ice cream sandwiches with cookies in place of bread, and ice cream in place of traditional sandwich contents (e.g., lunchmeat, lettuce, tomatoes, cheese). If asked whether they "like sandwiches," one person would say "yes" and the other would say "no." These people would appear to disagree about what food is good or bad, and to therefore have different first-order evaluative stances towards what food was "good." However, they actually have the exact same food preferences, but simply disagree about what a "sandwich" is. As a result, they do not actually have a genuine disagreement about what food is good or bad, just a different conception of what a "sandwich" is. Likewise, two people could share the same first-order moral standards about which actions are right or wrong, but disagree about what counts as a moral issue.

While this may seem implausible to some readers, this may be due to readers underestimating the extent to which ordinary people disagree with one another and with researchers when asked to classify various actions as moral or nonmoral. As Wright and colleagues have shown, participants exhibit considerable variation in their classification of issues as moral or nonmoral. In addition, a majority of participants routinely classify some issues as nonmoral that might surprise researchers. For instance, when Wright asked participants to classify issues as moral or nonmoral, she found that 22% of participants did not consider cheating on one's spouse to be a moral issue, and 86% did not consider eating factory-farmed meat to be a moral issue. It is one thing to consider adultery and practices that cause enormous animal suffering to be morally permissible; it is quite another to consider such issues to not even be moral issues at all. Of course, it is possible participants interpreted the classification task in unintended ways as well. Even so, participants do not reliably categorize issues as "moral" or

not in the same way as one another, both within (Wright, Grandjean, & McWhite, 2013) and between populations (Levine et al., 2021; Machery, 2018). This may be indicative if at least some conceptual variation in how people think about what does or doesn't constitute a moral issue, and pending evidence that people do think of morality in the same way, it remains an open possibility that people really do operate under different definitions of "morality" or what constitutes a moral issue, even if those people share similar attitudes about the normative status of the issue in question.

For instance, Wright found that 63% of participants categorized an abortion during the first trimester as a "personal" issue rather than a moral one, while only 36% considered it a moral issue. In short, ordinary people may think it possible that other people have different definitions of morality, and that they could express different stances towards a particular issue as a result of these differences, rather than a difference in their stance about the normative status of the action. If so, such disagreements may not be due to different moral standards per se, but to different conceptions of what it would even mean for something to be a moral issue. Just as people who disagree about the definition of a "sandwich" could disagree about whether "sandwiches are tasty," even if they have the same food preferences, so too could people have the same stances towards a particular action even if they have different conceptions of what it would mean for that action to be morally right or wrong.

### S4.7.4 Study 1B: Thematic analysis

The most common themes can be seen in **Figure S4.10** and **Table S4.4**.
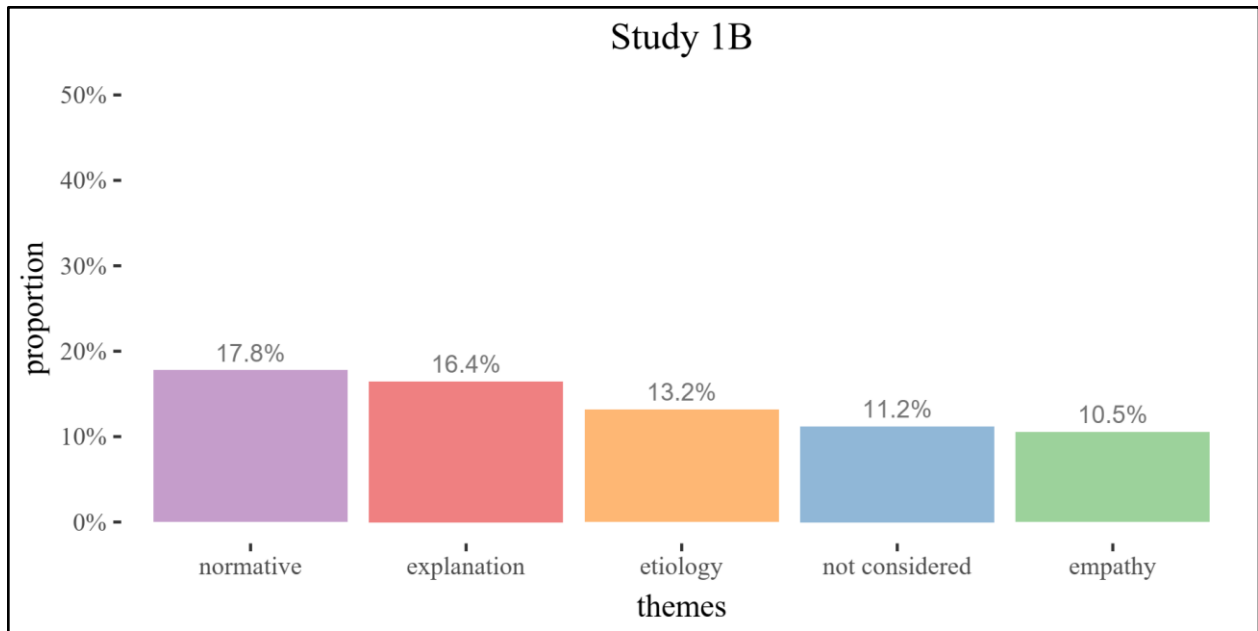
**Figure S4.10**

*Most common themes for Study 1B*



**Table S4.4**

*Most common themes for Study 1B*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *Normative* | Expressed moral judgment about the person who disagreed with them | 17.8% | 27 |
| *Explanation* | Offered an explanation for why the person might believe what they do | 16.4% | 25 |
| *Etiology* | Offered a causal account of how the person would respond as they did | 13.2% | 20 |
| *Not Considered* | Attributed disagreement to the other person not thinking about the scenario in an adequate way | 11.2% | 17 |
| *Empathy* | Expressed empathy or compassion towards someone who disagreed | 10.5% | 16 |

Like study 1A, *normative* responses were the most common. Context was no longer among the top five, though *explanation* and *etiology* remained common responses as in Study 1A. However, unlike Study 1A, *not considered* and *empathy* were among the five most common responses. *Not considered* refers to instances where the participant attributed the disagreement to the other person not considering one or more factors relevant to assessing the situation in the same way as the participant, which is in some respects similar to attributing the disagreement to nonmoral differences. Here are a pair of examples that exemplify this type of response:

> *Maybe the person is desperate for a job or hasn't considered the ramifications of not being qualified.*

> *Perhaps the other person didn't realize the implications fo getting an undeserved job for which one isn't qualified and won't be able to perform successfully.*

These responses suggest that the participant did not necessarily think that the person who disagreed with them has different moral standards. Instead, that person may have simply not adequately engaged with and thought about the question in the proper way. In fact, consider the first response. In suggesting that the other person may have because they did not consider the ramifications of not being qualified, this pragmatically implies that the person *would* agree if they *did* consider the ramifications. That is, they *would* agree with the participant *if* they considered the question in the same way as the participant. If anything, this implies that the participants think that the other person shares the same moral standards, but interpreted the question differently. This is a perfectly reasonable response, but it attributes the disagreement to something other than a genuine difference in moral values. *Empathy* refers to just what it appears to: any instance in which the participant empathized with or attempted to express understanding for what might motivate someone to disagree with them. For instance, one participant said:

> *I feel bad for the person who had to rob the bank for that reason but there are other ways to get money. If they are poor, may know what it's like (?)*

Such responses cannot tell us whether they believed the other person had different moral values.

### S4.7.4 Study 1C: Thematic Analysis

The most common themes can be seen in **Figure S4.11** and **Table S4.5**.
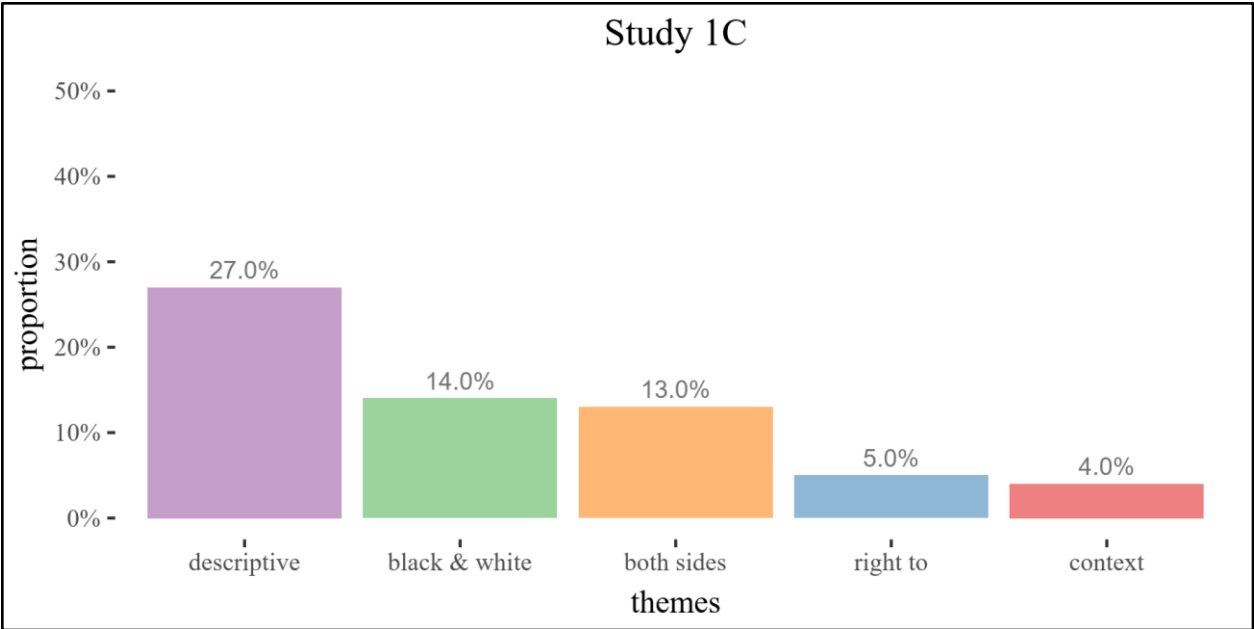
**Figure S4.11**

*Most common themes for Study 1C*



**Table S4.5**

*Most common themes for Study 1C*

| Theme | Explanation | Percentage | Frequency |
| --- | --- | --- | --- |
| *descriptive* | The view that different people or groups have different moral standards | 27.0% | 27 |
| *black and white* | Explicitly describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality | 14.0% | 14 |
| *both sides* | Both sides of a disagreement have part of the truth | 13.0% | 13 |
| *right to* | Each person considers their standards to be correct | 5.0% | 5 |

| *epistemic* | General appeals to epistemic considerations | 4.0% | 4 |

The most common theme was *descriptive*. In this study, this involved an appeal to the notion that different people have different moral beliefs or values. While differences in moral belief, or *descriptive relativism* may be a common belief, this is a mere descriptive fact that is consistent with both realism and antirealism (including relativism). While it is possible to infer metaethical relativism or realism *because* of the existence of descriptive differences in moral belief, neither necessarily follows, and, in any case, if this inference is not made explicit we cannot know whether the participant drew any particular metaethical inference. Instead, it would appear some participants simply took both sides being correct to *just mean* that they had different moral beliefs. While this is not an especially plausible reading if one is explicitly engaged with the question, it is possible participants did think the purpose of the question was to assess whether the participant thought people have different moral beliefs. Here are a few illustrative responses:

> *Each person has their own set of moral beliefs.The way moral beliefs work is that they can vary.*
>
> *Because they may think differently then the other person.*
>
> *Everyone believes different things*

None of these responses involve any considerations directly and clearly relevant to a metaethical rationale for selecting any particular response to the question, indicating that these participants may not have interpreted the question as intended. Next, 10.9% ($n = 14$) of participants referenced the notion of viewing morality as "black and white" or the notion of "grey areas," typically in order to deny that morality is black and white, and to affirm that there are "grey areas." Roughly, this captures the notion that there is a definitive, discrete, binary answer to every moral issue, a notion reminiscent in some ways of moral absolutism, though often with some allusions to a rigid, inflexible, definitive,

readily epistemically assessable, and strict moral code. Here are some examples of this pattern of response:

> *Not everything in the world is black and white and not everything is a clear positive or negative.*

> *cuz morality is bullshit binary thinking and has no basis in reality*

> *I don't think everything is black and white, but there are shades of grey, and there are multiple solutions and ways to looking at issues and problems.*

Unfortunately, a "black and white" perspective on morality is *not* the same thing as a realist perspective on morality. As such, participants who opted for this response appear to have not interpreted the question as intended. Take, for instance, the third response, which ends on, "[...] *and there are multiple solutions and ways to looking at issues and problems.*" This is true enough. For instance, there may be multiple ways to treat an illness, or multiple ways to build a bridge. However, there are still stance-independent (i.e., "objective") facts about what techniques effectively treat illnesses and what engineering methods will result in a standing bridge. Some participants seem to interpret the notion that two people can both be correct as the notion that two people could both conform with the same moral rules in different ways, which *is not* the same thing as relativism. As such, participants who favored this response seem to have favored the "relativist" response for understandable reasons, but reasons that nevertheless differ from an expression of relativism.

The next most common theme was *both sides* at 10.2% (*n* = 13). This is a special category of epistemic interpretation of the question. Such participants tended to explain their reason for answering as they did by appealing to the notion that both sides of the disagreement may have valid points to make, or have perspectives or that capture part of the truth. Here are a few examples:

> *Both sides can make valid points*

> *I think even with moral issues, there are often no complete absolutes, so there could be at least a degree of validity even on opposite sides of a moral issue, depending on how you approach it.*

> *Depending on the issue, I think there are two sides and opinions to every story/argument.One can not always be right and both sides could have a point.*

These examples suggest that these participants may have judged that both people could be correct not because two people with different moral standards could each be correct relative to their respective moral standards. Rather, each person could be *partially* correct, or offer a legitimate, correct, or valid perspective on a particular moral issue, even if there were a single, definitive, stance-independent fact of the matter. Thus, these responses do not seem to indicate that the participant interpreted the notion that both people could be correct as an expression of relativism, and their answers are therefore not necessarily indicative of an intended interpretation.

A handful of participants (3.9%, *n* = 5) appear to conflate the notion of two people being correct, and two people *considering* themselves to be correct, as reflected by the *right to* theme:

> *I chose the response because people believe they are right*

> *I think what is moral is something that is not set in stone. Different people believe issues are moral or immoral. So, two people could see the same issue differently, and from their own viewpoint believe they are right.*

A few (3.1%, *n* = 4) also appealed to uncertainty or a variety of other epistemic considerations:

> *It depends upon one's beliefs, especially when there is an unknown, unprovable variable.*

These kinds of responses imply *epistemic* interpretations rather than metaphysical or metaethical interpretations. Thus, they are more consistent with unintended interpretations, and are exactly the kind of responses we should expect if participants did not interpret the disagreement paradigm as intended. Of course, such responses were quite rare, comprising only a small proportion of responses. This would indicate that if epistemic conflations do occur, they are not very common, and that if participants are not interpreting questions as intended, this may be for reasons other than epistemic interpretations.

### S4.7.6 Study 2: Thematic analysis

The most common themes can be seen in **Figure S4.12** and **Table S4.6**.[202]

**Figure S4.12**

*Most common themes for study 2 (by condition)*



Study 2A: Direct realism

---
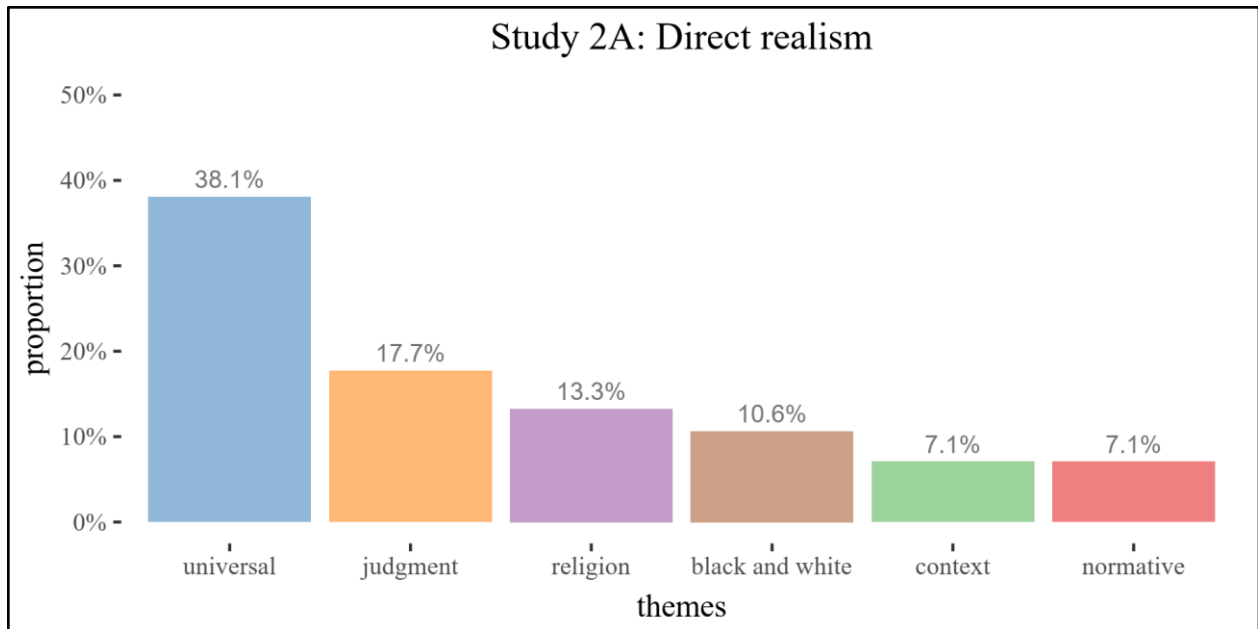
[202] Several categories were dropped from analyses. This includes "correct," "unclear," "other," and "no_answer." These are quasi-categories that can serve some forms of qualitative analysis but do not directly correspond to substantive themes. Roughly, "correct" was used whenever an item was flagged as seemingly intended, regardless of whether it indicated a 1|1 or 1|0 response. "Unclear" was used when a response was unclear, indicating a |0 coding. "Other" referred to responses that were unique or unusual, and defied any recurring theme or pattern. Such items may be of interest for qualitative analysis, but should be addressed individually. The theme "no_answer" captured instances where participants did not respond at all. This was a common occurrence in this particular study.

Study 2A: Direct relativism

Study 2B: Disagreement realism
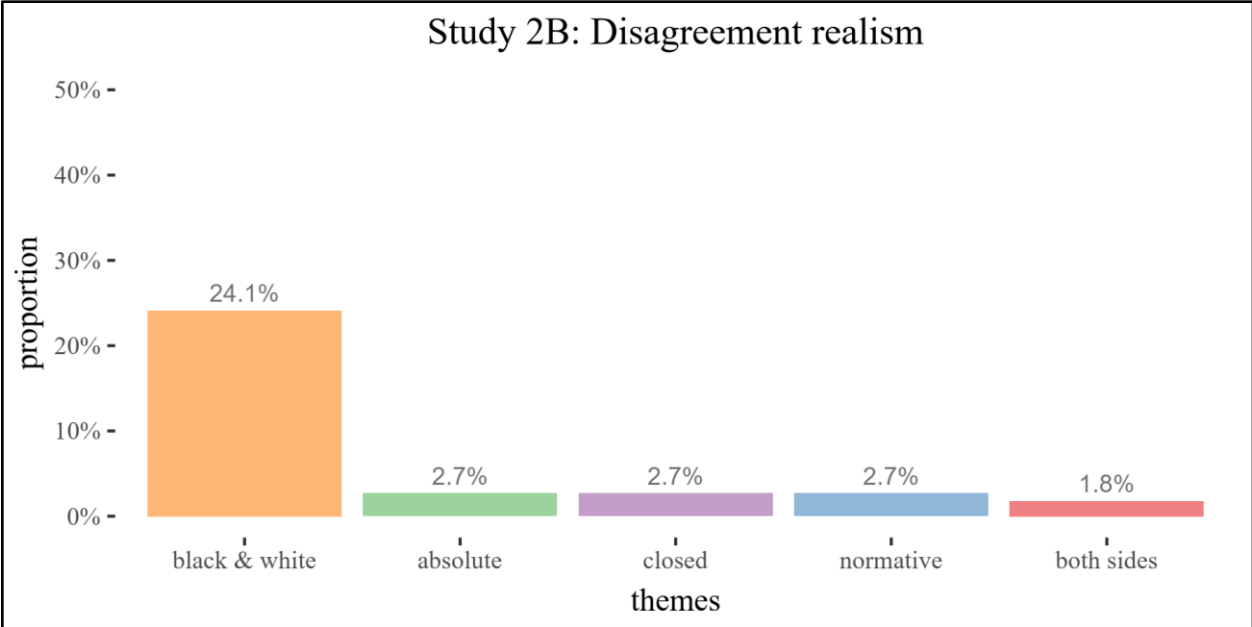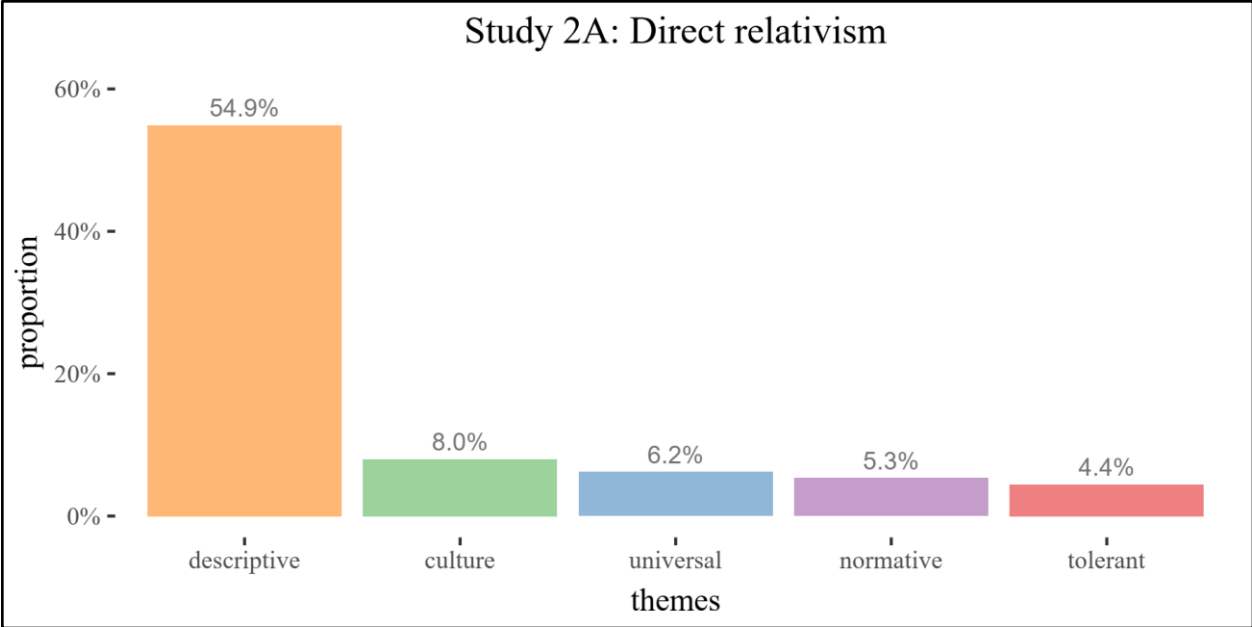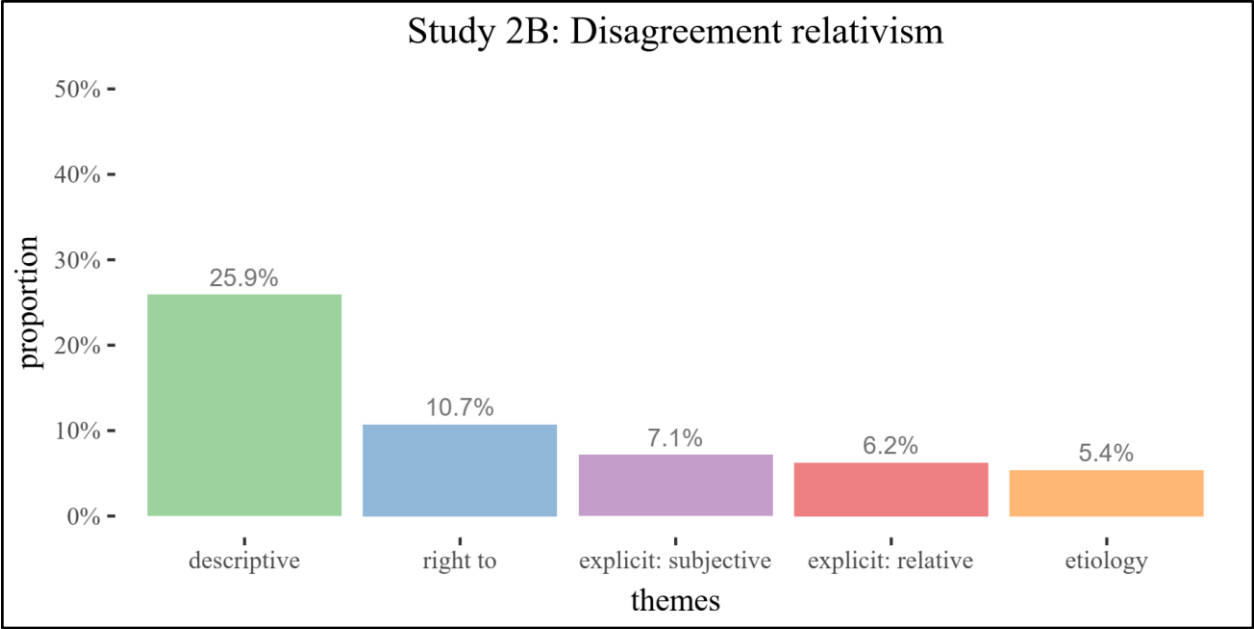
Table S4.6

*Most common themes for Study 2 (by condition)*

*S4.6.1 Direct realism*

| Theme | Explanation | Percentage | Frequency |
| --- | --- | --- | --- |
| *universal* | Normative position which holds that a given moral norm applies to everyone | 38.4% | 43 |
| *judgment* | Normative claim that people should act in accordance with the speaker, or that the speaker's views are the "only" way to act | 17.9% | 20 |
| *religion* | Refers to religion or religious beliefs | 13.4% | 15 |
| *black and white* | Explicitly describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality | 10.7% | 12 |
| *context\** | The view that whether an action is right or wrong depends on context/circumstances | 7.1% | 8 |

| Theme | Explanation | Percentage | Frequency |
|-------|-------------|------------|-----------|
| *normative** | Miscellaneous remarks about normative and evaluative conceptions of morality | 7.1% | 8 |

*Note.* Items with an asterisk (*) were tied for the most common response.

*S4.6.2 Direct relativism*

| Theme | Explanation | Percentage | Frequency |
|-------|-------------|------------|-----------|
| *descriptive* | The view that different people or groups have different moral standards | 54.9% | 62 |
| *culture* | Descriptive or etiological claim that attributes moral stance to a person's culture | 8.0% | 9 |
| *universal* | Normative position which holds that a given moral norm applies to everyone | 6.2% | 7 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 5.3% | 6 |
| *tolerant* | Normative claim that we should tolerate or respect other people or cultures (or their moral standards) or that we shouldn't judge others | 4.4% | 5 |

*S4.6.3 Disagreement realism*

| Theme | Explanation | Percentage | Frequency |
|-------|-------------|------------|-----------|
| *black and white* | Explicitly describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality | 24.1% | 27 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 2.7% | 3 |

| | | | |
|---|---|---|---|
| *closed* | View that an utterance indicates the speaker is close-minded or rigid in their thinking | 2.7% | 3 |
| *absolute* | Varies in meaning, but is associated with explicit use of term "absolute," exceptionless moral rules, black and white thinking, certainty, or being close-minded | 2.7% | 3 |
| *both sides* | The claim that both sides of a disagreement have part of the truth | 1.8% | 2 |

*S4.6.4 Disagreement relativism*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 25.9% | 29 |
| *right to* | Each person considers their standards to be correct | 10.7% | 12 |
| *explicit: subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 7.1% | 8 |
| *explicit: relative* | Explicit use of term "relative" (or related term, e.g., "relativism" | 6.3% | 7 |
| *etiology* | Offered a causal account of how the person would respond as they did | 5.4% | 6 |

It would be difficult to offer anything even approaching a thorough analysis of the recurring themes that appeared throughout the four different responses. Most of the themes described here are fairly self-explanatory or easy to work out from the summarized description. Hopefully it will suffice to briefly describe some of the highlights and to offer some general observations. First, as I discuss below, the *direct realism* condition would be better described as an expression of universalism and a

normative claim that people should act in accordance with the speaker's moral standards. Consistent with the poor validity of this item, a the two most common themes were *universal* and *judgment*, which provides evidence of the invalidity of the measure I used.

*Direct realism*

The most common themes for this condition were *universal* and *judgment*. This is consistent with an error in the operationalization of the realism condition. Realism holds that there are stance-independent moral facts. This is orthogonal to the *scope* of a moral fact, i.e., who that moral fact applies to. The wording used in this study is more consistent with moral *universalism*, which holds that one or more moral norms (or morality in general) applies to *everyone*. I conflated these concepts. This is corroborated by "universal" emerging as the most common theme. Participants often expressed this by directly using the term "universal," while others responded in a way that could be readily understood to convey universality:

> *He means there is a standard of morality that is universal.*

> *The respondent means that there should be only one standard that should apply to all.*

Second, I used explicit normative language ("should be judged"), which is consistent with another inappropriate conflation between metaethics and normative considerations. This is captured by the theme *judgment*, which describes a normative stance John takes towards people with different moral perspectives, typically that he judges them according to his standards and thinks everyone should conform to his standards. While this is not the same thing as realism, and is thus an unintended interpretation, it is understandable given how the item was worded:

> *I think that he means that he judges people by what he perceives as being right or wrong and if you do no agree with him, you are wrong.*

> *I think he believes that people should be judged (by him) in the manner that he feels is moral based on his own personal beliefs*

The third most common theme for the *direct realism* condition was *religion* (8.9%, *n* = 15), a theme that is not typically as common. Such participants often suggested that John was religious or was appealing to religious moral standards. Such remarks were often explicit and unambiguous:

*It sounds like some shit some old religious person would say*

*The only moral truth that exists is from the Bible/my religion*

*He is probably very religious (Christian or Muslim) and intolerant.*

Nothing about the item explicitly referenced religion or anything related to religion. This provides some indication that at least some people may associate expressions of universal moral standards as a signal that the speaker may be religious. Although religion was referenced by fewer than 10% of participants, it is still remarkable that this was a sufficiently salient consideration that participants were motivated to explicitly reference it. It seems plausible that, if prompted to predict whether the speaker was religious or not, the kind of response John offered would tend to be associated with religious belief. Some moral claims may be associated with people with particular demographic or ideological characteristics. Since inferences about those characteristics may vary across participants and populations, there is considerable potential for interpretive variation predicated on extraneous assumptions that vary in accordance with cultural and other forms of knowledge to influence response patterns in unintended ways.

However, the frequency with which religion was mentioned supports a more general observation: participants do not read statements or interpret stimuli in isolation. Rather, they make a variety of inferences based on background assumptions. Such background assumptions can vary from item to item, and can vary both within and across populations. And such background assumptions can also vary across items ostensibly intended to mean the same thing. For instance, suppose participants were asked to consider these sentences:

*"It is immoral to torture children for fun."*

*"It is immoral for gay people to get married."*

*"It is immoral to tell small and insignificant lies."*

*"It is immoral to express intolerance for people with different cultural practices."*

Would people interpret these remarks the same way? If they were philosophical robots they might. But each of these remarks may prompt a host of background assumptions about the goals and intentions of the speaker who would make such claims, and the characteristics of such a person, such as demographic variables. It would be remarkable, for instance, if people would not expect opposition to gay marriage to be associated with religiosity and political conservatism, which could in turn influence assumptions about what such a person was likely to mean. And we might think, for instance, that moral opposition to gay marriage is more likely to convey a realist stance towards morality than a statement, while opposition to cultural intolerance is more likely to be expressed by a relativist.

In short, much of the work that goes into people's efforts to interpret utterances relies on accessing background information that is culturally contingent and potentially variable within and between cultures. There may be no correct or canonical single literal meaning that all competent speakers of a language would attribute to a particular claim. We do not occupy the exact same communities, and our assumptions about what sentences, even an identical sentence expressed by native speakers of the same language, could vary in accordance with considerations extraneous to the formal and explicit content of the utterance itself. Researchers and philosophers who present participants with decontextualized toy sentences do not consider the role context and background information in evaluating the meaning of an utterance ignore such considerations at their peril, and risk poor study design and mistaken inferences.

Another notable theme to emerge from these conditions was *black and white*. I will address this theme now, since it is a fascinating composite folk term that emerged in numerous paradigms and seems to be a common and important way ordinary people think about morality. In this particular

study, the theme typically involved participants explicitly stating that John had a "black and white" view of morality, that John rejected the notion that morality had any "gray areas," or both. This is one of the few themes for which precise terms used to convey what the participant means are so consistent. Often a theme must be inferred from a variety of different ways of expressing it with no specific term or terms that serve as reliable cues. Not so with references to morality being *black and white*:

> *Morality is black or white.*

> *there are no grey areas*

> *I think he means there is right and wrong, it's black and white and clearly defined.*

> *That he believes morality is a question of absolutes, blacks and whites, with no gray areas.*

It is easy to identify even when less explicit:

> *There is right and wrong and no inbetween*

> *There is only one right or wrong in each situation, things are not ambiguous.*

In describing a "black and white" view of a moral issue, participants seem to roughly have in mind some combination of these qualities, with the particular combination varying between participants:

> *1. Absolutism*: exceptionless moral rules

> *2. Categoricity*: discrete rightness and wrongness, no mixed moral valence (both good and bad)

> *3. Epistemic ease:* Clear or unambiguous satisfaction conditions

> *4. Closed*: Close-mindedness or rigidity

**Absolutism**

Absolutism roughly conveys the notion of an exceptionless moral rule, or a moral standard that is insensitive to contextual considerations, such as the intentions of the agent, the expected outcome of the action, or other variables that could mitigate or alter the moral status of the action. For instance, someone who held an absolutist stance towards lying might claim that lying is always wrong, even if

the lie would save lives, or was intended to spare someone's feelings. The association between black

and white moral thinking and absolutism is often made explicit:

> *There is no gray area with morality, its a absolute*

> *That he believes morality is a question of absolutes, blacks and whites, with no gray areas.*

### Categoricity

Categoricity refers to the nature of the action in question: it is either discretely *good* or *bad*, *right* or

*wrong*. We might think some actions can have mixed moral valence. Suppose someone steals medicine

for their sick grandmother. We might think that stealing is morally bad, but that helping your

grandmother at great personal risk to yourself is morally good, even noble. As a result, we might feel

ambivalent about stealing to help someone: it is a "gray" moral action, in that it has both a morally

good element (altruism, concern for family) and a morally bad element (stealing other people's

property). Participants who regard a particular moral stance as black and white often seem to have in

mind the notion that someone rejects the ability for an action to have mixed moral valence. Rather,

all actions are either categorically *good* or *bad*. The phrase "black and white" is an apt metaphor, in that

it conveys that there can be no *mixing* of the moral status of an action, which would instead result in

some moral issues being "grey." Some people may think that if we were to examine a variety of actions

with moral valence, we'd judge some to be uniformly immoral, some to be uniformly good, and some

to have both good and bad. The absolutist is someone who denies the latter possibility, at least for the

moral issue in question.

### Epistemic ease

Some participants also include an epistemic element in their description of a moral perspective being

"black and white." This typically consists of suggesting that a black and white perspective towards

moral issues entails the perspective that there is a "clear" division between right and wrong, such that

on considering any particular action or situation, rendering a decisive moral judgment is a straightforward process for which one need engage in little deliberation or consideration of the details:

> *I think he means there is a clear division between right and wrong and everyone should follow those same rules.*

Alternatively, participants could mean that the standards, however they are formalized, are clear and unambiguous:

> *There is only one right or wrong in each situation, things are not ambiguous.*

> *I think he means there is right and wrong, it's black and white and clearly defined.*

### Closed

Finally, some participants associate black and white thinking with a closed, rigid, or inflexible approach towards the moral issue. That is, people who have a "black and white" perspective on a moral issue will be less open to changing their mind, considering situational factors others may deem relevant, or engaging with or listening to people with alternative perspectives. Note this participant's reference to "tunnel vision":

> *it is a statement that is centered on one right way and only way way to view actions. It is a tunnel vision and a black and white view of the world.*

> *I think he sees a fine line between right and wrong and isnt very open to change*

It's not always clear which (if any) of these particular characteristics participants have in mind, but references to each of these four traits in conjunction with use of terms like "black and white" or "gray areas" typically emphasizes each in descending order, with absolutism and categoricity appearing most frequently. Overall, participants tend to see someone who has a "black and white" view of moral issues as someone who thinks that, for a given moral issue, there is an unambiguous fact of the matter about whether the act in question is right or wrong, that this fact is insensitive to contextual consideration, and they are not receptive to considering other people's point of view. For instance, a person who displays a black and white moral stance towards abortion may think that *all* abortions are wrong

regardless of the circumstances (such as incest or threats to the mother's health), and they have no interest in hearing excuses or rationalizations or arguments for why at least some abortions might be justified; roughly, their view is simply that "abortion is always immoral, *period*."

## Direct relativism

The most common theme was *descriptive*. This simply refers to *descriptive relativism*, the empirical observation that different people or groups have different moral standards. As noted in the main text, participants may conflate metaethical relativism (stance-independent moral facts are true or false only relative to different evaluative standards), and descriptive relativism. This theme highlights how such a conflation could readily emerge even when presented with a descriptive of metaethical relativism. Note that John did not make any direct empirical claims such as "different people and societies have different moral beliefs." Instead, John said:

> John: *"There is no single standard of moral truth. Different societies must be judged by different moral standards".*

Nevertheless, 39.0% (*n* = 62) participants were coded as expressing the descriptive theme. This is exactly what we'd expect if ordinary people struggled to distinguish metaethical and descriptive considerations. Numerous examples illustrate the tendency for participants to interpret expression of metaethical relativism as simple descriptive claims about variation in moral belief:

> *I think he means that different people have different way of thinking about standards. Not everyone adheres to the same moral standards.*

> *That different people have different moral standards.*

> *That morality is judged differently and viewed differently by different societies.*

One surprising finding was the comparative lack of reference to tolerating or respecting other cultures. Despite the connection between moral relativism and tolerance (Collier-Spruel et al., 2019), only 3.1% (*n* = 5) participants referenced tolerance:

*I think he means that you have to be tolerant of other cultures and understand that they live their lives by different standards.*

*That a person's morals and values are shaped by the society they live in and other societies cannot fault them for that.*

Given the close association between relativism and tolerance, it seems reasonable to expect more people to reference tolerance or abstaining from judging or imposing our standards on others without prompting, yet participants in this sample did not do so.

### Disagreement realism

Only two themes stood out in this condition. First, many participants restated what John said in various ways without expressing any substantive interpretation, as indicated by the theme *repeat*, a theme that emerged in 39.5% ($n = 51$) of responses. For example:

*In a discussion, no more than one person can be right.*

*That only one person can be right in a moral issue disagreement.*

This may highlight a shortcoming with the item in question; it may be that insufficient context or information was given, and that participants had trouble understanding what *else* I was asking. Yet 20.9% ($n = 27$) of participants described John's view as *black and white*. Note that viewing morality as black and white *does not* indicate moral realism. A moral antirealist could be just as rigid, absolutist, pigheaded, and judgmental as a moral realist. Such attitudes are conceptually orthogonal to one's views on the nature of moral of moral truth.

### Disagreement relativism

The *descriptive* theme was also the most common theme for the disagreement relativism condition, at 19.2% ($n = 29$) of responses. This unintended interpretation is again unsurprising if people don't have a clear conception of metaethical relativism. A handful of participants indicated that what John meant was that people were correct *according to their own perspective*, a theme reflected by the category *right to*. It is clear on examination that these responses do *not* express metaethical relativism:

*Each can see themselves as moral.*

*In their own mind each person may be right according to their moral compass.*

*He means that each side thinks they are correct. A moral standard is individual. One may think they are correct based on what their standards are, so how can they actually be incorrect?*

These views are distinct from relativism. Relativism is the view that people are correct relative to their standards. But *right to* more properly conveys the notion that different people *think* they are correct, and is thus better construed as an epistemic rather than metaphysical position. This was not especially common at 7.9% ($n = 12$), yet it illustrates one of the more subtle ways people could conflate relativism with other concepts, and may even represent a way people could seem to understand relativism even when they do not.

A handful of participants even used explicit metaethical language, including "subjective" and "relative." Some of these responses were spot on:

*I think John means that morality is subjective and that if one person things something is moral and another thinks it is immoral then those individuals can both be correct because morality is subjective.*

However, even when people use explicit metaethical language, they sometimes include additional remarks that indicate that they are using these terms in ways that don't match their meaning in contemporary academic metaethics. For instance, one participant stated:

*I think he means that perceptions of morality can be subjective and defined differently for different people. From each person's point of view they are both correct.*

Note their emphasis on something being true or false *from* different points of view, e.g., Alex believes X is correct, and Sam believes not-X is correct. This is not an indication that they *are* correct, just that they *consider* themselves correct. This *is not* relativism. In some cases, they use explicit language yet the rest of their remarks make it hard to interpret what they mean:

*There is no one true answer to a tough question. Sometimes the path to morality is subjective unless it harms someone else.*

This person indicates that there is no true answer, which is inconsistent with subjectivism, and that something is subjective "unless it harms someone else," which seems to impose constraints on subjectivism that make it unclear whether this conveys a genuine understanding of subjectivism or some sophisticated hybrid conception of metaethics. Regardless, the surrounding remarks render this comment too ambiguous to clearly categorize as an intended or unintended interpretation.

### S4.7.7 Study 3: Thematic analysis

The most common themes can be seen in **Figure S4.13** and **Table S4.7**.

**Figure S4.13**

*Most common themes for Study 3 (by condition)*
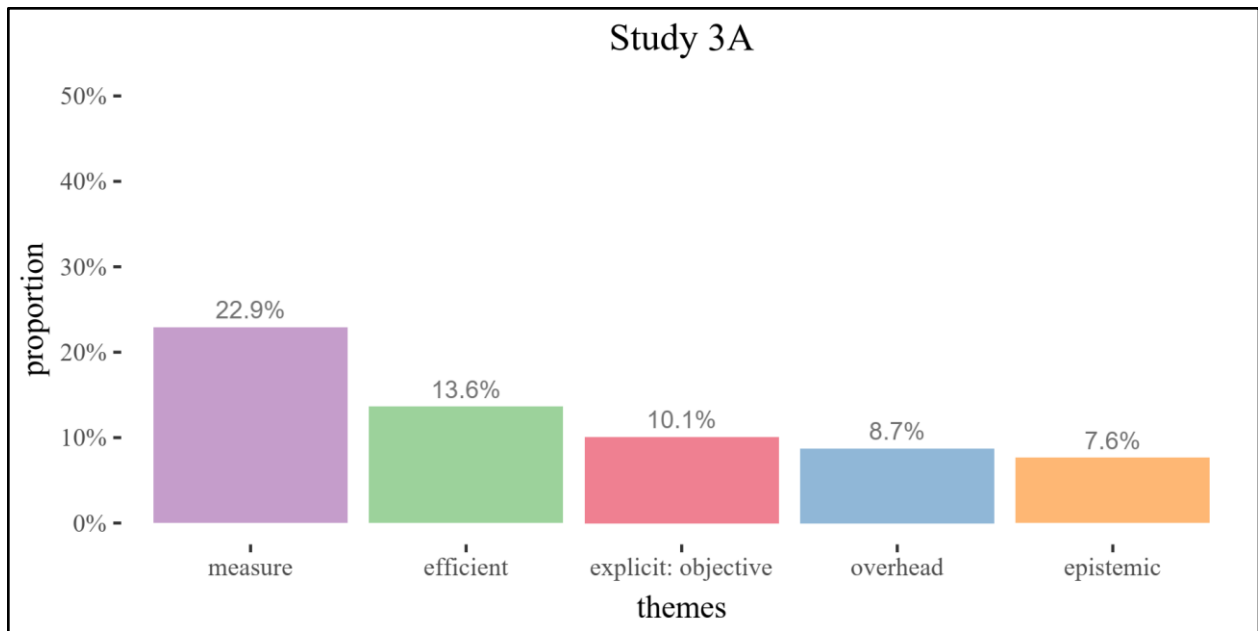
Study 3B
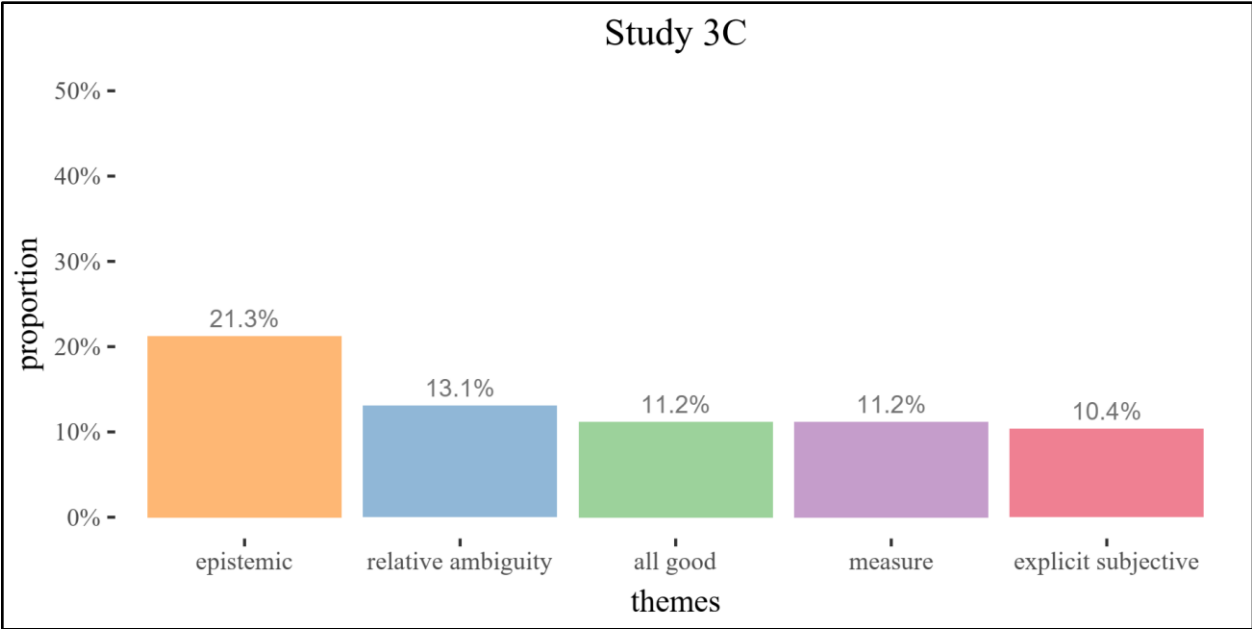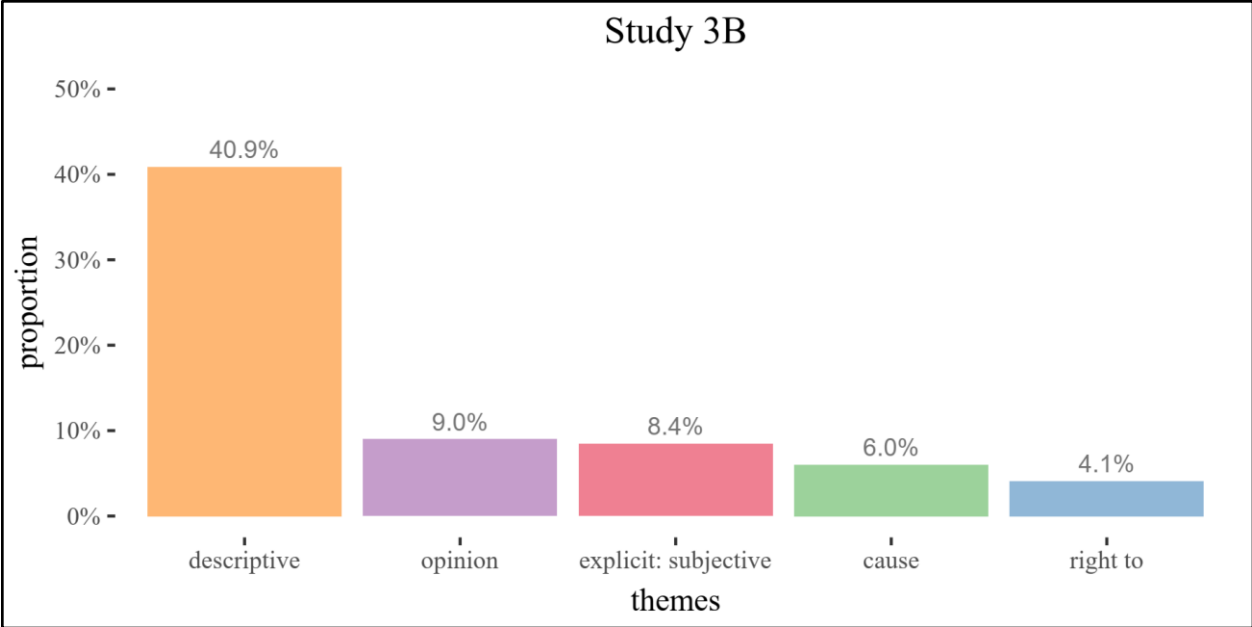


Study 3C

**Table S4.7**

*Most common themes for Study 3 (by condition)*

*Table S4.7.1 Study 3A: Realism*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *measure* | The view that something is "objective" if it can be measured or quantified | 22.9% | 84 |
| *efficient* | The view that some charities are more efficient than others at helping people or managing finances or resources | 13.6% | 50 |
| *explicit: objective* | Explicit use of term "objective" (or related term, e.g., "objectivism") | 10.1% | 37 |
| *overhead* | Reference to the overhead costs of charitable causes | 8.7% | 31 |
| *epistemic* | General comments or appeals to epistemic considerations | 7.6% | 28 |

*Table S4.7.2 Study 3B: Relativism*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 40.9% | 150 |
| *opinion* | The view that something is a matter of opinion | 9.0% | 33 |
| *explicit: subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 8.5% | 31 |
| *cause* | The claim that which charity is best depends on the cause of the charity | 6.0% | 22 |
| *right to* | Each person considers their standards to be correct | 4.1% | 15 |

*Table S4.7.3 Study 3C: Noncognitivism*

| Theme | Explanation | Percentage | Frequency |
|-------|-------------|------------|-----------|
| *epistemic* | General appeals to epistemic considerations | 21.0% | 78 |
| *relative ambiguity* | Conflating the claim that there are no moral/normative facts with the claim that moral/normative facts are relative/subjective, or any instance in which a response is ambiguous between relativism and noncognitivism/nihilism | 13.1% | 66 |
| *all good* | The claim that all charities do some kind of good | 11.2% | 41 |
| *measure* | The view that something is "objective" if it can be measured or quantified | 11.3% | 41 |
| *explicit: subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 10.4% | 38 |

**Realism condition**

*Measure* was the most notable theme to emerge in the realism condition. This theme refers to any instance in which the participant made reference to our ability to measure or quantify the impact of different charities in order to assess their effectiveness.

> *There is some standard to measure the good it does, and by that standard it does more good than others.*

> *It means that there is an objective measure of how much good a charity does.*

> *It means that there is some kind of objective measure that shows which charity does more to improve the world and/or society, and you can determine which charity thus does the most good according to that metric, which isn't based on opinion but is quantifiable.*

> *It's stating that there's statistical or objective evidence that can be presented to "prove" that some charities do more good, rather than being a subjective statement.*

Note that none of these responses indicate an interpretation that suggests they understood the statement to convey that there are stance-independent moral facts. Rather, they seem to interpret the original question in a completely understandable way, but one which is unrelated to the sense in which moral realists tend to regard moral facts as "objective." A realist interpretation of the claim that "it is a fact," in the realist sense, "that some charities do more good than others, not a matter of personal beliefs and values," would entail that there are facts about which charities are better than others that are not *made true* by the beliefs or values of individuals or groups. This has nothing to do with whether what the charities do is *measurable* or *quantifiable*. After all, many moral realists believe that moral facts can be ascertained *a priori* or even that they are *self-evident* or that we acquire knowledge via moral perception or moral intuition; knowledge of stance-independent moral facts is not typically (or, to my knowledge, *ever*) acquired by measuring or quantifying the facts in question That is, stance-independent moral facts just aren't the sorts of things one needs to measure or quantify.[203]

Yet there is a commonsense understanding of "objective" that means something like "subject to public evaluation" and for which there are precise means of quantification. For instance, one person might report that "it seems to be about 25 C°" and another might say that "it seems to be 20 C°," they might resolve this disagreement by looking at a thermometer. Perhaps the thermometer indicates that it's 23 C°. If so, the thermometer provides an "objective" measure of what the temperature is, where this is understood to mean some sort of quantifiable, measurable standard of what the temperature is that isn't contingent on and subject to the evaluations of a particular individual, who may be subject to some bias or error do to private error, and so on. Roughly, then, this conception of something being objective is, in some ways, similar to the notion of stance-independence, but it is not a *metaphysical* thesis, it is an *epistemic* one. This conception of objectivity presupposes that there is some

---

[203] Though once we know what the moral facts are, *if* they are natural facts, measurement or quantification may become relevant, e.g., utilitarians may engage in some type of measurement or calculus, but the fact that morality is about maximizing utility is not *itself* subject to measurement or quantification.

fact of the matter, then holds that there may be some criteria or method that can resolve the issue via some sort of third-personal standards or criteria that are publicly accessible and not reducible to the private judgments of individuals. This is *close* to a notion of stance-independence, but it is not the same thing. This is because such standards can be intersubjectively stipulated, and this can result in the relevant facts about which there are some measurable/quantifiable criteria being little more than *descriptive* facts. For instance, if we agree that our standard of evaluation is "number of quality adjusted life years (QALYs) added per dollar spent," then we could in principle provide a precise measure of this in unambiguous, quantifiable terms. We could, for instance, discover that there are two charities:

Charity A: $50 per 1 QALY

Charity B: $7 per 1 QALY

We can then say that Charity B produces more QALYs per dollar, and is therefore superior by this metric. Yet this would not entail that there are any stance-independent normative facts about which charities are "better," since this would represent little more than a conditional (or non-categorical) conception of a charity being better. That is, we could only say that if we use this standard (QALYs per dollar), *then* there are facts about which charities are better or worse according to this standard. But it would not follow that this is the correct standard, independent of our goals, standards, or values. Such stipulative, intersubjective standards may be perfectly appropriate for practical purposes, but they do not entail the existence of stance-independent moral facts. As such, such responses typically resulted in a clear unintended or unclear interpretation of the question, rather than a clear intended one, though on occasion participants included additional comments that suggested a stance-independent reading of the question.

Note that the *measure* theme emerged in around half of the participants who provided a clear unintended interpretation. This is a remarkably consistent tendency for people to interpret a question in an unintended way. Notably, given this particular item's emphasis on the relevant facts being true

regardless of people's personal beliefs or values, it is arguably an especially face valid representation of realism, as well, and yet participants *still* didn't interpret as intended.

However, the most important insight about the frequency of the *measure* theme, is that it did not make it into the top five for any other attempts to assess interpretations of realism. This illustrates that the particular ways in which participants can interpret statements about realism in unintended ways can be highly context-specific and vary across conditions. This is itself evidence of the highly context-specific way people interpret questions; they attend, rightly so, to contextual considerations that may be relevant to that specific formulation that aren't present in other situations. This is a perfectly sensible, natural way for people to think that will tend in practice to be highly effective. And it drives home the problem with a great deal of research in moral psychology, including research on metaethics: researchers operate under the mistaken presumption that one can readily solicit meaningful judgments about what something means, or prompt people to understand a question or set of instructions, after stripping away all or most contextual information. Yet this isn't how people think about moral issues in the real world. Real-world moral judgment is deeply embedded in the social contexts in which such judgment and reasoning occurs, and real moral judgments occur in a variety of contexts in which a rich panoply of situational factors may not merely be incidental or tangentially relevant to what the utterance means, or what the person engaging in moral reasoning is thinking, but *constitutive* of such judgments. In short, a great deal of conventional moral psychology may suffer a serious deficit in ecological validity (Gaesser, Campbell, & Young, 2022; Navarro-Plaza et al., 2020; cf. Holleman et al., 2020; Lewkowicz, 2001).

### Relativism condition

The most common theme in the relativism condition was, unsurprisingly, *descriptive*. This is unsurprising because people seem to struggle to distinguish between the notion that moral facts can be true or false relative to different moral standards, and the more mundane descriptive claim that

people have different moral beliefs. Like many other conditions that express moral relativism, participants once again interpreted a statement intended to reflect relativism, understood as a metaethical stance, in descriptive terms. Examples include:

*Each person places different value on what "good" means, so this will differ according to different people.*

*Everyone has their own perspective, and some people are more into some causes then others.*

*What you consider important will determine what you believe is the "most good"*

None of these statements indicate that moral facts are true or false relative to these different moral standards, they simply state that people have different moral standards. Such descriptive facts are orthogonal to and unrelated to whether moral realism is true, and suggest that participants did not understand that the question was about the truth status of moral claims, not an empirical claim about the existence of moral disagreement.

The most common theme in the noncognitivism condition was *epistemic*. This is a general category that captures any interpretations that appeal to epistemological concerns. Such concerns typically indicate a clear unintended interpretation, since whether there are stance-independent moral facts is a metaphysical claim[204] that does not depend on epistemic considerations.[205] For comparison, whether there is a stance-independent fact about whether there was ever life on Mars does not depend on whether we possess the tools to know whether this is true. It may be that there was life, but no trace of that life is available to us. Nevertheless, there would still have been life on Mars, even if nobody will ever know. Just the same, whether there are moral facts does not depend on whether we

---

[204] However, Scanlon and Parfit articulate realist accounts that purportedly lack metaphysical commitments (see (Veluwenkamp, 2017). Even so, this is consistent with the epistemic theme generally reflecting unintended interpretations of questions or stimuli related to metaethics.

[205] Though some moral realists may argue that moral realism is only true if we have access to these moral facts. Again, however, whether there are stance-independent moral facts is orthogonal to epistemic considerations, even if epistemic access is a necessary condition to earn the honorific label of a "moral realist" account. That is, the moral realist who includes an epistemic access condition would still regard a metaethical position which holds that there are stance-independent moral facts, but we have no way of accessing them, to be a view that holds that there are stance-independent moral facts; this is just a tautology and cannot be denied; they just wouldn't *label* these as "realist" accounts. This is little more than a terminological difference, not a substantive philosophical one (Sinnott-Armstrong, 2009).

can *know* that there are moral facts. Nevertheless, many participants interpreted the question in the noncognitivism condition to be concerned with epistemic considerations. For instance:

> *There are no statistics whatsoever to back up which charities do the most good. So there's know way of knowing which charities are best.*

> *You cant really tell which charity does the most good until you really visit them and research them fully.*

> *No one can be absolutely sure which charity does the most for people, it's all an opinion and hard to justify*

Note that none of these responses indicate that moral considerations with respect to which charities do more good are irresolvable because there is a way for such claims to be true or false. Rather, such reactions take the question to suggest that it is difficult to know whether one charity is better than another. If anything, this implies that there *is* some fact of the matter, but that it's hard (or impossible) to know what that fact is. What it does not indicate is that there is no fact at all. Such interpretations simply do not reflect any recognizable *metaethical* interpretation.

Although it only occurred in 8.4% ($n = 66$) cases, *relative ambiguity* is a noteworthy theme that appeared only in this particular dataset. Such participants seemed to conflate the notion that there is "no fact of the matter" with *subjectivism*. Yet subjectivism does hold that there are facts of the matter about whether moral claims are true or false; it simply regards them as true or false relative to the standards of different individuals. Nevertheless, many participants interpreted this question to reflect subjectivism:

> *It means the same thing as above, that based on what you think is most important makes it the most good so there's not one specific answer*

> *that it's a subjective opinion and it depends on the person.*

> *Charities effectiveness is dependent on personal beliefs and not facts.*

> *There is no objective basis for which charities are doing the most good. There is a subjective basis for determining how well charities are doing.*

Note the use of "subjective" and cognitivistic language, such as "belief," and even the claim that there is a "subjective basis for determining how well charities are doing," none of which is consistent with a noncognitivist reading of the notion that there is no fact of the matter; indeed, it even suggests there can be facts of the matter, just not a single, objective fact of the matter. While this may suggest that people conflate noncognitivism with subjectivism, I suspect the problem may be due at least in part to researcher error in how I operationalized noncognitivism. To say that there is no "fact of the matter" could be interpreted to mean no *single* fact of the matter, or no *stance-independent* fact of the matter, which could have unintentionally prompted some participants to interpret the question to reflect relativism/subjectivism. As such, this particular conflation may be due to poor phrasing in the question, rather than an inherent difficulty ordinary people have with distinguishing noncognitivism from relativism/subjectivism.

**Noncognitivism condition**

*Epistemic* was by far the most common theme, at 21.0% (*n* = 78). Many participants appeared to interpret the idea that there is no fact of the matter as the idea that we have *no way to know* which charity was better:

> *There is no proof that one charity is better than another.*
>
> *there is no way to tell which charity is better*
>
> *There's no way to determine which charities do the most good.*

This is *not* what philosophers mean when they claim that there are no facts about a given issue; they mean that there is literally no fact, not that there is a fact but we just can't know about it. Others conflated noncognitivism with relativism, which was the second most common theme, *relative ambiguity*. This may be due in part to researcher error. Part of the issue is that I took "no fact of the matter" to mean that, literally, there was no fact, but participants may have interpreted this as the notion that there is no *stance-independent* fact of the matter, or a fact about which there is a stance-

independent fact if we agree on a particular set of evaluative standards (which is a little different than a more straightforward stance-independence). However, it's not clear whether this should be characterized as researcher error or precisely the kinds of ambiguities that make adequate specification of a metaethical concept difficult. This theme is likely related to the fifth theme, *explicit: subjective*.
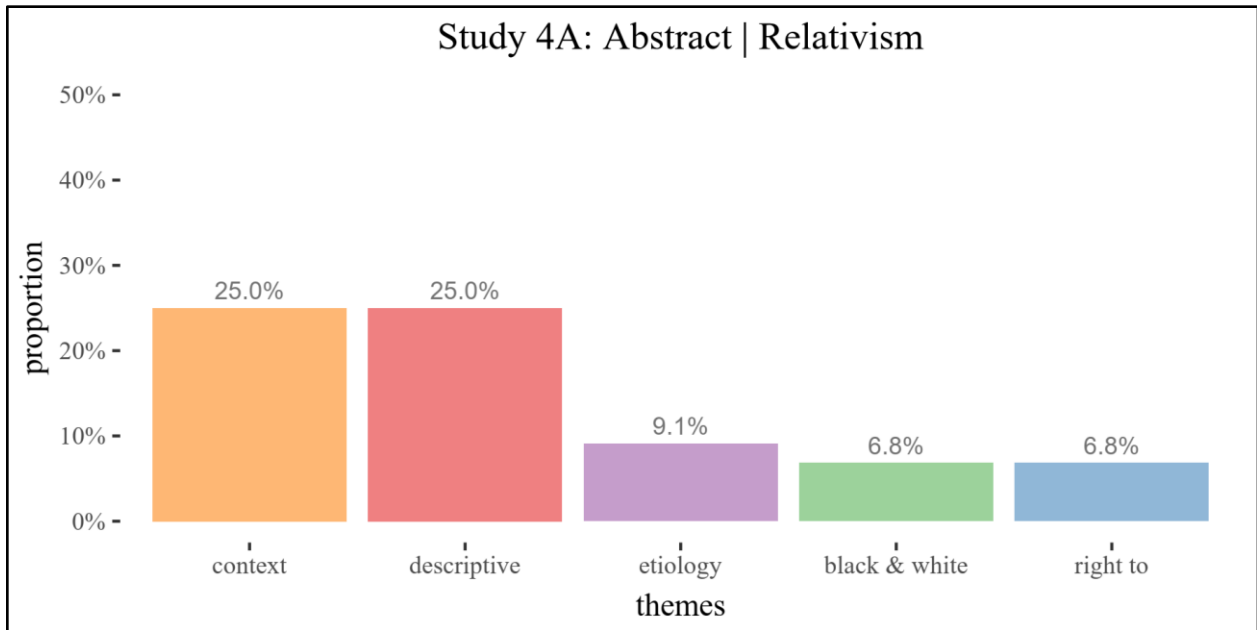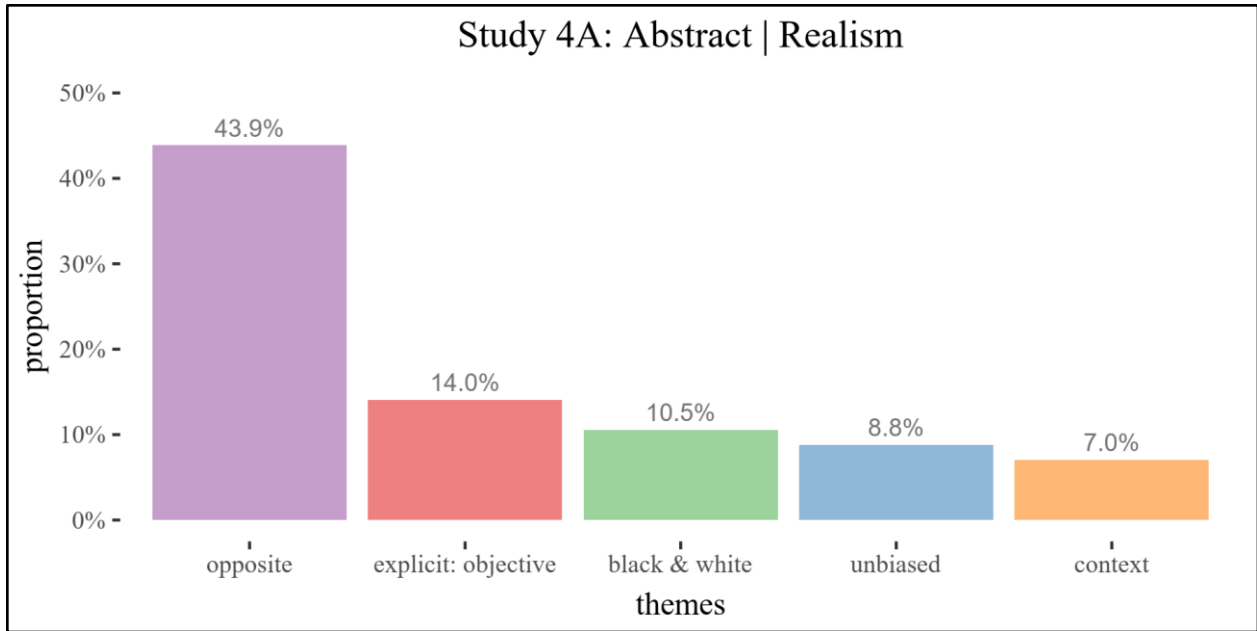
Finally, *measure* was a common theme in this case, likely owing to the use of a concrete issue such as charities that specifically deals in quantifiable outcomes (e.g., number of lives saved) and handling money. Interestingly, this theme highlights one of the ways ordinary people's thinking about an issue being "objective" seems to differ from what philosophers typically have in mind. Many people appear to think that a matter is objective to the extent that we can employ some public or quantifiable standard of evaluation to the matter. For instance, if we have a ruler, or a scale, or a thermometer, these measures provide "objective," and "quantifiable" measures of whatever it is that's being measured. This is *not* the same thing as objectivism in the sense of being stance-independent. Rather, it seems to be a matter of judging that *if* there is some intersubjective agreement on a standard of evaluation, *then* whatever it is we're evaluating can be judged according to that standard in a way that is "objective," in some respect (e.g., quantifiable, unbiased, and so on; it's not clear what exactly people have in mind). Notably, even if such a standard were regarded as stance-independent, this still wouldn't resolve whether there were any stance-independent facts requiring us to apply or conform to this standard. It could simply be a matter of intersubjective agreement. In which case what we might be dealing with is a rudimentary form of folk constructivism, *not* realism. I'm skeptical we can go so far as to attribute proto-constructivism to ordinary people. Like metaethics more broadly, I suspect people have no determinate stances or commitments on this matter, either, and that it is likewise indeterminate.
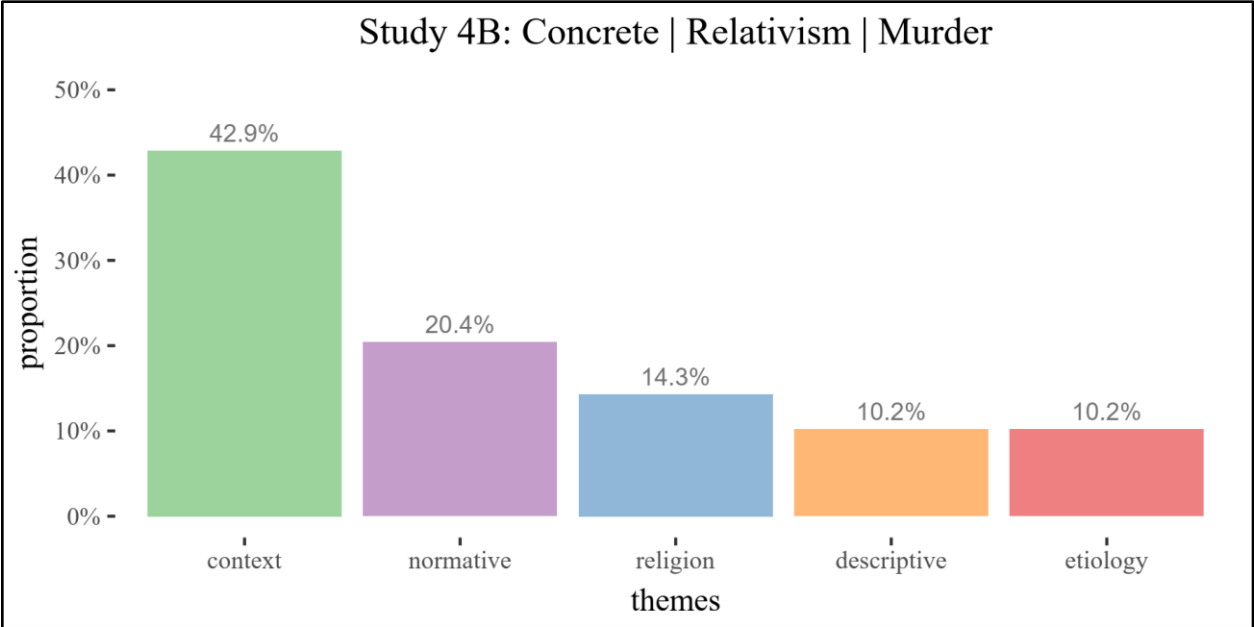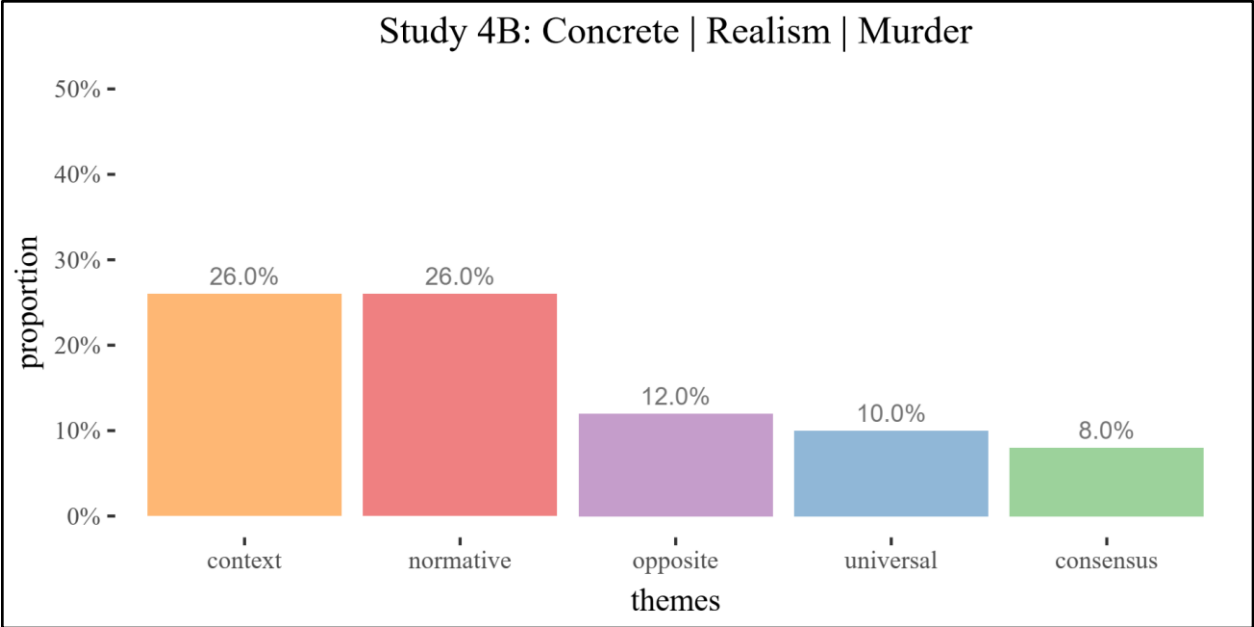
### S4.7.8 Study 4: Thematic analysis

The most common themes can be seen in **Figure S4.14** and **Table S4.8**.

**Figure S4.14**

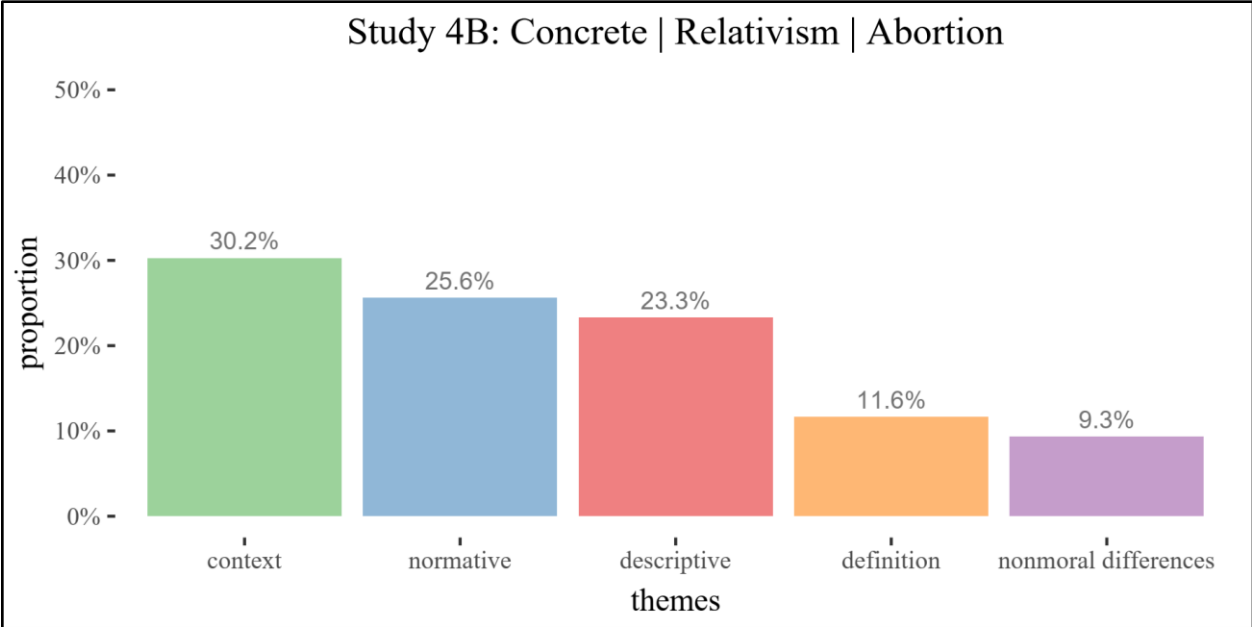*Most common themes for Study 4 (by condition)*

Study 4B: Concrete | Realism | Murder



Study 4B: Concrete | Relativism | Murder

Study 4B: Concrete | Realism | Abortion



Study 4B: Concrete | Relativism | Abortion
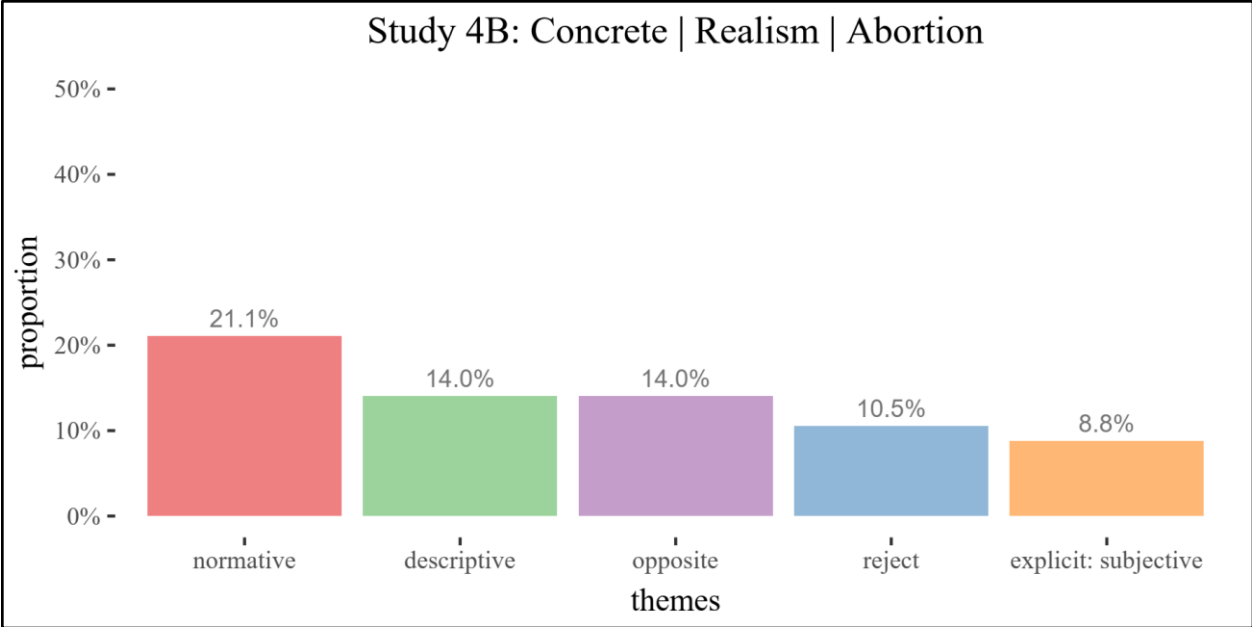
**Table S4.8**

*Most common themes for Study 4 (by condition)*

*S4.8.1 Study 4A: Abstract | Realism*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *opposite* | Participant describes the opposite/contrary metaethical view (e.g., the question is about realism, and they describe antirealism) | 43.9% | 25 |
| *explicit: subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 14.0% | 8 |
| *black and white* | Explicitly describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality | 10.5% | 6 |
| *unbiased* | View that something is "objective" when it is unbiased / impartial | 8.8% | 5 |
| *context* | The view that whether an action is right or wrong depends on context/circumstances | 7.0% | 4 |

*Table S4.8.2 Study 4A: Abstract | Relativism*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 25.0% | 11 |
| *context* | The view that whether an action is right or wrong depends on context/circumstances | 25.0% | 11 |
| *etiology* | Offered a causal account of how the person would respond as they did | 9.1% | 4 |
| *right to* | Each person considers their standards to be correct | 6.8% | 3 |
| *black and white* | Explicitly describes morality as "black and white" or mentions "grey areas." Typically | 6.8% | 3 |

conveys a rigid, absolutist, or inflexible approach towards morality

*Table S4.8.3 Study 4B: Concrete | Realism | Murder*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 26.0% | 13 |
| *context* | The view that whether an action is right or wrong depends on context/circumstances | 26.0% | 13 |
| *opposite* | Participant describes the opposite/contrary metaethical view (e.g., the question is about realism, and they describe antirealism) | 12.0% | 6 |
| *universal* | Normative position which holds that a given moral norm applies to everyone | 10.0% | 5 |
| *consensus* | View that a moral claim is objective when most/everyone agrees about it | 8.0% | 4 |

*Table S4.8.4 Study 4B: Concrete | Relativism | Murder*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *context* | Disagreement attributed to different assumptions about the circumstances in which the action was performed | 42.9% | 21 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 20.4% | 10 |
| *religion* | Refers to religion or religious beliefs | 14.3% | 7 |
| *etiology* | Offered a causal account of how the person would respond as they did | 10.2% | 5 |
| *descriptive* | The view that different people or groups have different moral standards | 10.2% | 5 |

*Table S4.8.5 Study 4B: Concrete | Realism | Abortion*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 21.1% | 12 |
| *opposite* | Participant describes the opposite/contrary metaethical view (e.g., the question is about realism, and they describe antirealism) | 14.0% | 8 |
| *descriptive* | The view that different people or groups have different moral standards | 14.0% | 8 |
| *reject* | Participant expresses disagreement with stimuli | 10.5% | 6 |
| *explicit subjective* | Explicit use of term "relative" (or related term, e.g., "relativism") | 8.8% | 5 |

*Table S4.8.6 Study 4B: Concrete | Relativism | Abortion*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *context* | The view that whether an action is right or wrong depends on context/circumstances | 30.2% | 13 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 25.6% | 11 |
| *descriptive* | The view that different people or groups have different moral standards | 23.3% | 10 |
| *definition* | Conflates metaethical considerations with issues related to definitions | 11.6% | 5 |
| *nonmoral differences* | Conflates metaethical considerations with differences in nonmoral beliefs | 9.3% | 4 |

### Abstract | Realism

One of the most puzzling themes to emerge in any of the thematic analyses conducted here was the unusual occurrence of the *opposite* theme. A response was coded as *opposite* whenever a clear intended interpretation would reflect a specific metaethical stance, but the participant interpreted the stimuli in a way that conflicted with or even reflected the *opposite* understanding of the relevant metaethical stance; e.g. interpreting the claim that moral truth is "relative" to mean that there are stance-independent moral facts, or to interpret the claim that moral truth is "objective" to mean that there is no moral truth, or that it is true or false relative to the standards of people or cultures. For whatever reason, this occurred 24.3% (*n* = 25) of the time in the abstract|realism condition. My initial reaction was to suspect that I had mislabeled the conditions, and that responses coded for the abstract|realism condition were actually responses to a relativist condition. However, this does not appear to be the case. Several examples explicitly refer to or repeat portions of the question itself, and make use of the term "objective" when doing so. Such cases unambiguously illustrate that participants were responding to the realism condition:

> *Moral truth is objective because what is moral can vary from person to person, across different societies across the world.*

> *Moral truth is objective because everyone has different views on what is moral or not. What might be immoral to one person might be perfectly moral to another person.*

> *Moral truth is objective because there are many morals that are dependant on culture, life experience and religion. Since we don't all share the same cultures and religions, moral truths will vary. For example, to a devout Catholic, abortion is morally wrong. However, to an atheist, abortion may be acceptable in some situations.*

> *Morals are right and wrong. Some people may think the wrong thing is the right thing. This is objective. It depends obthe person morals*

As these responses illustrate, participants clearly interpreted the claim that moral truth is *objective* in line with moral relativism and related concepts, such as descriptive relativism. It's unclear why this

occurred, though it may be that participants mistakenly thought that the term they were asked about was "subjective," or they correctly read the term "objective" but thought that it meant "subjective." Regardless of how this occurred (unless it was coding error!), it indicates that ordinary people, at least under conditions in which they are inattentive, may conflate or mistake the term "objective" for "subjective" or in some other way interpret "objective" in a way opposite to its intended meaning.

**Abstract | Relativism**

The most common themes in the abstract|relativism condition were consistent with the themes that typically emerge in other relativist conditions. Participants were most likely to conflate relativism with the claim that different people and societies have different moral beliefs 17.2% ($n = 11$) or to conflate relativism with the notion that whether something is right or wrong depends on the context, 17.2% ($n = 11$) which is consistent with nonrelativist positions and indicates a clear unintended interpretation about what relativism entails. Examples of unintended interpretations coded as *descriptive* include:

> *What one person find "right" or just, another may not.*

> *People have their own sets of values and beliefs which they base decisions of what is right and what is wrong off of.*

> *Different people have different morals given their different religions or lack of belief in any.*

Examples of unintended interpretations coded as *context* include:

> *It means that morality cannot be applied systemically and must take situational factors into account.*

> *To say that moral truth is relative is to say that morals change according to circumstances.*

> *That is depends on the situation. A person's actions and/or behaviors in certain situations may be deemed moral but in other situations, the same actions and/or behaviors may be deemed immoral.*

These remarks speak for themselves. These are not merely instances in which someone may not have interpreted the claim that morality is relative clearly enough for us to know how they interpreted it. Rather, these are unambiguous instances in which people interpret the notion that moral truth is relative in a perfectly reasonable, but is simply not researchers studying folk metaethics are asking

about. In this particular sample, clear unintended interpretations like these were slightly more than twice as likely as clear intended interpretations. If we can reliably identify double the number of instances in which people interpret stimuli in unintended ways compared to intended ones, this should give serious pause to any research employing such stimuli.

These are exactly the kinds of themes we should expect to frequently emerge if participants are not merely offering unclear responses, but if they are actively interpreting explicit references to moral relativism in unintended ways. Perhaps the most surprising result is absence of explicit references to respect or tolerance. Support for relativism is associated with increased moral tolerance (Collier-Spruel et al., 2019), and when references to relativism do appear in public exchanges, relativism is often conflated with or entangled with normative implications related to tolerance, respect, or non-interference.

### Concrete | Realism | Murder

Like the abstract realism condition, a handful of participants interpreted "objective" opposite to its intended meaning, to refer to moral relativism or descriptive variation in moral belief, though only 7.7% (*n* = 6) of participants did so. However, the most notable themes were *normative* and *context*, which tied for first place at 16.7% (*n* = 13) each. *Context* is exactly the kind of response we should expect when asking about "murder." While I describe this condition as a "concrete" one, in that it specifies a particular type of moral violation, the extent to which a moral issue is concrete is a matter of degree, with moral issues falling on a spectrum from completely unspecified (e.g., "someone did something immoral") to maximally detailed (e.g., comprehensive legal report documenting all details relevant to a crime that comprises dozens or even hundreds of pages, a recorded interview of the perpetrator, and so on). A simple reference to murder may be more concrete than talking about moral truth in the abstract, but simply referencing "murder" hardly fills in the details in a meaningful way. There are all sorts of events we might describe as murders, and we might regard some as clearly

immoral (such as a serial killer kidnapping and dismembering people for fun) while we might see others as morally justified (e.g., assassinating a crazed dictator who is about to order the genocide of an entire ethnic group). The details of the so-called "murder" are relevant. Without such details, people may be reluctant to state whether there is an objective truth about whether "murder is morally wrong," not because they deny that there are stance-independent moral facts, but because their response essentially boils down to, "it depends." This is exactly how some people responded:

> *Murder can be morally wrong, but it also cannot be. It will depend on the circumstance of the case and a lot of variables.*

> *I think it means that killing people is wrong no matter the circumstance, I don't don't agree that it's objective though.*

> *The objectivity comes under the circumstances of the murder, and what is the person's definition of murder.*

> *I think the statement is objective because there are ways to murder with cause that is not morally wrong (war, defending self, etc.).*

Many participants struggle to distinguish between some general category of action such as murder being morally acceptable in some cases but not others with the notion that whether it is moral or immoral depends on the standards of people or cultures. Normative considerations likewise show that participants struggle to disentangle normative from metaethical considerations:

> *that murder is a bad thing*

> *Murder IS morally wrong, it's taking the life of someone who has more life to live. It's taking their right to live their full life.*

> *In my own words, murder IS morally wrong. There is no objection when it comes to it, a human should not take the life of another human no matter what the circumstances are. I personally do not believe in the death sentence due to this.*

> *To say that denying someone else's equally valid claim to life as that of the murderer would be morally wrong.*

None of these responses have anything to do with metaethics. They simply reiterate the first-order moral claim that murder is bad. While no predictions were made about what the most common themes

would be, these results are unsurprising. Note that the *normative* interpretation occurs whenever a participant is inclined to cast judgment on someone else or express a first-order moral judgment. I argued in **Chapter 2** that participants may conflate metaethical and normative considerations for a variety of reasons, including a desire to signal their moral opposition to the act in question. If, as I maintain, morality primarily serves a variety of sociofunctional purposes associated with promoting one's welfare and the welfare of allies at the expense if one's competitors, we should expect moral judgments to be largely oriented around a motivation to appear good and virtuous and to denigrate outgroups, *not* to conform to logical consistency with respect to abstract normative moral principles or to convey a coherent set of metaethical commitments. Just as ordinary usage of math is best understood in light of its practical use, and not in terms of its metaphysical entailments, moral is a social technology that facilitates the practical ends of individuals and groups, it is not about conforming to philosophical ideals. Ordinary people want to have a nice meal, get laid, and be popular enough that nobody is motivated to bash their skull in with a rock while they sleep; they're not especially interested in appeasing the ghosts of Plato and Aristotle.

### Concrete | Relativism | Murder

*Context* and *normative* were also the most common themes in the relativism version of the concrete murder condition, though in the concrete|relativism|murder condition context was the decided winner, coming in at 29.2% (*n* = 21) responses, while only 13.9% (*n* = 10) were coded as *normative*. Once again, these are precisely the kinds of interpretations we should expect if people likewise do not interpret direct references to relativism as intended. Many participants interpreted "relative" to mean that whether murder is right or wrong depends on the circumstances, a response that makes far more sense for "relative" than for "objective":

> *I believe that means that murder is morally wrong only dependent on the situation and the context in which the murder takes place.*

> *It means what type of murder. Is self defense murder? Is death penalty murder?*

> *it means that murder can be more or less "wrong" depending on other circumstances*

Such responses suggest that people do not interpret explicit references to moral truth being "relative" as intended. The same holds for *normative* responses:

> *Murder is not something a human being should do.*

> *Murder is wrong! Yes its morally wrong. It breaks one of Gods 10 commandments. To me, its the worst sin we can commit.*

> *Because we live by morals as a person,we are supposed to care and look out for each other not kill each other*

Such responses suggest that metaethical considerations were not salient when responding to the question, and that first-order normative considerations loomed large in how participants reacted to these questions. It is possible that if the metaethical nature of the question were made more explicit participants would pick up this and respond accordingly, so the fact that many participants expressed a normative attitude does not necessarily entail that they don't or can't think in metaethical terms, or that they have no determinate metaethical stances or commitments. However, it is consistent with the possibility that studies interpreted to measure folk metaethical views may require sufficient instructions to adequately prompt metaethical responses.

**Concrete | Realism | Abortion**

*Normative* reactions were also common in the abortion condition, comprising 13.3% (*n* = 12) responses:

> *its wrong if the baby has a hearbeat*

> *There are a lot of reasons why abortion is wrong. The strongest reason being it is truly murder of an unborn child and sometimes the child is born before they are dead. It's never right to take an innocent life.*

> *to be objective about abortion is to stand on the fact when you have an abortion you are killing someone*

Nothing about these responses indicates that these participants interpreted "objective" to refer to the notion that there is a stance-independent fact about whether abortion is immoral. Once again, a

handful of participants, 8.9% ($n = 8$) interpreted "objective" in a way opposite to its intended meaning, and a sizeable minority raised objections to the statement itself, a pattern of response fairly similar to expressing a normative judgment:

> *there is no truth in this statement. But, if I need to supply an answer, it would be that some believe that abortion is akin to murder.*

Notably, participants who rejected the statement did so *because* they interpreted as a question about metaethics, and were rejecting the metaethical presuppositions expressed by the claim:

> *Are moral claims actually objective? The statement is subjective*

> *To me, this is a subjective statement. Each person has a different set of moral values and it is up to them to view right and wrong as they choose.*

> *No, I think it's completely subjective. Often this claim is based on church authorities, usually celibate males who have no right to tell others what to do.*

This high rate of participants objecting to the claim itself is not hard to interpret in light of the fact that many participants are likely to disagree that "abortion is morally wrong," while this is less likely in the case of murder and other moral issues for which there is broad consensus. That several participants made a point of objecting to the claim that abortion is wrong, but this did not occur in the murder conditions, supports the notion that the themes that emerge from the data reasonably reflect the sorts of themes we might expect to emerge given the content of the stimuli.

***Concrete | Relativism | Abortion***

Results for the concrete|relativism|abortion were likewise unsurprising, with the most common themes consisting of context 20% ($n = 13$) *normative* 16.9% ($n = 11$) and *descriptive* 15.4% ($n = 10$). Once again, this is consistent with precisely the kinds of unintended interpretations we would expect if people conflated relativism with non-metaethical considerations. Many participants interpreted the notion that relativism about whether "abortion is morally wrong" is true meant that whether abortion is right or wrong depends on the specific circumstances of the abortion:

> *It is relative because there are cases when it might be acceptable to have abortions (like when a woman is raped or the birth might kill her). It's not so black and white.*

> *That you have to consider other factors and you can't just say that all abortions are morally wrong.*

Still others took the opportunity to express agreement with the claim that abortion is morally wrong, which is irrelevant to what the question was about:

> *Because abortion is about selfishness rather than need.*

> *Because one is taking the life of another living being.*

> *Abortion is morally wrong because you are killing a life.*

Still others interpreted the question to be asking whether different people have different moral beliefs:

> *It means that what people perceive as right and wrong varies from society to society and even person to person.*

> *It means peoples morals are diffrent.*

Again, we see the same patterns emerge. People reliably construe the notion that morality is "relative" in a variety of ways, but frequently do so in ways that are difficult to circumvent in conventional studies on metaethics. When explicitly asked about moral relativism, people simply do not consistently interpret this in the way they would need for such questions to be valid measures of metaethical belief.

### S4.7.9 Study 5A: Thematic analysis

Thematic analysis was conducted for all ten items. Since there were too few responses per item, it made little sense to assess the frequency of themes on a per-item basis. Instead, results for the five most frequent themes were aggregated across all ten items. The most common themes can be seen in **Figure S4.15** and **Table S4.9**.

**Figure S4.15**

*Most common themes for Study 5A (aggregated)*



Study 5A: Explanation
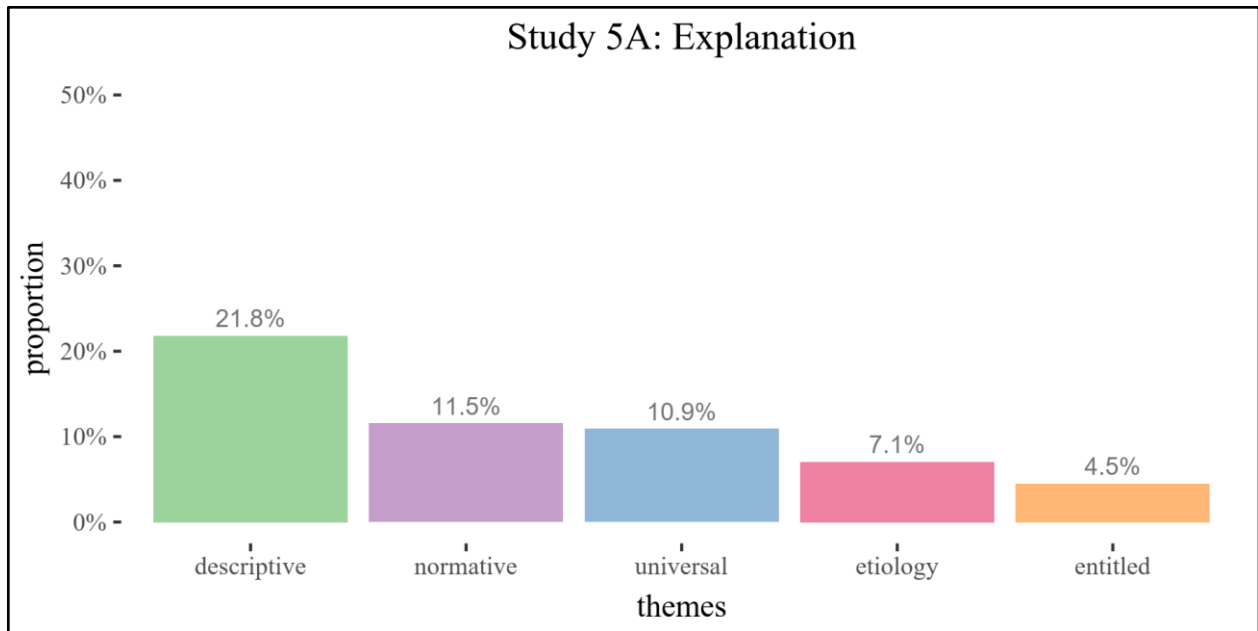
**Table S4.9**

*Most common themes for Study 5A (aggregated)*

| | | | |
|---|---|---|---|
| *descriptive* | The view that different people or groups | 21.8% | 34 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 11.5% | 18 |
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 10.9% | 17 |
| *etiology* | Offered a causal account of how the person would respond as they did | 7.1% | 11 |
| *entitled\** | The claim that people are entitled to their beliefs | 4.5% | 7 |
| *epistemic\** | General comments or appeals to epistemic considerations | 4.5% | 7 |
| *explicit subjective\** | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 4.5% | 7 |

| *personal** | Describes morality as a matter of personal belief | 4.5% | 7 |
|---|---|---|---|

Consistent with the results of many other studies, the most common theme to emerge by far was *descriptive*. Once again, participants frequently interpreted items intended to reflect relativism to instead reflect the mundane, non-metaethical descriptive claim that different people and cultures have different moral beliefs:

> *Different cultures can have different morals, for instance some muslim cultures believe that the woman should keep their faces covered while others don't.*

> *not everyones belifes are the same*

> *Each culture has its differences in regard to what constitutes morally right behavior.*

Note that in this case, participants were asked to explain *why* they agreed or disagreed with the item in question. Yet the existence of moral disagreement does not justify or explain why someone would endorse relativism. Moral realists readily acknowledge that people disagree about what is morally right or wrong; indeed, that is a large part of what motivates them to endorse moral realism! After all, if everyone shared their moral standards, there would be no need to go around advocating for moral realism in the first place, any more than one would need to go around advocating for the claim that water is wet. Think about how bizarre it would be for someone to deny that there are stance-independent moral facts *because* different societies have different moral standards. While, in principle, participants could be expressing, in some inchoate form, some kind of rationale for moral antirealism predicated on moral disagreement, which is indeed a position one could take, a more conservative interpretation of what participants are expressing is that they agreed with the item *because they interpreted the item itself to express descriptive claims about the existence of moral diversity* and *not* because they interpreted as an expression of a metaethical stance, for which moral diversity was offered as a justification for their view.

Typical of many studies, many participants made normative remarks:

> *everyone should treat everyone with love respect and care. its better for life*

> *We should have morals, or else we become animals.*

> *I strongly agree with this because it's what the right thing to do. If a muslim woman loves a jewish man, they have the right to be together regardless of what extreme muslim/jewish people believe.*

Such remarks suggest participants may have interpreted questions ostensibly intended to reflect metaethical positions to instead have various implications for how we should act, e.g., whether we should have moral standards, and how we should treat other people.

In a few cases, participants expressed the notion that morality is *universal*. In some instances, such responses were accompanied by enough of an explanation that they appeared to convey an intended metaethical stance. However, this was not always the case. Some responses either merely state that morality is universal, which, without qualification cannot be interpreted clearly, or conveyed descriptive claims that people in different cultures have (at least some) shared moral standards, neither of which adequately conveys an appropriate metaethical rationale for their level of agreement with such items. For instance:

> *I think most societies agree on a few basic rules where morals come into play.*

> *The moral belief that humans should not murder seems pretty universal. Stealing tends to be seen as morally wrong regardless of demographic as well.*

> *Different cultures have different ideas of what is right and what is wrong. But some things, such as murder and incest, are considered wrong in nearly all cultures.*

Just as variation in moral belief does not entail relativism, the universality of moral belief does not entail realism. After all, that most people like chocolate or French fries does not entail that there are stance-independent gastronomic facts about what food we should eat that don't depend on our preferences. Nor, for that matter, does thinking everyone *should* like chocolate or French fries entail that there is a stance-independent gastronomic fact. The same holds for moral standards. A moral

antirealist could think that "everyone should be honest" without being committed to some type of realism. They could simply express the preference that everyone be honest, or the imperative that everyone be honest, or in some other way express a stance towards what everyone ought to do without supposing that such people would be making a moral error if they did not.

### S4.7.10 Study 5B: Thematic analysis

The most common themes can be seen in **Figure S4.16** and **Table S4.10**.

**Figure S4.16**

*Most common themes for study 5B (aggregated)*



**Table S4.10**

*Most common themes for study 5B (aggregated)*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 20.5% | 32 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 15.4% | 24 |

| | | | |
|---|---|---|---|
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 11.5% | 18 |
| *epistemic* | General appeals to epistemic considerations | 6.4% | 10 |
| *personal* | Describes morality as a matter of personal belief | 5.8% | 9 |

Over a fifth of participants conflated items on the MRS with descriptive claims, a finding consistent with many other measures. In many cases, this is *all* people said, indicating a straightforward conflation:

> *That people can believe different things without one being right and the other wrong.*
>
> *Some people have different morals.*
>
> *something can be considered okay in one culture while not in another*

This reveals how, even in cases where items are purportedly validated by experts and exhibit a host of encouraging indicators of validity, many people *still* don't interpret items as intended.

The frequency of the *universal* theme is also no surprise, since all three reverse-coded items point to universalism, with two also including references to stance-independence, and one exclusively indicating universalism. The *universalism* theme appeared most frequently for these items, but since the number of responses per item is low it is hard to draw any firm conclusions.

The *normative* theme was also surprisingly common, emerging primarily for the three universalism/realism items, #6, #9, and #10. This is again unsurprising, given that #6 and #10 both use explicitly normative language with the use of "should." This suggests that, despite the content of these items referencing universalism and realism, that normative concerns loom large and tend to frequently serve as the central focus of people's responses. Note how people's emphasis in some cases shifts fully to normative considerations:

*Some actions are just wrong.*

*Means people of different cultures (races, religions, etc) should still behave and demonstrate 'goodness'.*

*Basic beliefs of just being kind to each other.*

### S4.7.11 Study 6A: Thematic analysis

The most common themes can be seen in **Figure S4.17** and **Table S4.11**.
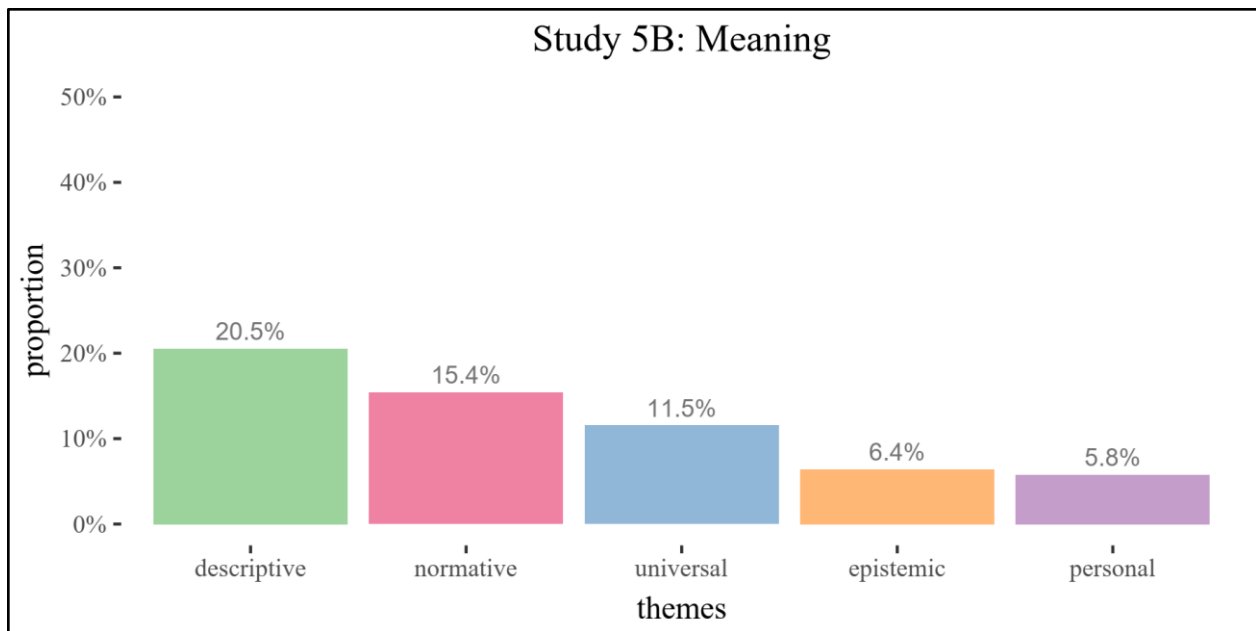
**Figure S4.17**

*Most common themes for Study 6A*
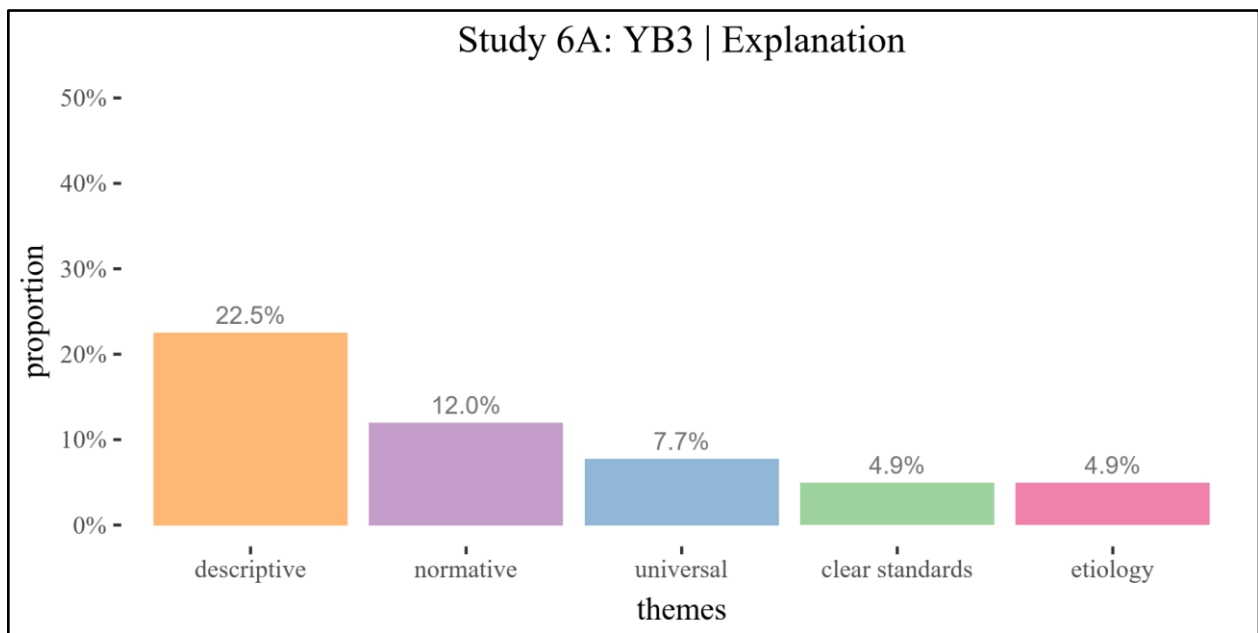


**Table S4.11**

*Most common themes for Study 6A*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 22.5% | 32 |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality | 12.0% | 17 |

| | | | |
|---|---|---|---|
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 7.7% | 11 |
| *etiology* | Offered a causal account of how the person would respond as they did | 4.9% | 7 |
| *clear standards* | The claim that moral standards are made clear (by e.g., a formal codification of ethics) | 4.9% | 7 |

Consistent with other scale items used to assess beliefs about relativism, participants frequently interpreted questions about relativism in descriptive terms. The most prevalent theme, by far, was *descriptive*: 15.9% (*n* = 32) of participants explained why they expressed the level of agreement they did with one of the three items by referencing differences in moral beliefs. While moral disagreement could be used as a reason in favor of a particular metaethical position, most responses did not explicitly express inferences of this kind. Rather, they appear to have simply interpreted questions to *just mean* that different people have different moral beliefs. Here are a handful of illustrative examples:

> *people have different beliefs*

> *Every persons ethics are different and beliefs as well. One person might not agree with the code of ethics and live that outside of the workplace.*

> *Because everyone has a different idea of what is moral or not.*

Imagine that you were asked to explain why you agreed that moral claims are true or false relative to the standards of different individuals or groups. Is this what you would say? That seems unlikely. This is far more plausibly a reiteration of what the participant took it to mean to agree with the item; that is, they agreed with the item *because* they interpreted it as an expression of descriptive relativism. This is the most reasonable way to make sense of the *descriptive* theme also being the most common theme when participants were asked to explain what the statement means, which was slightly higher, at 19.4% (*n* = 37; see below). Like other items conveying relativism, the normative and universal themes were

also quite common, and etiology once again made an appearance. However, the unique theme *clear standards* deserves special consideration. All instances of this theme occurred for a single item: item #1, and did not occur for any other items in any other scales or measures. This is notable in part because it supports my contention that the themes I've identified in the open response data are not simply the top-down imposition of my own expectations, but organically emerge on an ad hoc (in a good way!) basis to accommodate local idiosyncrasies. Consider item #1: *There are no ethical principles that are so important that they should be a part of any code of ethics.* The *clear standards* theme involves remarks about how there may be a variety of practical benefits to formally codifying a set of moral standards. One can imagine a society which has accumulated a host of rules and conventions. Eventually, that society may find it expedient to formalize those rules in an official code of conduct. Such events have famously occurred in both the ancient world through the present. It does not take a degree in history for people to be aware of such cultural watersheds: Hammurabi's Code, the Ten Commandments, the Magna Carta, and so on have all worked their way into popular consciousness, and we are all familiar with lists of rules. Many participants appeared to interpret this item to refer to something like the notion that denial that there could be some rule that is so important we ought to put it on any codified set of moral rules. Consider some of the replies indicative of this interpretation:

*I believe ethic codes make standards more clear.*

*I think that in order to ensure human rights, we need to make sure that they are in writing so that they are harder to violate*

*There should always be an ethics code, otherwise morality will be a diminished quality in society*

Others took the remark to concern whether any given institution (such as a workplace) would benefit from having a formal code of ethics, or whether a code of conduct would necessarily reflect moral considerations:

*Code of ethics are important and every institution needs guidelines. There are many ethical behaviors, that for some, need to be spelled out. In addition, all foundations of ethical behavior are important and should be defined by every organization that deals with the public.*

*Ethical principles indeed seem relevant as potentially something to include a code of conduct. A code of conduct by definition seems Ike it would involve ethics as the basis for the code.*

These responses have nothing to do with metaethics, and any participant who interpreted the item in this way's response does not reflect their views about metaethics.

**S4.7.12 Study 6B: Thematic analysis**

The most common themes can be seen in **Figure S4.18** and **Table S4.12**.

**Figure S4.18**

*Most common themes for Study 6B*



Study 6B: YB3 | Meaning

**Table S4.12**

*Most common themes for Study 6B*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 26.1% | 37 |
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 4.9% | 7 |
| *explicit subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 4.2% | 6 |
| *clear standards\** | The claim that moral standards are made clear (by e.g., a formal codification of ethics) | 3.5% | 5 |
| *definition\** | Conflates metaethical considerations with issues related to definitions | 3.5% | 5 |
| *don't know\** | Explicitly states that they don't know or understand | 3.5% | 5 |
| *personal\** | Describes morality as a matter of personal belief | 3.5% | 5 |

The only theme that really stands out is the descriptive conflation, which appeared in more than a quarter of responses. This illustrates the common tendency to interpret items on the YB3 in descriptive terms. This is somewhat surprising, since these items were selected in part because they *don't* straightforwardly consist of descriptive claims. I'm not sure what to make of this, other than that it may be that the tendency to lean on descriptive rather than normative claims is so strong that it emerges even in cases where it doesn't reflect a reasonable interpretation.

### S4.7.13 Study 6C: Thematic analysis

The most common themes can be seen in **Figure S4.19** and **Table S4.13**.
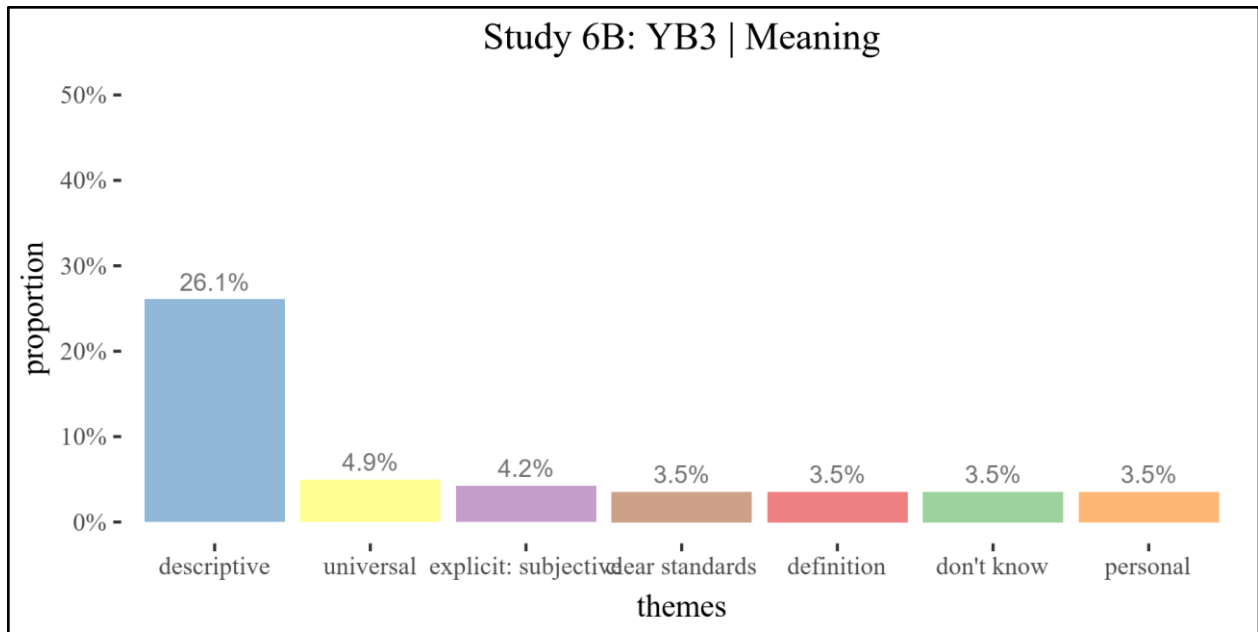
**Figure S4.19**

*Most common themes for Study 6C*
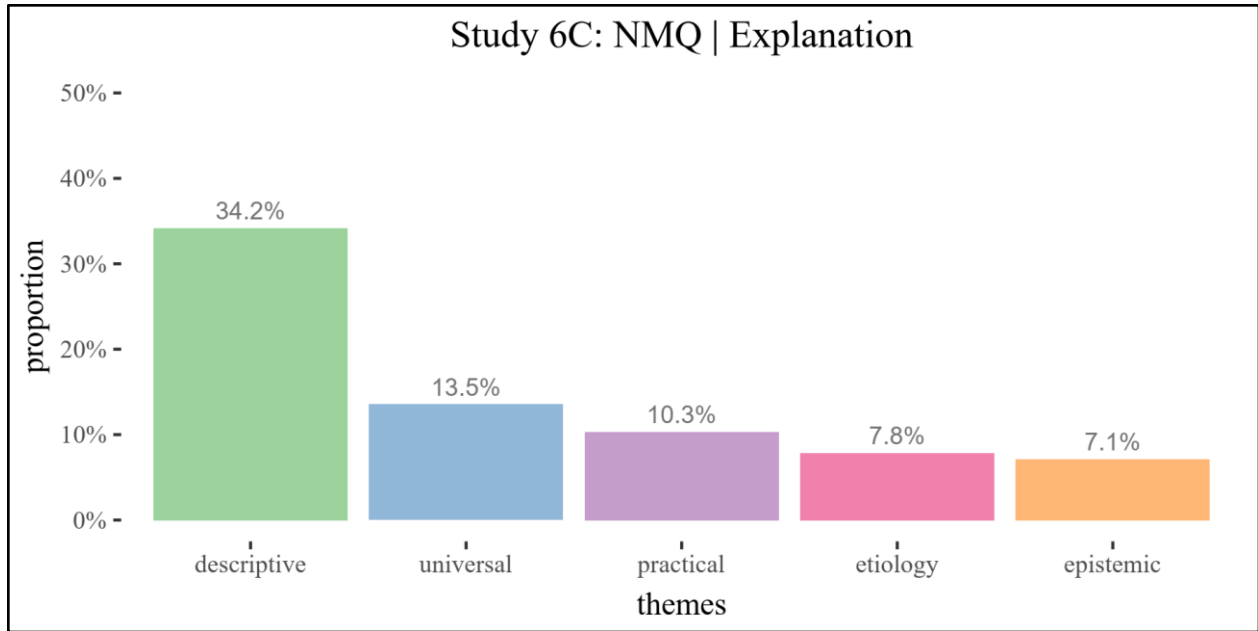


**Table S4.13**

*Most common themes for Study 6C*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The view that different people or groups have different moral standards | 34.2% | 96 |
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 13.5% | 38 |
| *practical* | Normative claim that things go better when people hold certain moral standards (e.g., the same standards as one another) | 10.3% | 29 |
| *etiology* | Offered a causal account of how the person would respond as they did | 7.8% | 22 |
| *epistemic* | General appeals to epistemic considerations | 7.1% | 20 |

Typical of so many other open response questions, the *descriptive* theme was the most common theme, at 21.1% (*n* = 96) of responses overall. This pattern held even for items on the realism subscale, not just the relativism subscale. That is, even when the item was not an ambiguous remark that could easily be interpreted as a descriptive claim that different people have different moral beliefs, participants *still* tended to interpret a variety of remarks in descriptive terms. For instance, consider this item: *Fundamental moral principles are universally valid; therefore they can be transferred from one society to another without difficulty.* Now consider some responses to this item:

> *Different societies have different beliefs and values , can't be just transferred.*

> *I think different cultures have different ideas about what is right and wrong. This is why we see lots of religious and political differences across the world.*

> *Some cultures view gender rights on a moral basis. Some are more restrictive and others more free. It's far from universal what people view as right and wrong for genders.*

Note that these responses have nothing to do with moral realism, for or against. They simply consist of claims about the existence of differences in moral belief, with one response also adding in that it isn't easy to transfer moral standards from one society to another *because* of those differences (which also has nothing to do with realism or antirealism), which is a sensible addition, since the item itself focuses on the ease with which we can "transfer" moral standards from society to another. Of course, this isn't surprising, since the item is not a face valid measure of realism to begin with. How easy it is to transfer the normative standards of one society to another is completely orthogonal to questions about realism and antirealism.[206] The remaining themes are not especially interesting. Several items (#5 and #6) discuss morality being "universal," a term that was echoed or referenced mostly in

---

[206] In addition the item is double-barreled, universalism does not entail realism, and it's unclear what a "fundamental" moral principle is or what it would mean for it to be "valid," both of which sound like technical terms, but have no obvious and unambiguous interpretation even to someone with training in metaethics and philosophe like myself. I have no idea what nonphilosophers would make of these terms. In short, this item is vague, ambiguous, and unrepresentative of realism in so many ways I genuinely have no idea what it would mean to agree or disagree with it, or why that would have anything to do with realism.

response to these items. Finally, the *practical* theme made a rare appearance in the top five. The *practical* theme refers to any instance in which the participant discusses the practical benefits of people having (or not having) some set of normative moral standards, or holding certain kinds of beliefs or attitudes. With few exceptions, almost all instances of this theme occurred in response to item #4: *What makes it possible for people to live together in harmony is the fact that fundamental moral rules do not differ from person to person.* Examples of responses with the *practical* theme include:

> *In order for people to get along and not be offended, they must share same beliefs if living together and/or have a serious and meaningful relationship*

> *People would get along better if they had similar morals and were not being immoral.*

Such responses concern the practical impact of various possible states of affairs, which has nothing to do with moral realism and antirealism. Yet far from illustrating that ordinary people struggle to understand metaethical concepts, these responses reveal inadequacies in the items themselves. I have gone out of my way to emphasize that these are *unintended interpretations*, not *misinterpretations*. I have done so for a reason. These are not unreasonable or confused interpretations of the item. The problem, in this case, is not how participants interpreted the item, but the item itself. The item just doesn't have anything to do with moral realism or antirealism. Achieving clear intended interpretations is a two-way street: not only must ordinary people interpret questions about realism and antirealism as intended, the questions must accurately reflect the relevant metaethical distinction in the first place.

**S4.7.14 Study 6D: Thematic analysis**

The most common themes can be seen in **Figure S4.20** and **Table S4.14**.

**Figure S4.20**

*Most common themes for Study 6D*
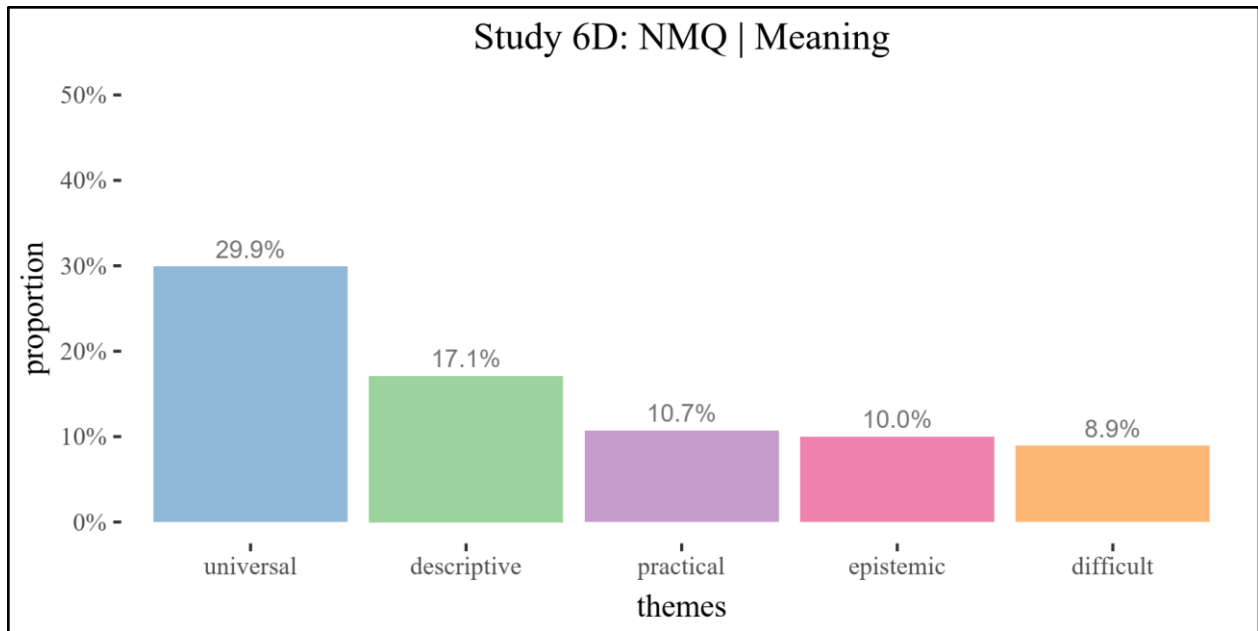


**Table S4.14**

*Most common themes for Study 6D*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *universal* | Normative position which holds that a given moral norm applies to everyone *or* explicit reference to morality being "universal" | 29.9% | 84 |
| *descriptive* | The view that different people or groups have different moral standards | 17.1% | 48 |
| *practical* | Normative claim that things go better when people hold certain moral standards (e.g., the same standards as one another) | 10.1% | 30 |
| *epistemic* | General appeals to epistemic considerations | 10.0% | 28 |
| *difficult* | View that it is easy or difficult to achieve some moral goal (e.g., persuading others or reaching consensus) | 8.9% | 25 |

Consistent with Study 6C, universal, descriptive, and practical were the most common themes, though *universal* and *descriptive* swapped places, with universal accounting for 19.4% ($n = 84$) of responses, and *descriptive* account for 11.1% ($n = 48$). Once again, participants appear to interpret questions ostensibly intended to reflect forms of realism and antirealism in ways that are not related to either, e.g., as descriptive claims, or as claims that moral standards apply to everyone. The *universal* theme makes an especially unfortunate appearance. When someone says that a norm applies to everyone, it is unclear what this means. This *could* entail or be consistent with moral realism, but it might not.

There is just no way to be sure. This is because the scope of a moral concern does not directly entail that the norm in question is stance-independent. For instance, if one believes all moral facts depend on God's will, one could believe the same standards apply to everyone, even if they are stance-dependent; the same holds true for ideal observer theory and for relation-designating accounts more generally, and may hold for any constructivist approach which grounds the legitimacy of moral facts in some procedure for generating such rules.

People could also be expressing some universal account on purely descriptive grounds, i.e., that it is *in fact true* that people abide by, or endorse, the same moral standards everywhere. For instance, consider this response:

> *The idea of what can be considered moral is different in different societies and in different situations. Therefore, there's no such thing a true, universal morality that applies to everyone in exactly the same way.*

> *It implies that everyone lives by the same 'code'*

The most plausible interpretation of these responses is *descriptive universalism*, the opposite of descriptive relativism. These are not metaethical positions.

Or they might endorse a type of *prescriptive universalism*: i.e., that it's *beneficial* or that we *should* conform to the same moral standards everywhere. One could endorse prescriptive universalism on pragmatic or normative moral grounds without endorsing moral realism. That is, someone could simply think

it's practically beneficial for us to adopt the same moral standards, which isn't even necessarily a moral position at all. For example, consider these responses:

*That people should think and believe the same things*

*Everyone should follow the same moral laws.*

These aren't metaethical positions.

They could also hold the first-order normative stance that we *morally ought* to share the same moral standards. Both such views are consistent with antirealism. I'm a moral antirealist, after all, and I think everyone should abstain from murder and torture and slavery, yet this simply reflects my personal stance. It has nothing to do with thinking that they ought to do so *independent of whether this would be consistent with their goals, standards, or values*, which is what you would need to show is the specific position people have in mind when interpreting questions about morality being "universal" if you're to conclude such items reflect some position on moral realism, rather than a mere normative stance on the scope of a given moral norm. Consider these responses:

*There are certain aspects of morality that are universal such as murder and stealing. No matter what culure, religion a person has, these things are universally negative and a no no.*

*Moral laws in my opinion would refer to life laws. Things that are moral, I believe, would apply to anyone. I cannot imagine that there would be cultures that are doing things that immoral. But then, maybe I'm not thinking outside the box on this.*

Do these responses clearly indicate that these universal moral rules are true independent of the goals, standards, or values of individuals or groups? No. The participants who offered these responses *might* think this, but saying that certain things are "universally negative" is consistent with believing that they are negative in virtue of shared intersubjective goals and values. I'm a moral antirealist, and I think torture and murder are a "universal negative," in the trivial, descriptive sense that these are generally considered bad (i.e., "negative") everywhere, and generally regarded as a "no-no." That is, I hold various stances about the descriptive facts (e.g., most people think torture is immoral), and stances

about what moral standards we should agree on (e.g., rules against torture). In short, the mere fact that someone thinks a moral rule "applies to everyone" or that it's "universal" does not provide enough information to conclude that they think such moral norms are stance-independently true.

Finally, we simply do not know whether (a) how people interpret the term "universal," or associated terms, e.g., "applies," (b) whether they interpret these terms in the same way as one another (c) whether or not, and to what extent, such interpretations are consistent with the interpretations intended by researchers. Without substantial descriptive evidence that their interpretations are a reliable indicator of their views towards realism and antirealism, interpreting such remarks as measures of their metaethical stances or commitments is highly questionable at best.

### S4.7.15 Study 7A: Thematic analysis

The most common themes can be seen in **Figure S4.21** and **Table S4.15**.

**Figure S4.21**
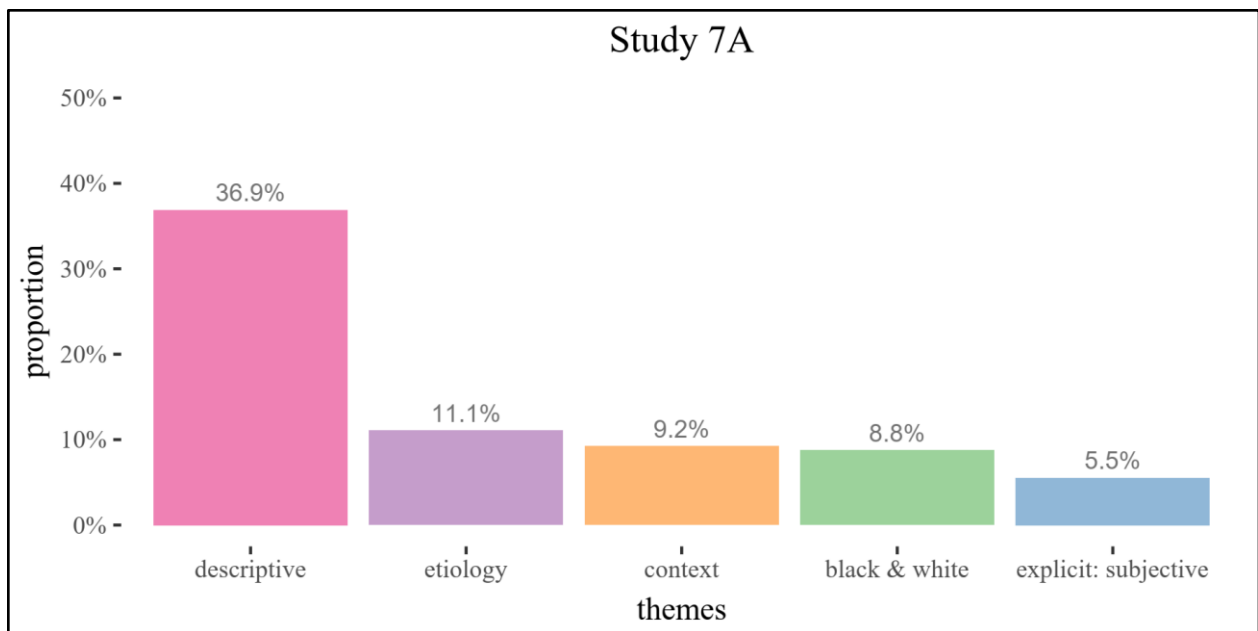
*Most common themes for Study 7A*

**Table S4.15**

*Most common themes for Study 7A*

| Theme | Explanation | Percentage | Frequency |
|---|---|---|---|
| *descriptive* | The claim that different people or groups have different moral standards | 36.9% | 80 |
| *etiology* | Provides a causal account of how the person would respond as they did | 11.1% | 24 |
| *context* | The claim that whether an action is right or wrong depends on context/circumstances | 9.2% | 20 |
| *black and white* | Describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality | 8.8% | 19 |
| *explicit subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") | 5.5% | 12 |

Once again, the descriptive theme was by far the most common one to emerge from the data, accounting for nearly 40% of responses. In this case, this is unsurprising, since three of the four items appear to conflate metaethics and descriptive claims. Only one of the items, which consists of an itemized version of the disagreement paradigm, is not subject to a formal descriptive conflation. Nevertheless, many participants still interpreted even this item as descriptive, though less frequently than for other items. Overall, these results are more consistent with the items having poor face validity than with ordinary people lacking determinate metaethical stances or commitments. Nevertheless, interpretation rates are not substantially better for most other items, even when those items have better face validity. It almost seems as though participants aren't going to interpret simple one-sentence items in metaethical terms regardless of what you include in an item. Perhaps richer and more sophisticated language would succeed, though perhaps at the cost of confusing participants. And

perhaps richer and more detailed items would succeed where simple one-sentence items fail, though again there'd be disadvantages to taking such an approach.

Etiology was the second most common theme. Etiology is closely related to the descriptive theme. It captures claims about the *origins* of our moral beliefs. Many participants made remarks along these lines:

> *Maybe people have different personal experiences or were raised differently and might sway them one way or the other, especially if is grey area type moral issue.*

> *Different people are grown up differently in their beliefs and therefore their morals can be different too.*

> *Based on someone's upbringing is going to determine their moral compass. A child raised in a traditional Christian home is going to have different morals than a child raised in a Palestinian home where they are taught to hate Jews and think that killing a Jew is a good thing.*

The descriptive theme captures claims that people have different moral standards, but such remarks don't need to offer explanations for the causal origins of differences (or similarities) in moral belief. This reveals an interesting feature of open response questions, and a possible methodological shortcoming in my approach. When I ask people to explain what items on metaethics scales mean, or why people agree or disagree with them, people may interpret these not as requests to explain what the item means, but to provide some rationale or motivation for the item, or to explain why one would be inclined to make such a claim.

For instance, if I ask someone to explain why they agree that people from different societies have different beliefs, a perfectly reasonable response is "because they were raised to have different beliefs." This is a sensible interpretation of the question, even though it does not express any metaethical content. This is one reason why future studies should employ a wider and more sophisticated array of questions. This could involve more sophisticated qualitative procedures, e.g., interviews, or pivot towards multiple choice or other quantitative measures designed to assess interpretation, e.g., asking people whether they agree or disagree that an item is asking about whether

there are facts about what is right or wrong that don't depend on our moral standards. I take the latter approach in **Chapter 5**, and participants do not perform any better, but future studies could reveal that these results were due to methodological inadequacies.

### S4.7.16 Study 7B: Thematic analysis

The most common themes can be seen in **Figure S4.22** and **Table S4.16**.

**Figure S4.22**

*Most common themes for Study 7B*
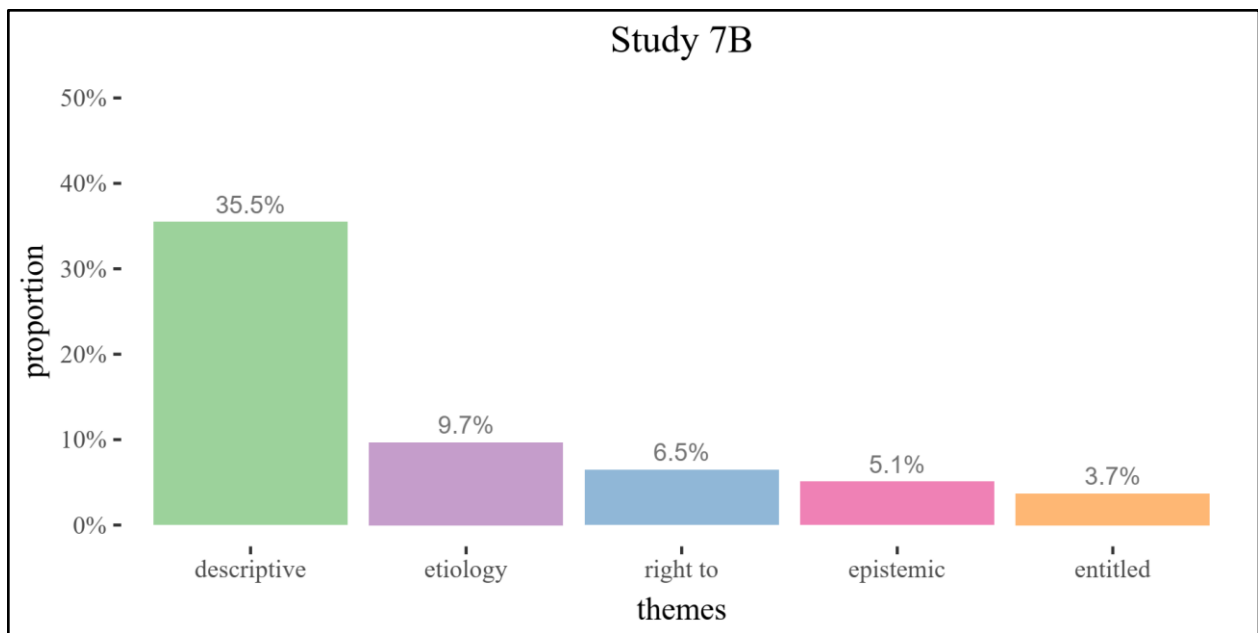


**Table S4.16**

*Most common themes for Study 7B*

| Theme | Explanation | Percentage | Frequency |
|-------|-------------|------------|-----------|
| *descriptive* | The claim that different people or groups have different moral standards | 35.5% | 77 |
| *etiology* | Provides a causal account of how the person would respond as they did | 9.7% | 21 |
| *right to* | The claim that each person considers their standards to be correct | 6.5% | 14 |

| *epistemic* | General comments or appeals to epistemic considerations | 5.0% | 11 |
|---|---|---|---|
| *nihilism* | The claim that there is no moral truth | 3.8% | 8 |

Once again, *descriptive* was the most common theme by far, comprising 35.5% (*n* = 77) of responses. It should be fairly clear, at this point, that many people take items intended to reflect metaethical relativism to instead reflect the descriptive claim that different people and societies have different moral standards. Only 12.4% (*n* = 27) of participants provided clear intended interpretations, while nearly three times that many were coded as *descriptive*. The *explain* condition provides far stronger evidence that people's responses are clearly unintended. This is because, when someone is asked to explain why they agree or disagree with a statement, their response could simply fail to explicitly reference the meaning of the item itself. If you ask me why I went to the store, my response won't necessarily reveal that I understand what a "store" is. If, however, you explicitly ask me to explain what a "store" is, and I fail to do so, this provides at least some indication that I don't know what a store is. The descriptive theme was common for all four items, though it was especially common for items #2 and #3. Here are illustrative examples of responses coded as *descriptive* for each of the four items:

> *Some people believe in different meanings or scripturesor they may take it in a different way.*

> *It means people hold different views on what's morally right or wrong. Not everyone shares the same opinion on topics and issues. They are spread across a wide spectrum.*

> *Different cultures have different norms and beliefs*

> *People, cultures, and societies can have different ideas of what is morally right or wrong. For instance one person may be against capital punishment because they think it is morally wrong, whereas another person may not see it as morally wrong. The same with cultures and societies.*

None of these responses have anything to do with metaethics. These participants did not interpret items on the FMO in the way Zijlstra intended.

The *etiology* (9.7%, *n* = 21) and *right to* (6.5%, *n* = 14) themes were the second and third most common themes, and closely conceptually linked to the *descriptive* theme. The *etiology* theme refers to instances where the participant explains how people came to hold their moral beliefs, e.g.:

> *A society in which a person lives and the society where he is stays are both very important to determine what is morally right or wrong*

> *It indicates that if a person were brought up in a different place that there beliefs may differ accordingly.*

These responses do allude to people having different moral beliefs and standards, but they go beyond this, by pointing to the causal origins of those beliefs and standards. Many participants appear to interpret questions about relativism not as questions about whether different people have different moral beliefs, but to concern *how* people come to have different moral beliefs. Typically, people will attribute the causal origins of our moral beliefs to our culture and our personal experiences. *Right to* is another recurring theme conceptually related to descriptive variation. Participants will once again go beyond simply pointing to the descriptive fact that different people have different moral beliefs by adding an additional detail: that each person *thinks they are correct*, or that certain moral beliefs are "true to" or "right to" the people who hold those beliefs:

> *it means that everyone can be right in their own view*

> *Becasue individuals have different morals, what is morally right to one person may be morally worng to another.*

> *Each society or culture decides what is immoral or moral to them.*

Note the pattern here: each of these participants thinks of moral claims in terms of them being right *according to their views*. This is not relativism, since relativism would hold that those people *are in fact right* in some relative sense, not that they *consider themselves* right. Consider a common way people describe other people's views:

> *According to young earth creationists, the earth is only 6,000 years old*

> *According to her, it's* not *true that the moon landing really took place.*

We can speak about what's true *to* people without supposing that we, ourselves, believe it to be true. This is what responses coded as *right to* appear to express. This is a plausible, interesting, and natural way to respond to items ostensibly intended to reflect relativism, once again highlighting that participants aren't simply incompetent or foolish in some straightforward way. Rather, the very fact that they respond in the precise ways highlights the rich, context sensitive, and highly flexible way people interpret statements. Absent sufficient context, there are a variety of ways people could interpret items intended to reflect metaethical claims. If, as I suspect, they don't have the relevant metaethical concepts, they are left with little choice but to suppose that these items mean *something*. Landing on the notion that a moral standard can be "correct" according to one person but "incorrect" according to another is at least in the conceptual vicinity of relativism. Unfortunately, it isn't relativism, and conceptual relatedness isn't sufficient to warrant judging such responses to be genuine reflections of relativism.

**S4.8 Full theme list**

**Table S4.17**

*Theme descriptions for tables*

| Theme | Explanation |
|---|---|
| *absolute* | Varies in meaning, but is associated with explicit use of term "absolute," exceptionless moral rules, black and white thinking, certainty, or being close-minded |
| *all good* [charity] | The claim that all charities do some kind of good |
| *black and white* | Describes morality as "black and white" or mentions "grey areas." Typically conveys a rigid, absolutist, or inflexible approach towards morality |
| *both sides* | The claim that both sides of a disagreement have part of the truth |
| *cause* [charity] | The claim that which charity is best depends on the cause of the charity |
| *clear standards* | The claim that moral standards are made clear (by e.g., a formal codification of ethics) |

| | |
|---|---|
| *closed* | The claim that whoever endorses the moral position in question is close-minded or rigid in their thinking |
| *consensus* | The claim that something is objective because most or all people agree |
| *context* [*disagreement*] | Disagreement attributed to different assumptions about the circumstances in which the action was performed |
| *context* | The claim that whether an action is right or wrong depends on context/circumstances |
| *culture* | Descriptive or etiological claim that attributes moral stance to a person's culture |
| *definition* [disagreement] | Attributes disagreement to different definitions of moral terms/concepts |
| *definition* | Conflates metaethical considerations with issues related to definitions (such as the definition of "moral") |
| *descriptive* | The claim that different people or groups have different moral standards |
| *difficult* | The claim that it is easy or difficult to achieve some moral goal (e.g., persuading others or reaching consensus) |
| *efficient* [charity] | The claim that some charities are more efficient than others at helping people or managing resources |
| *empathy* [disagreement] | Attributes empathy or compassion towards someone who disagreed |
| *entitled* | The claim that people are entitled to their beliefs |
| *epistemic* | General comments or appeals to epistemic considerations |
| *etiology* | Provides a causal account of how the person would respond as they did |
| *explanation* | Provides an explanation for why the person might believe what they do |
| *explicit: objective* | Explicit use of term "objective" (or related term, e.g., "objectivism") |
| *explicit: relative* | Explicit use of term "relative" (or related term, e.g., "relativism") |
| *explicit: subjective* | Explicit use of term "subjective" (or related term, e.g., "subjectivism") |
| *judgment* | The normative claim that people should act in accordance with the speaker, or that the speaker's views are the "only" way to act |
| *measure* | The claim that something is "objective" if it can be measured or quantified |
| *nihilism* | The claim that there is no moral truth |
| *normative* [disagreement] | Expresses a moral judgment about the person who disagreed with them |
| *normative* | Miscellaneous remarks about normative and evaluative conceptions of morality |

| | |
|---|---|
| *not considered* [disagreement] | Attributes disagreement to the other person not thinking about the scenario in an adequate way |
| *opinion* | The claim that something is a matter of opinion, typically explicitly using the term "opinion" with little or no qualification |
| *opposite* | Describes the opposite/contrary metaethical view (e.g., the participant appears to interpret "objective" to mean "subjective") |
| *overhead* [charity] | Comments on the overhead costs of charities |
| *personal* | Describes morality as a matter of personal belief |
| *practical* | The normative claim that things go better when people hold certain moral standards (e.g., the same standards as one another) |
| *reject* | Expresses disagreement with stimuli |
| *relative ambiguity* | Interprets the claim that there are no moral/normative facts with the claim that moral/normative facts are relative/subjective, or any instance in which a response is ambiguous between relativism and noncognitivism/nihilism |
| *religion* | Refers to religion or religious beliefs |
| *right to* | The claim that each person considers their standards to be correct |
| *tolerant* | The normative claim that we should tolerate or respect other people or cultures (or their moral standards) or that we shouldn't judge others |
| *unbiased* | The claim that something is "objective" when it is unbiased / impartial |
| *universal* | The claim that a given moral norm applies to everyone *or* explicit reference to morality being "universal" |

*Quasi-themes*

| **Quasi-theme** | **Explanation** |
|---|---|
| *complaint* | Complains about some feature of the study (e.g., inadequate compensation for open response questions) |
| *correct* | A clear intended response |
| *no answer* | The participant did not answer or gave a response that could not be construed as an answer, such an emoji or a "." |
| *other* | The response was unusual and could not be classified using standard themes. These items should in principle receive unique codes appropriate to the response |

| | |
|---|---|
| *repeat* | The participant either (a) repeated significant proportions of the stimuli or (b) stated some trivial metacommentary (e.g., "this is the same as my previous response") |
| *unclear* | An unclear response |

**SUPPLEMENT TO CHAPTER 5**

**S5.1 Additional commentary on limitations**

*S5.1.1 Study 1D: Additional discussion and limitations*

One issue to address is the nonsignificance of the realism item in study 1D, once participants who failed the comprehension check were removed. Notably, this was an item I created. My comparatively greater, if inconsistent success (my other item performed worse than average compared to items in other studies), suggests that there may be *some* signal in all the noise, and points to the possibility that many other studies may rely on invalid measures not because people have no determinate views, but because other studies lack face validity.

This is consistent with an examination of the content of these items, which appear to conflate realism, antirealism, and relativism with other, unintended concepts. It is also consistent with the possibility that measures used in other studies are invalid not because people have no determinate views, but because these studies lack face validity. This is supported by examination of the content of these items, which appear to conflate realism, antirealism, and relativism with other, unintended concepts. Yet performance was far below what would be necessary for a valid measure even for my items with greater face validity, despite the nonsignificance of one of the results. Even though I did not obtain unequivocal evidence that correct responses are uniformly below 50%, this outcome was avoided by a hair's breadth. It's worth pausing to reflect on how far this is from a vindication of folk metaethics research, and how little this does to warrant significant hope that valid scale items could be constructed.

First, this could simply be Type II error (i.e., a false negative). Type II errors are to be expected when running many analyses. This is due in part to the fact that this study was not adequately powered to detect small differences. Regardless of the reason why this result became nonsignificant, 41.8% with the upper bound of the 95% confidence interval being a mere 50.4% is consistent with just *barely*

more than half of participants crossing the finish line once we exclude those who fail comprehension checks. At best, this would mean that we've identified one item that could scrape past the threshold of half of participants selecting the correct response. Yet it is worth emphasizing that this is already an incredibly low threshold that, if passed, is still not good enough to declare an item to be valid. My goal in using such a low threshold is to illustrate that items cannot meet even this very low bar. I did not select this low of a bar because it reflected a genuine finish line that, if crossed, would indicate that a measure was valid. It was selected because it was so far below what would be necessary for a valid measure that the fact that most measures couldn't reach even this low of a bar served as a rhetorical point to illustrate just how atrocious existing measures actually are. I'm not surprised one of my own items may have just scraped by. I could have just as reasonably set the bar at 55% or 60%, which still wouldn't be good enough for a valid measure, and this item would have decisively failed such a test.

In any case, even if it's possible one item just barely manages to do so, this should be cold comfort to efforts to devise valid measures of realism. It is worth noting, in addition, that my potential success at creating an item that outperforms all existing items goes some way in vindicating my objections to those items and pointing to my own competence at constructing valid items. In other words, if the best we can do is stumble onto items that slightly outperform other items, but we're still left with nearly half of participants not understanding what they're being asked, this is hardly evidence that people can reliably interpret questions about metaethics as intended.

### S5.1.2 Study 2: Additional discussion and limitations

One concern with Study 2 is that it's also possible that I constructed stimuli in a way that inappropriately biased participants towards one or another interpretation of quantum mechanics, but that a more neutral description would have resulted in a more equal preference for each. My description of the Many Worlds interpretation deliberately emphasized its strange implications, which

may have made it a less appealing choice. However, the implications are accurate, and in any even if hamming up my description of the Many Worlds interpretation did drive participants towards the Copenhagen interpretation, this could simply serve to illustrate how, in ordinary experiments, researcher bias could influence the content of stimuli in ways that result in nonrandom response patterns, even if in the absence of such biases results would be random. Far from illustrating that indeterminacy should lead us to predict an equal distribution, this simply reveals one of the mechanisms that would lead us to not make such predictions: e.g., non-neutral descriptions of stimuli could influence which responses people favor.

**SUPPLEMENT TO CHAPTER 6**

**S6.1 Study 1**

**S6.1.1 Study 1: Exploratory factor analysis for metaethics scale items**

I conducted exploratory factor analysis on the *other* (third person) and *self* (first person) versions of the metaethics scales, which may be seen in **Table S6.1.1** and **Table S6.1.2**, respectively. Both EFAs were conducted using Jamovi 2.3.21 (The jamovi project, 2022). Both EFAs were conducted using principal axis extraction and oblimin rotation, and the number of factors displayed was based on parallel analysis. A Kaiser-Meyer-Olkin (KMO) was conducted for the other and the self conditions. For the *other* condition overall KMO = 0.861, indicating the sampling adequacy of the scale. For the *self* condition overall KMO = 0.911, which likewise indicates the overall sampling adequacy of the scale. Bartlett's test of sphericity was also conducted for the *other* and *self* scales. For the *other* scale, $\chi 2(66) =$ 1805, $p < .001$, indicating adequacy for factor analysis. For the *self* scale results were also sufficient for factor analysis, $\chi 2(66) = 2920$, $p < .001$.

Results from both EFAs indicate that the realism and universalism items are strongly loaded onto the same factor. While a bottom-up approach to understanding a putative variable that such items would be intended to measure may suggest that they should be collapsed into a single factor, such an impulse should be resisted. Realism and universalism are conceptually distinct. If people treat them as more or less the same, this *could* be interpreted as some fascinating feature of human psychology. However, I find it more plausible that the failure to distinguish the two is best attributable to participants not interpreting the items (either together or collectively) as intended. That is, the fact that they load onto the same factor indicates problems with the measures. For comparison, if results from a study that asked people to identify cars and trees loaded onto the same factor, I would not conclude that people had some kind of chimeric car-tree concept driving their judgments. I'd first wonder whether they interpreted the questions as I intended.

**Table S6.1.1**

*Factor loadings for other condition*

| | Factor | | Uniqueness |
| | 1 | 2 | |
|---|---|---|---|
| oth_obj1 | 0.625 | | 0.499 |
| oth_obj2 | 0.671 | | 0.536 |
| oth_obj3 | 0.430 | | 0.845 |
| oth_uni1 | 0.798 | | 0.305 |
| oth_uni2 | 0.798 | | 0.382 |
| oth_uni3 | 0.783 | | 0.406 |
| oth_rel1 | | 0.604 | 0.480 |
| oth_rel2 | | 0.900 | 0.217 |
| oth_rel3 | | 0.729 | 0.452 |
| oth_cog1_R | | -0.394 | 0.866 |
| oth_cog2 | | | 0.939 |
| oth_cog3 | 0.678 | | 0.577 |

*Note.* Extraction method: Principal axis; Rotation: oblimin. Factors below 0.3 are not displayed.

**Table S6.1.2**

*Factor loadings for self condition*

| | Factor | | Uniqueness |
| | 1 | 2 | |
|---|---|---|---|
| self_obj1 | 0.908 | | 0.207 |
| self_obj2 | 0.645 | 0.304 | 0.326 |
| self_obj3 | 0.559 | 0.320 | 0.574 |
| self_uni1 | 0.944 | | 0.150 |
| self_uni2 | 0.711 | | 0.329 |
| self_uni3 | 0.903 | | 0.197 |
| self_rel1 | -0.470 | 0.412 | 0.311 |
| self_rel2 | | 0.872 | 0.242 |
| self_rel3 | | 0.773 | 0.404 |
| self_cog1_R | | -0.453 | 0.735 |
| self_cog2 | | 0.303 | 0.891 |
| self_cog3 | 0.575 | | 0.481 |

*Note.* Extraction method: Principal axis; Rotation: oblimin. Factors below 0.3 are not displayed.

The relativism subscale performed reasonably well, loading onto the same factor in both conditions (though the first item did not perform exceptionally well on either, and was especially poor in the *self* condition). This at least gestures at the potential for distinguishing realism and universalism from relativism, and future efforts could perhaps improve on these results with a larger initial pool of items and a larger sample. Finally, the noncognitivism subscale performed terribly, with the items not loading onto the same factor for either version of the scale. I suspect this is because, while philosophers may understand these items be asking about a singular concept (whether moral claims are truth-apt), participants are probably not interpreting such items in this way, nor do they have any implicit competence in doing so in a way responsive to the wording used for these items. I am not optimistic about the prospects of devising a noncognitivism scale that would perform adequately, but it would at least be worth putting more effort into the task using a new and larger set of items.

This is a case where it would be especially helpful to assess how participants were interpreting these items, and to put effort more generally into qualitatively assessing what sorts of phrasing could be used to prompt the desired interpretation. This may be *extremely* difficult. Epistemic conflations, social desirability, and other factors may all play an especially large role when it comes to noncognitivist items. It would be difficult to disentangle, for instance, a desire to hedge on epistemic grounds in judging that a particular view is "neither true nor false," or "can't" be true or false, which could be understood to convey the participant's agnosticism or uncertainty, and a judgment about whether moral claims (as a class) express propositions.

Of course, if I am correct that ordinary people don't generally have determinate metaethical positions, this could be a case where they simply don't have a position on whether moral claims are truth-apt. Having a position on the matter would require thinking about morality in a highly structured way: one would have to recognize a conceptually distinct domain of norms, the *moral domain*, and one would have to have a principled stance on the metanormative characteristics of the domain as a whole.

In this case, it would revolve around whether utterances associated with the domain were, as a feature of their semantic characteristics, either consist analytically of propositional claims, or are primarily or exclusively used to express propositional claims. This would further require some distinction (however implicit) between propositions and nonpropositional utterances. It's certainly possible for people to have internalized sophisticated features of their language such that they can competently make judgments in accordance with an implicit utilization of all of these distinctions, but it could simply turn out at numerous junctures in this chain of characteristics needed for people to have some uniform stance or commitment towards moral issues regarding their truth-aptness that people just don't speak or think in a way that would yield a uniform and determinate stance on the matter. That people do speak and think in a way that could result in the judgment that moral claims are truth-apt requires a host of assumptions about language and psychology that could simply be false. Overall, such commentary may achieve little more than recapitulating my skepticism about the determinacy of folk metaethics. Such matters are best resolved by new and better empirical research.

### S6.1.2 Study 1: Additional graphs and tables

**Table S6.1.3**

*One-Way ANOVA (Welch's)*

|  | F | df1 | df2 | p |
|---|---|---|---|---|
| other \| realism | 11.535 | 2 | 258 | < .001 |
| other \| universalism | 17.303 | 2 | 255 | < .001 |
| other \| relativism | 9.221 | 2 | 258 | < .001 |
| self \| realism | 0.659 | 2 | 258 | 0.518 |
| self \| universalism | 1.273 | 2 | 258 | 0.282 |
| self \| relativism | 2.417 | 2 | 258 | 0.091 |

**Figure S6.1.1**

*Study 1: Third person|realism condition*
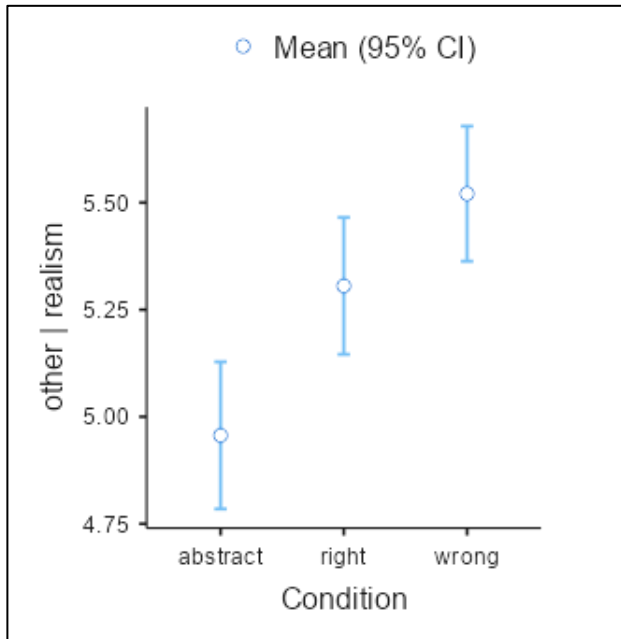


**Figure S6.1.2**

*Study 1: Third person|universalism condition*
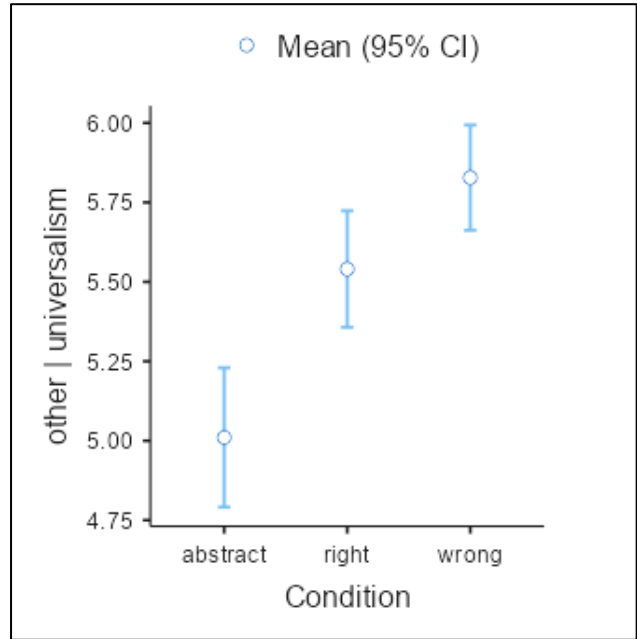


**Figure S6.1.3**

*Study 1: Third person|relativism condition*
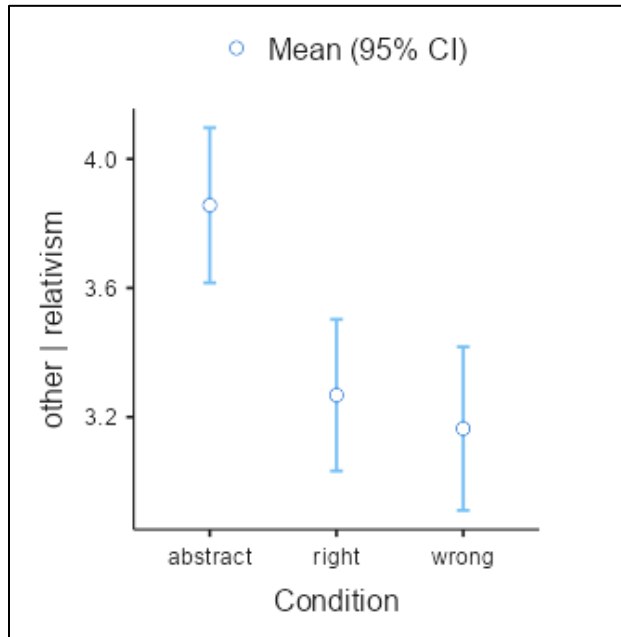


**Figure S6.1.4**

*Study 1: First person|realism condition*
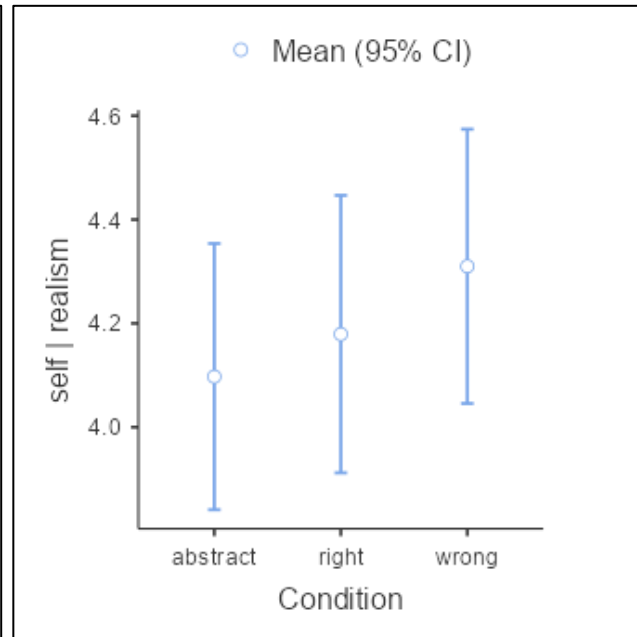
**Table S6.1.4**

*Study 1: Group descriptives for first person and third person realism, universalism, and relativism conditions*

|  | Condition | n | Mean | SD | SE |
|---|---|---|---|---|---|
| other \| realism | abstract | 130 | 4.96 | 0.990 | 0.0868 |
|  | right | 132 | 5.31 | 0.928 | 0.0808 |
|  | wrong | 128 | 5.52 | 0.902 | 0.0797 |
| other \| universalism | abstract | 130 | 5.01 | 1.264 | 0.1108 |
|  | right | 132 | 5.54 | 1.064 | 0.0926 |
|  | wrong | 128 | 5.83 | 0.946 | 0.0837 |
| other \| relativism | abstract | 130 | 3.86 | 1.385 | 0.1215 |
|  | right | 132 | 3.27 | 1.364 | 0.1188 |
|  | wrong | 128 | 3.16 | 1.449 | 0.1281 |
| self \| realism | abstract | 130 | 4.10 | 1.478 | 0.1296 |
|  | right | 132 | 4.18 | 1.552 | 0.1351 |
|  | wrong | 128 | 4.31 | 1.511 | 0.1335 |
| self \| universalism | abstract | 130 | 3.81 | 1.750 | 0.1535 |
|  | right | 132 | 4.04 | 1.790 | 0.1558 |
|  | wrong | 128 | 4.16 | 1.866 | 0.1649 |
| self \| relativism | abstract | 130 | 4.69 | 1.466 | 0.1285 |
|  | right | 132 | 4.33 | 1.571 | 0.1367 |
|  | wrong | 128 | 4.35 | 1.554 | 0.1374 |

**Figure S6.1.5**

*Study 1: First person|universalism condition*
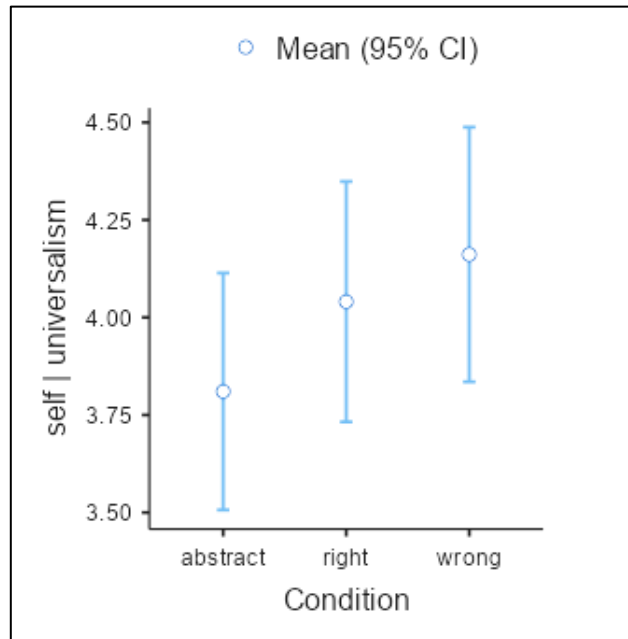
**Figure S6.1.6**

*Study 1: First person|relativism condition*
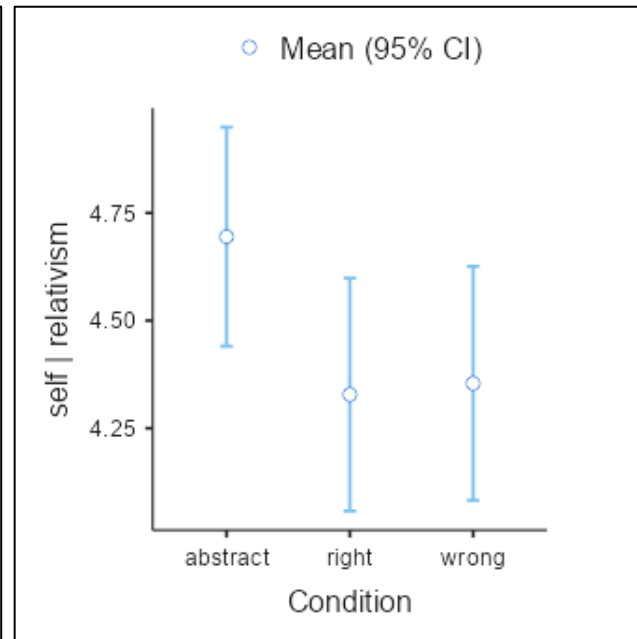




**S6.2 Study 2: Third person paradigm with concrete moral issues**

**Methods**

***Participants.*** I aimed to recruit 700 participants. 701 participants completed the study. Participants consisted of 701 adult US residents on Amazon's Mechanical Turk (329 females, 365 males, 6 other, 1 unreported, $M_{age}$ = 40.9, $SD_{age}$ = 12.3, age range = 20-77).

***Procedure.*** Unlike all other studies in this chapter, Study 2 employed a within-subjects design. All participants were assigned to all conditions. The order of conditions was partially randomized. First, all participants were assigned to the no statement condition employed in Study 1.[207] Then participants were presented with four conditions, presented in random order, in which a person is having a

---

[207] I had planned to also include the abstract right and wrong statements used in Study 1 as well, but I dropped these due to resource constraints and concerns with participant fatigue. In particular, a pretest of the study yielded a high drop rate, indicating that some participants may have found the repetitive nature of the task onerous and dropped out. I felt it best to proceed with testing the concrete moral issues that this study focuses on, with some minimal replication of the previous study's conditions as well.

discussion and then makes a moral statement. The moral statements were drawn from a pool of 45

pretested items that were designed to reflect a broad range of moral issues. The four items were:

*(1) It is morally wrong to eat meat.*

*(2) It is morally wrong for a professor to give a bad grade to a student just because they dislike the student.*

*(3) It is morally wrong for a person to go to a funeral to mock the deceased person in front of their family.*

*(4) It is morally wrong for a woman who knows she is pregnant to drink alcohol.*

Items were chosen based on their severity and plausibility as moral transgressions, in order to avoid

focusing exclusively on rare, unrealistic, or implausible moral transgressions that are commonly

employed in moral psychological research (such as sacrificial dilemmas). Each item also varied in terms

of pretested levels of participant agreement (the claim that it is morally wrong to eat meat had the

lowest agreement of the 45 items, while the other three had high agreement), perceived realism (i.e.,

whether participants endorsed a moral realist response with respect to the issue; only the alcohol

condition had high perceived realism), and perceived consensus (i.e., whether participants thought

most people agreed about the issue; every issue was moderately high in perceived consensus except

for eating meat, which was the lowest of the 45 items).

Participants were first asked to judge the metaethical beliefs and character of a typical person

in their society, and were then presented each of the four concrete moral conditions in random order,

and were again asked to judge the metaethical beliefs of the person who made each statement, along

with judgments about that person's character and quality as a social partner. Finally, participants were

asked about their own metaethical beliefs and to report their age and gender.

**Measures.** Study 2 employed measures drawn from the same measures used in Study 1. However, to

reduce participant fatigue with a within-subjects design, this study only employed one measure from

each of the realism, universalism, and relativism scales (the noncognitivism scale was dropped entirely

due to its poor performance in Study 1). In particular, I chose the following items from the realism, universalism, and relativism scales, respectively:

Realism
*They believe moral truth is independent of cultural standards and personal beliefs.*

Universalism
*They believe the same moral rules apply to everyone.*

Relativism
*They believe that things are only morally right or wrong according to different points of view.*

First person versions of these items were presented after all of the third person measures as well. These three items were used in all five third person conditions and in the first person condition. I also included a handful of character measures for exploratory purposes for the third person measures. Participants were asked to judge the other person's moral character (1 = Very morally bad, 7 = Very morally good), empathy (1 = Not empathic at all, 7 = Very empathic), how seriously they take morality (1 = Not seriously at all, 7 = Very seriously), and how good it would be to have such a person as a coworker, neighbor, or close friend (1 = Not at all, 7 = Very good).

**Results**

My primary interest was in evaluating differences in perception of the target's metaethical positions (realism, universalism, and relativism) and character traits (moral character, empathy, moral seriousness, and how good of a social partner they would be) across the five conditions: (1) no statement, (2) meat (3) grade, (3) mock, and (4) alcohol. To test for these differences, I conducted a repeated measures ANOVA for each of the seven primary measures.

I conducted a repeated measures ANOVA to compare the effect of statement condition on perceived objectivism. However, Mauchly's test revealed that the sphericity assumption was not met,

$\chi^2(9) = 195.22, p < .001.$[208] Since the sphericity assumption was violated and $\varepsilon = 0.864$, I report Huynh-Feldt corrected results (van den Berg, 2022).[209] There was a significant main effect of statement condition on perceived objectivism, $F(3.48, 2418.78) = 22.40$ , $p < .001$ , $\eta^2_p = 0.031$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.1**. Descriptive statistics are available on **Table S6.2.2**. Estimated marginal means appear in **Table S6.2.3** and **Figure S6.2.1.**

---

[208] Initial analysis was conducted using Jamovi. However, Jamovi only provided the *p*-value, so Mauchly's test was also conducted in JASP to obtain all test results. There was some inconsistency in what version of JASP was used, but the latest version was JASP 0.17, and results were checked again in this version.

[209] Van den Berg (2022) suggests reporting Huynh-Feldt corrected results when $\varepsilon > 0.75$.

**Table S6.2.1**

*Post Hoc Comparison – objectivism*

| Comparison | | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|---|
| objectivism | | objectivism | | | | | |
| no statement | - | meat | 0.1363 | 0.0925 | 696 | 1.474 | 0.282 |
| | - | grade | -0.5524 | 0.0789 | 696 | -7.005 | < .001 |
| | - | mock | -0.3931 | 0.0849 | 696 | -4.628 | < .001 |
| | - | alcohol | -0.3400 | 0.0826 | 696 | -4.119 | < .001 |
| meat | - | grade | -0.6887 | 0.0975 | 696 | -7.062 | < .001 |
| | - | mock | -0.5294 | 0.0997 | 696 | -5.312 | < .001 |
| | - | alcohol | -0.4763 | 0.0994 | 696 | -4.793 | < .001 |
| grade | - | mock | 0.1593 | 0.0696 | 696 | 2.286 | 0.068 |
| | - | alcohol | 0.2123 | 0.0714 | 696 | 2.975 | 0.012 |
| mock | - | alcohol | 0.0531 | 0.0733 | 696 | 0.725 | 0.469 |

**Table S6.2.2**

*Descriptives – objectivism*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 697 |
| Mean | 4.37 | 4.23 | 4.92 | 4.76 | 4.71 |
| 95% CI mean lower bound | 4.25 | 4.07 | 4.79 | 4.61 | 4.57 |
| 95% CI mean upper bound | 4.48 | 4.39 | 5.05 | 4.90 | 4.85 |
| Standard deviation | 1.52 | 2.09 | 1.78 | 1.96 | 1.88 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

**Table S6.2.3**

*Estimated Marginal Means – objectivism*

| Objectivism | Mean | SE | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| no statement | 4.37 | 0.0574 | 4.25 | 4.48 |
| meat | 4.23 | 0.0792 | 4.07 | 4.39 |
| grade | 4.92 | 0.0676 | 4.79 | 5.05 |
| mock | 4.76 | 0.0742 | 4.61 | 4.90 |
| alcohol | 4.71 | 0.0712 | 4.57 | 4.85 |

**Figure S6.2.1**

*Estimated marginal means for perceived objectivism*
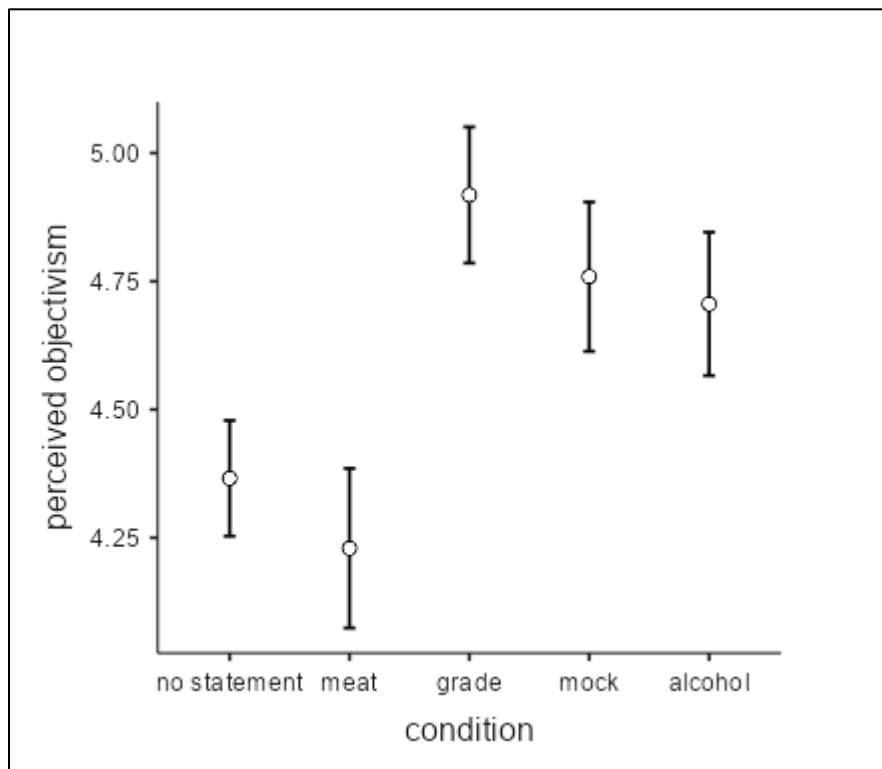


Next, I conducted a repeated measures ANOVA to compare the effect of statement condition on perceived universalism. Sphericity was violated, $\chi^2(9) = 177.53$, $p < .001$. Since $\varepsilon = 0.885$, I report

Huynh-Feldt corrected results. There was a significant main effect of statement condition on perceived universalism, $F(3.56, 2476.99) = 29.10$ , $p < .001$ , $\eta^2_p = 0.040$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.4**. Descriptive statistics are available on **Table S6.2.5**. Estimated marginal means appear in **Table S6.2.6** and **Figure S6.2.2.**

**Table S6.2.4**

*Post Hoc Comparison – universalism*

| Comparison | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|
| universalism | universalism | | | | | |
| no statement | - meat | -0.0129 | 0.0796 | 696 | -0.162 | 0.871 |
| | - grade | -0.4132 | 0.0709 | 696 | -5.830 | < .001 |
| | - mock | -0.4763 | 0.0738 | 696 | -6.452 | < .001 |
| | - alcohol | -0.5854 | 0.0689 | 696 | -8.492 | < .001 |
| meat | - grade | -0.4003 | 0.0815 | 696 | -4.909 | < .001 |
| | - mock | -0.4634 | 0.0810 | 696 | -5.724 | < .001 |
| | - alcohol | -0.5725 | 0.0768 | 696 | -7.457 | < .001 |
| grade | - mock | -0.0631 | 0.0576 | 696 | -1.095 | 0.547 |
| | - alcohol | -0.1722 | 0.0592 | 696 | -2.911 | 0.015 |

**Table S6.2.5**

*Descriptives – universalism*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 697 |
| Mean | 5.12 | 5.13 | 5.53 | 5.60 | 5.71 |
| 95% CI mean lower bound | 5.01 | 5.00 | 5.42 | 5.48 | 5.60 |
| 95% CI mean upper bound | 5.23 | 5.27 | 5.65 | 5.71 | 5.81 |
| Standard deviation | 1.52 | 1.77 | 1.53 | 1.58 | 1.38 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

**Table S6.2.6**

*Estimated Marginal Means – universalism*

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| Universalism | Mean | SE | Lower | Upper |
| no statement | 5.12 | 0.0574 | 5.01 | 5.01 |
| meat | 5.13 | 0.0672 | 5.00 | 5.00 |
| grade | 5.53 | 0.0580 | 5.42 | 5.42 |
| mock | 5.60 | 0.0598 | 5.48 | 5.48 |
| alcohol | 5.71 | 0.0523 | 5.60 | 5.60 |

**Figure S6.2.2**

*Estimated marginal means for perceived universalism*
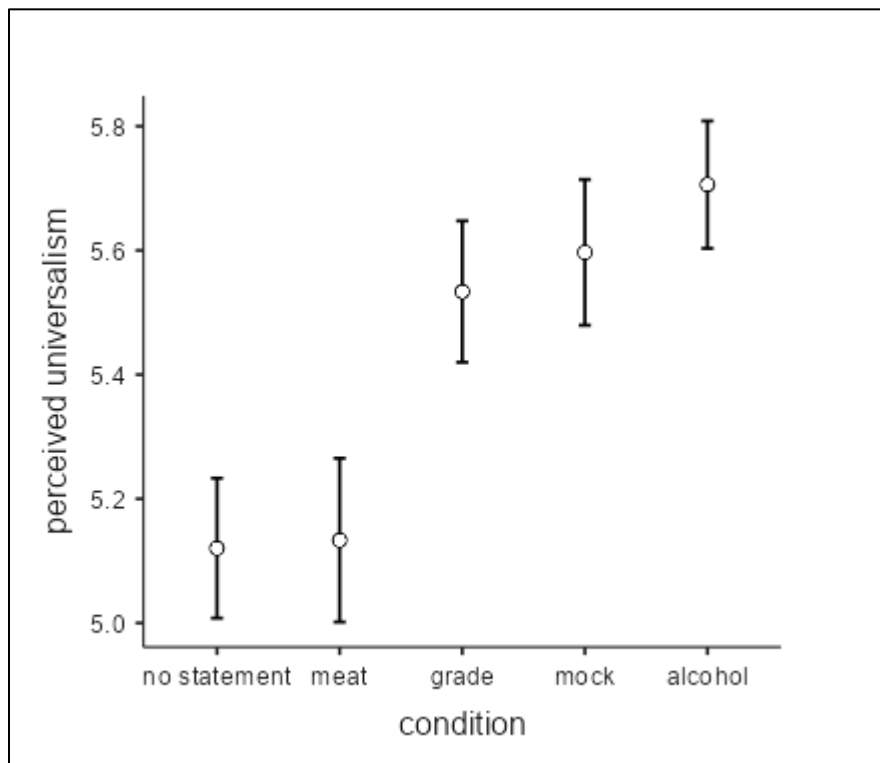


I also conducted a repeated measures ANOVA to compare the effect of statement condition on perceived relativism. Sphericity was violated, $\chi^2(9) = 132.50, p < .001$ . Since $\varepsilon = 0.904$, I report

Huynh-Feldt corrected results. There was a significant main effect of statement condition on perceived universalism, $F(3.64, 2532.74) = 19.4$ , $p < .001$ , $\eta^2_p = .027$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.7**. Descriptive statistics are available on **Table S6.2.8**. Estimated marginal means appear in **Table S6.2.9** and **Figure S6.2.3.**

**Table S6.2.7**

*Post Hoc Comparisons – relativism*

| Comparison | | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|---|
| relativism | | relativism | | | | | |
| no statement | - | meat | 0.48924 | 0.0783 | 696 | 6.250 | < .001 |
| | - | grade | 0.36298 | 0.0736 | 696 | 4.929 | < .001 |
| | - | mock | 0.49785 | 0.0736 | 696 | 6.764 | < .001 |
| | - | alcohol | 0.58967 | 0.0700 | 696 | 8.421 | < .001 |
| meat | - | grade | -0.12626 | 0.0837 | 696 | -1.508 | 0.528 |
| | - | mock | 0.00861 | 0.0853 | 696 | 0.101 | 0.920 |
| | - | alcohol | 0.10043 | 0.0805 | 696 | 1.247 | 0.528 |
| grade | - | mock | 0.13486 | 0.0630 | 696 | 2.141 | 0.163 |
| | - | alcohol | 0.22669 | 0.0638 | 696 | 3.553 | 0.002 |

**Table S6.2.8**

*Descriptives – relativism*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 697 |
| Mean | 3.79 | 3.30 | 3.42 | 3.29 | 3.20 |
| 95% CI mean lower bound | 3.67 | 3.16 | 3.29 | 3.15 | 3.06 |
| 95% CI mean upper bound | 3.90 | 3.44 | 3.55 | 3.43 | 3.33 |
| Standard deviation | 1.58 | 1.87 | 1.75 | 1.84 | 1.77 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

**Table S6.2.9**

*Estimated Marginal Means – relativism*

| Relativism | Mean | SE | 95% Confidence Interval | |
| --- | --- | --- | --- | --- |
| | | | Lower | Upper |
| no statement | 3.79 | 0.0599 | 3.67 | 3.90 |
| meat | 3.30 | 0.0710 | 3.16 | 3.44 |
| grade | 3.42 | 0.0661 | 3.29 | 3.55 |
| mock | 3.29 | 0.0699 | 3.15 | 3.43 |
| alcohol | 3.20 | 0.0670 | 3.06 | 3.33 |

**Figure S6.2.3**

*Estimated marginal means for perceived relativism*
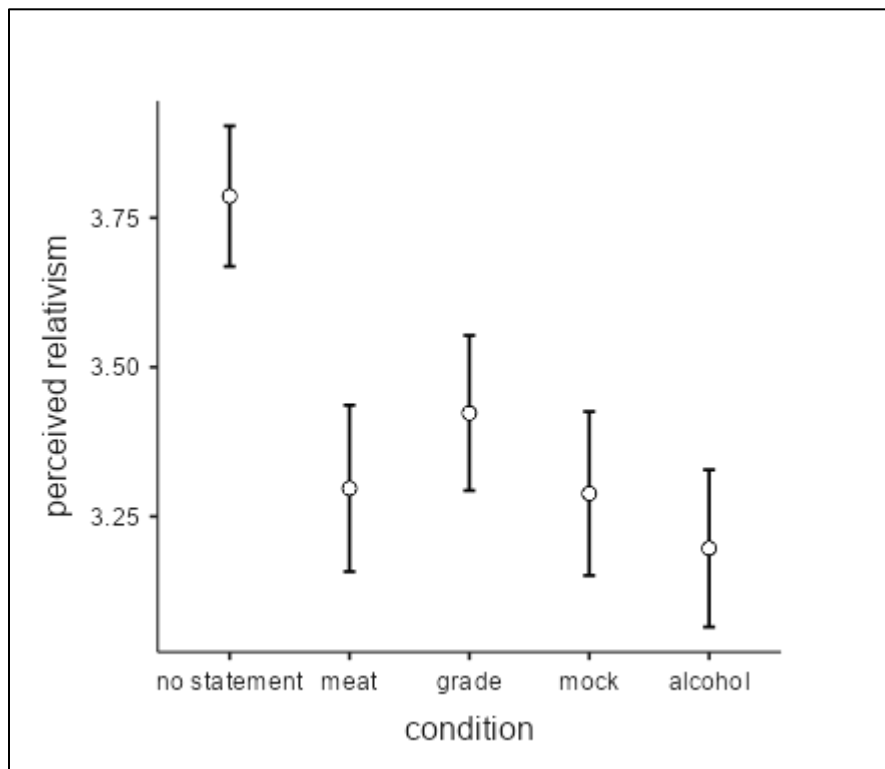


In addition to the three metaethical measures, I also assessed character judgments. First, I conducted a repeated measures ANOVA to compare the effect of statement condition on perceived moral character. Sphericity was violated, $\chi^2(9) = 216.55$, $p < 0.001$. Since $\varepsilon = 0.847$, I report Huynh-

Feldt corrected results. There was a significant main effect of statement condition on perceived universalism, $F(3.41, 2367.41) = 112, p < .001$ , $\eta^2_p = .139$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.10**. Descriptive statistics are available on **Table S6.2.11**. Estimated marginal means appear in **Table S6.2.12** and **Figure S6.2.4.**

**Table S6.2.10**

*Post Hoc Comparisons – character*

| Comparison | | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|---|
| character | | character | | | | | |
| no statement | - | meat | 0.2687 | 0.0569 | 695 | 4.721 | < .001 |
| | - | grade | -0.7227 | 0.0629 | 695 | -11.491 | < .001 |
| | - | mock | -0.6667 | 0.0695 | 695 | -9.592 | < .001 |
| | - | alcohol | -0.6968 | 0.0604 | 695 | -11.529 | < .001 |
| meat | - | grade | -0.9914 | 0.0663 | 695 | -14.947 | < .001 |
| | - | mock | -0.9353 | 0.0732 | 695 | -12.782 | < .001 |
| | - | alcohol | -0.9655 | 0.0660 | 695 | -14.629 | < .001 |
| grade | - | mock | 0.0560 | 0.0521 | 695 | 1.076 | 0.847 |
| | - | alcohol | 0.0259 | 0.0507 | 695 | 0.510 | 1.000 |

**Table S6.2.11**

*Descriptives – character*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 696 | 697 | 697 | 697 |
| Mean | 4.65 | 4.39 | 5.37 | 5.32 | 5.35 |
| 95% CI mean lower bound | 4.57 | 4.29 | 5.27 | 5.20 | 5.25 |
| 95% CI mean upper bound | 4.74 | 4.48 | 5.48 | 5.44 | 5.45 |
| Standard deviation | 1.14 | 1.29 | 1.40 | 1.59 | 1.33 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

**Table S6.2.12**

*Estimated Marginal Means – character*

| Character | Mean | SE | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| no statement | 4.65 | 0.0434 | 4.57 | 4.74 |
| meat | 4.39 | 0.0489 | 4.29 | 4.48 |
| grade | 5.38 | 0.0530 | 5.27 | 5.48 |
| mock | 5.32 | 0.0603 | 5.20 | 5.44 |
| alcohol | 5.35 | 0.0504 | 5.25 | 5.45 |

**Figure S6.2.4**

*Estimated marginal means for perceived character*
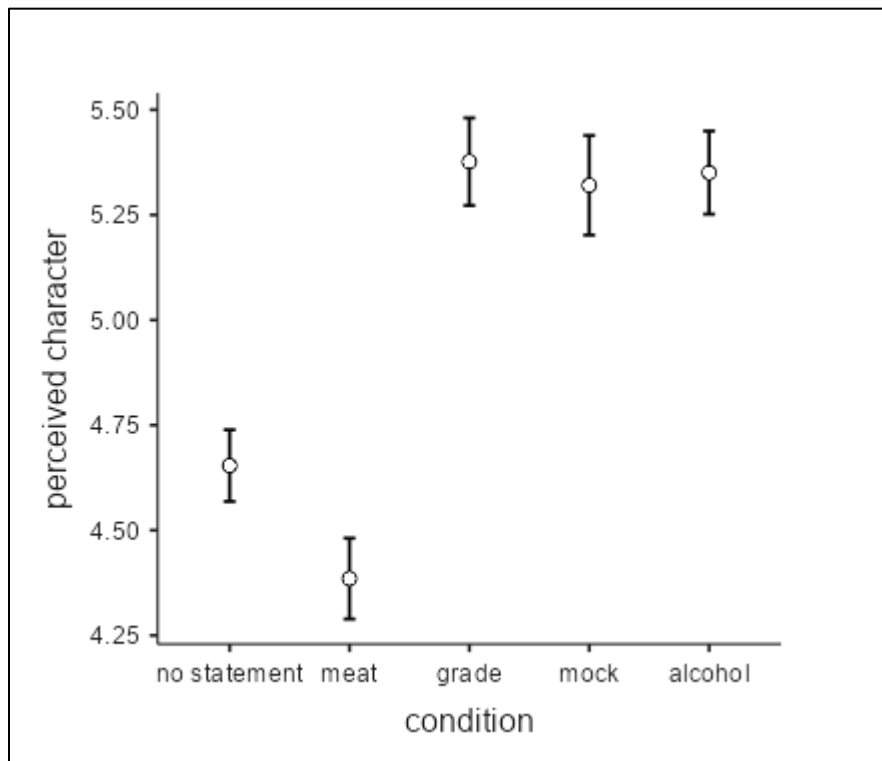


Next, I conducted a repeated measures ANOVA to compare the effect of statement condition on perceived empathy. Sphericity was violated, $\chi^2(9) = 222.55$, $p < .001$ Since $\varepsilon = 0.856$, I report Huynh-Feldt corrected results. There was a significant main effect of statement condition on

perceived universalism, $F(3.44, 2394.98) = 91.2$ , $p < .001$ , $\eta^2_p = 0.116$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.13**. Descriptive statistics are available on **Table S6.2.14**. Estimated marginal means appear in **Table S6.2.15** and **Figure S6.2.5.**

**Table S6.2.13**

*Post Hoc Comparisons – empathy*

| Comparison | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|
| empathy | empathy | | | | | |
| no statement | - meat | 0.0775 | 0.0751 | 696 | 1.03 | 0.302 |
| | - grade | -0.8623 | 0.0649 | 696 | -13.29 | < .001 |
| | - mock | -0.9842 | 0.0739 | 696 | -13.32 | < .001 |
| | - alcohol | -0.5179 | 0.0641 | 696 | -8.08 | < .001 |
| meat | - grade | -0.9397 | 0.0798 | 696 | -11.77 | < .001 |
| | - mock | -1.0617 | 0.0879 | 696 | -12.08 | < .001 |
| | - alcohol | -0.5954 | 0.0790 | 696 | -7.54 | < .001 |
| grade | - mock | -0.1220 | 0.0563 | 696 | -2.17 | 0.061 |
| | - alcohol | 0.3443 | 0.0612 | 696 | 5.62 | < .001 |

**Table S6.2.14**

*Descriptives – empathy*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 697 |
| Mean | 4.28 | 4.21 | 5.14 | 5.27 | 4.80 |
| 95% CI mean lower bound | 4.19 | 4.08 | 5.04 | 5.14 | 4.69 |
| 95% CI mean upper bound | 4.37 | 4.33 | 5.25 | 5.39 | 4.91 |
| Standard deviation | 1.22 | 1.73 | 1.42 | 1.71 | 1.52 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.
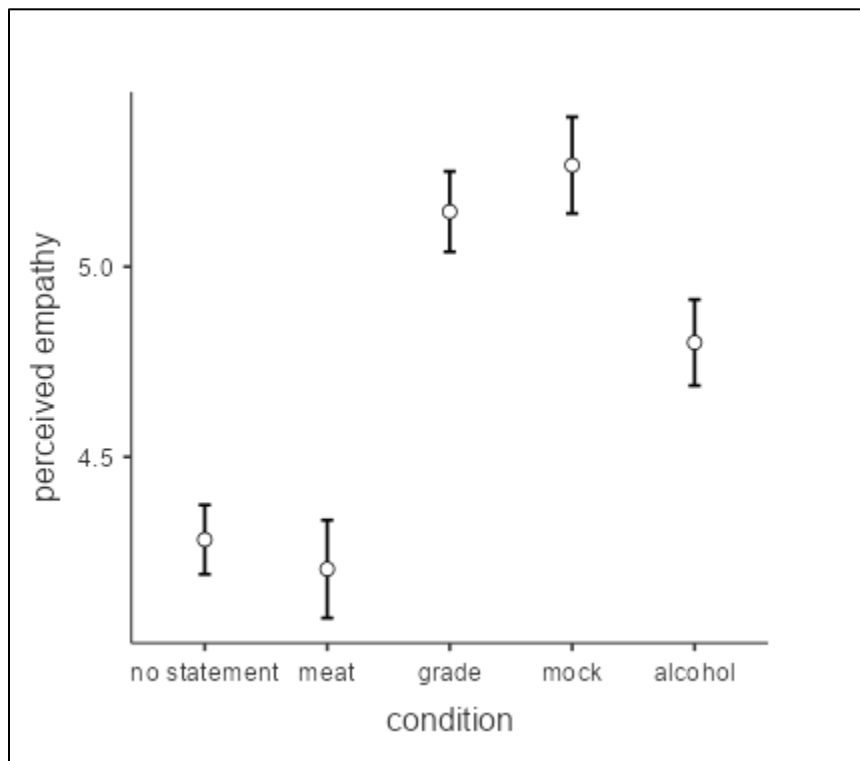
**Table S6.2.15**

*Estimated Marginal Means – empathy*

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| Character | Mean | SE | Lower | Upper |
| no statement | 4.28 | 0.0464 | 4.19 | 4.37 |
| meat | 4.21 | 0.0655 | 4.08 | 4.33 |
| grade | 5.14 | 0.0538 | 5.04 | 5.25 |
| mock | 5.27 | 0.0646 | 5.14 | 5.39 |
| alcohol | 4.80 | 0.0576 | 4.69 | 4.91 |

**Figure S6.2.5**

*Estimated marginal means for perceived empathy*



I also conducted a repeated measures ANOVA to compare the effect of statement condition on perceived moral seriousness. Sphericity was violated, $\chi^2(9) = 237.47$, $p < .001$. Since $\varepsilon = 0.857$, I report Huynh-Feldt corrected results. There was a significant main effect of statement condition on

perceived universalism, $F(3.45, 2398.16) = 44.6$, $p < .001$ , $\eta^2_p = .060$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.16**. Descriptive statistics are available on **Table S6.2.17**. Estimated marginal means appear in **Table S6.2.18** and **Figure S6.2.6.**

**Table S6.2.16**

*Post Hoc Comparisons – seriousness*

| Comparison | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|
| **Seriousness** | **seriousness** | | | | | |
| no statement | - meat | -0.4878 | 0.0722 | 0.0722 | -6.752 | < .001 |
| | - grade | -0.6399 | 0.0634 | 0.0634 | -10.086 | < .001 |
| | - mock | -0.6643 | 0.0718 | 0.0718 | -9.254 | < .001 |
| | - alcohol | -0.8106 | 0.0635 | 0.0635 | -12.763 | < .001 |
| meat | - grade | -0.1521 | 0.0705 | 0.0705 | -2.157 | 0.082 |
| | - mock | -0.1765 | 0.0798 | 0.0798 | -2.212 | 0.082 |
| | - alcohol | -0.3228 | 0.0720 | 0.0720 | -4.486 | < .001 |
| grade | - mock | -0.0244 | 0.0511 | 0.0511 | -0.477 | 0.633 |
| | - alcohol | -0.1707 | 0.0524 | 0.0524 | -3.260 | 0.006 |
| mock | - alcohol | -0.1463 | 0.0595 | 0.0595 | -2.458 | 0.057 |

**Table S6.2.17**

*Descriptives – seriousness*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 697 |
| Mean | 4.72 | 5.21 | 5.36 | 5.39 | 5.54 |
| 95% CI mean lower bound | 4.63 | 5.09 | 5.26 | 5.27 | 5.43 |
| 95% CI mean upper bound | 4.82 | 5.33 | 5.47 | 5.51 | 5.64 |
| Standard deviation | 1.28 | 1.62 | 1.39 | 1.58 | 1.39 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.
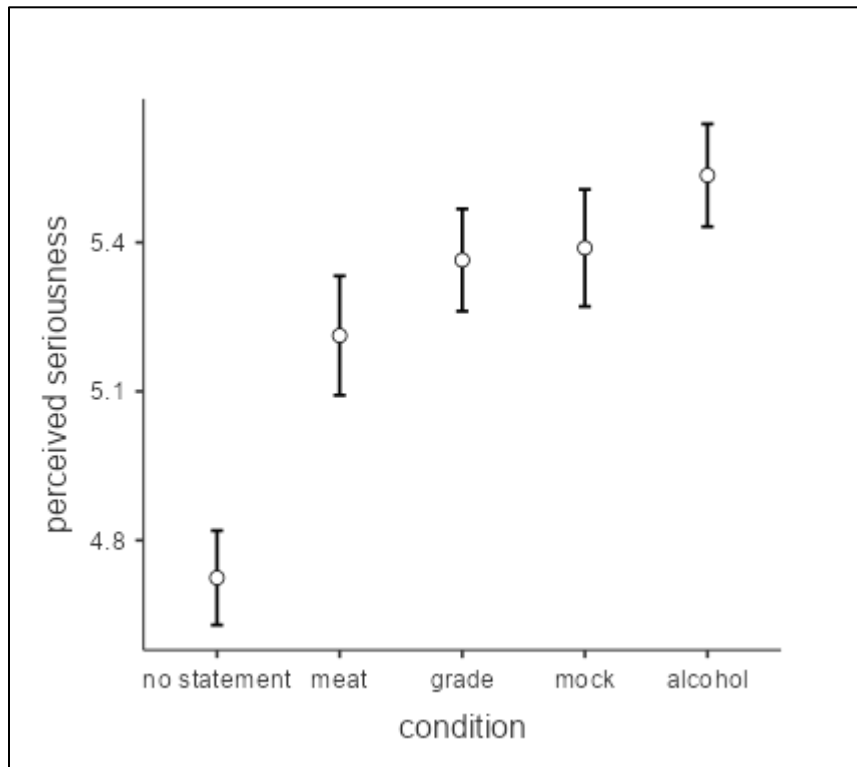
**Table S6.2.18**

*Estimated Marginal Means – seriousness*

| Character | Mean | SE | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| no statement | 4.72 | 0.0484 | 4.63 | 4.82 |
| meat | 5.21 | 0.0612 | 5.09 | 5.33 |
| grade | 5.36 | 0.0525 | 5.26 | 5.47 |
| mock | 5.39 | 0.0600 | 5.27 | 5.51 |
| alcohol | 5.54 | 0.0527 | 5.43 | 5.64 |

**Figure S6.2.6**

*Estimated marginal means for perceived moral seriousness*



Lastly, I conducted a repeated measures ANOVA to compare the effect of statement condition on desirability as a social partner. Sphericity was violated, $\chi^2(9) = 321.16$, $p < .001$. Since ε

= 0.793, I report Huynh-Feldt corrected results. There was a significant main effect of statement condition on perceived universalism, $F(3.19, 2216.28) = 142$, $p < .001$ , $\eta^2_p = .170$. Holm-adjusted pairwise comparisons may be seen on **Table S6.2.19**. Descriptive statistics are available on **Table 6.2.20**. Estimated marginal means appear in **Table S6.2.21** and **Figure 6.2.7.**

**Table S6.2.19**

*Post Hoc Comparisons – partner preference*

| Comparison | | | Mean Difference | SE | df | t | $p_{holm}$ |
|---|---|---|---|---|---|---|---|
| partner preference | | partner preference | | | | | |
| no statement | - | meat | 0.713 | 0.0666 | 695 | 10.69 | < .001 |
| | - | grade | -0.684 | 0.0636 | 695 | -10.75 | < .001 |
| | - | mock | -0.553 | 0.0710 | 695 | -7.79 | < .001 |
| | - | alcohol | -0.447 | 0.0633 | 695 | -7.06 | < .001 |
| meat | - | grade | -1.397 | 0.0761 | 695 | -18.35 | < .001 |
| | - | mock | -1.266 | 0.0854 | 695 | -14.82 | < .001 |
| | - | alcohol | -1.159 | 0.0754 | 695 | -15.37 | < .001 |
| grade | - | mock | 0.131 | 0.0515 | 695 | 2.54 | 0.023 |
| | - | alcohol | 0.237 | 0.0526 | 695 | 4.51 | < .001 |
| mock | - | alcohol | 0.106 | 0.0619 | 695 | 1.72 | 0.087 |

**Table S6.2.20**

*Descriptives – partner preference*

| | no_statement | meat | grade | mock | alcohol |
|---|---|---|---|---|---|
| N | 697 | 697 | 697 | 697 | 696 |
| Mean | 4.47 | 3.75 | 5.15 | 5.02 | 4.91 |
| 95% CI mean lower bound | 4.37 | 3.64 | 5.05 | 4.90 | 4.80 |
| 95% CI mean upper bound | 4.56 | 3.87 | 5.25 | 5.14 | 5.02 |
| Standard deviation | 1.28 | 1.54 | 1.37 | 1.63 | 1.45 |

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.
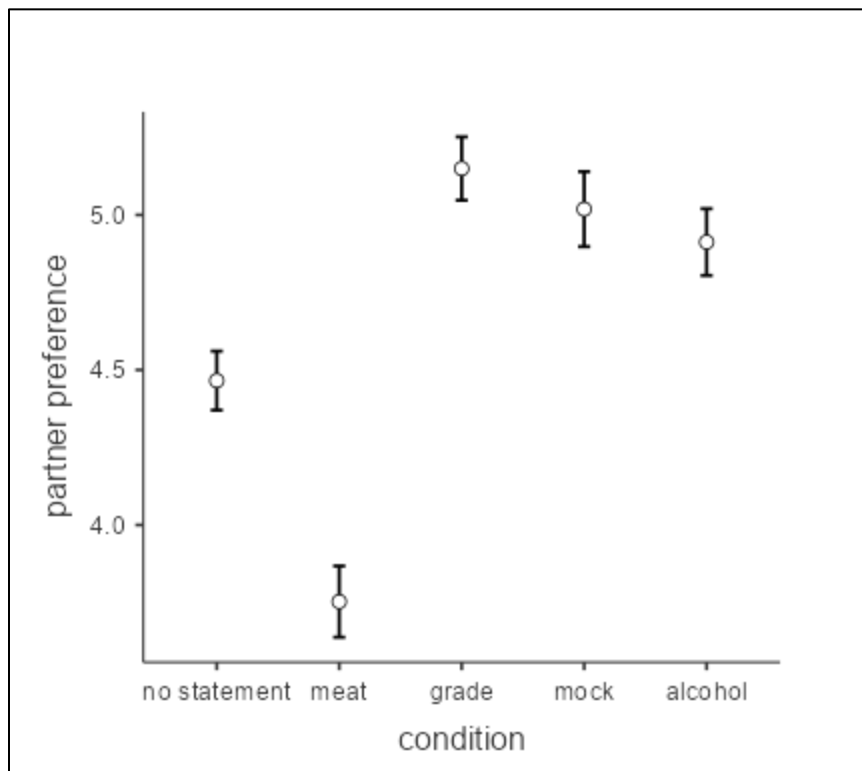
**Table S6.2.21**

*Estimated Marginal Means – partner preference*

| Character | Mean | SE | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| no statement | 4.47 | 0.0485 | 4.37 | 4.56 |
| meat | 3.75 | 0.0585 | 3.64 | 3.87 |
| grade | 5.15 | 0.0520 | 5.05 | 5.25 |
| mock | 5.02 | 0.0616 | 4.90 | 5.14 |
| alcohol | 4.91 | 0.0548 | 4.80 | 5.02 |

**Figure S6.2.7**

*Estimated marginal means for desirability as a social partner*



**Discussion**

The results of Study 2 largely corroborate the results of Study 1. In Study 1, the mean score for perceived objectivism and universalism was significantly above the midpoint, while perceived

relativism was significantly below the midpoint, indicating that participants were generally inclined to regard others in their society as having a tendency towards objectivism and universalism and against relativism. These results were obtained in spite of the fact that participants didn't perceive *themselves* this way, and instead were more ambivalent about endorsing objectivism, universalism, and exhibited a mild tendency to endorse relativism on average.

The results of Study 2 are generally consistent with these findings. Participants judged a typical person in their society to be inclined towards realism and universalism, and disinclined towards relativism, regardless of whether they made no statement at all, or expressed a moral stance towards one of the four concrete moral issues. Yet these effects were stronger for some of the concrete moral issues. In particular, participants were especially likely to perceive someone to endorse objectivism and universalism, and to reject relativism, when that person asserted that it is wrong to give a student a bad grade just because they dislike the student, to go to a funeral to mock the person who died in front of their family, and for a woman who knows she is pregnant to drink alcohol. Perceived objectivism and universalism were lower and perceived similarly to one another in the no statement eating meat conditions.

However, perceived relativism was about the same for all of the concrete conditions (including eating meat), all of which resulted in substantially lower perceived relativism scores than the no statement condition. This could indicate that expressing a concrete moral stance towards any moral issue at all has, all else being equal, some tendency to encourage the perception of objectivism and universalism, even if this perception can vary based on the content of the moral issue. However, it could also simply indicate comparatively greater hesitance in attributing metaethical stances to people in the absence of any specific information about their moral views.

Participants also judged various character traits in the no statement and concrete conditions. Although the mean score for perceived character was above the midpoint in all conditions, it was

lowest for the condition in which participants were asked to judge a person who said it was morally wrong to eat meat and was significantly lower than the no statement condition. The *meat* condition performed at about the same level as the no statement condition for perceived empathy, both just barely above the midpoint. All three of the other concrete moral issues prompted greater perceived empathy, suggesting a similar pattern to perceived character. This same pattern was strongest for partner preference. Participants were asked to judge how good the person in the no statement condition or a person making each of the four concrete moral statements would be to have as a coworker, neighbor, or close friend. Across all conditions, perceived desirability as a social partner was above the midpoint, indicating that people had, on average, a positive perception of the social desirability of the target, *except for* the person who said it was wrong to eat meat. Not only was perceived social desirability significantly lower in this condition than all others (including the no statement condition), it was the *only* instance of a perceived character trait dropping below the midpoint out of all of the measures.

In contrast, this disparity between the *meat* condition and the three other conditions was not present for perceived moral seriousness, where participants judged the person who made the statement to be significantly more morally serious than in the no statement condition, but mean levels of perceived seriousness did not differ by much across conditions. Overall, this pattern of results suggests that participants are especially likely to derogate someone's character when that person expresses moral opposition to eating meat. There are several reasons why this might be the case. Whatever the cause, it is unlikely that a significant proportion (much less a majority) of participants are vegans or vegetarians. As such, it is likely most participants *themselves* eat meat, and are unlikely to regard it as a serious moral transgression (unlike the other concrete moral issues). As such, the claim that it is morally wrong to eat meat would reflect negatively on participants themselves, given the high probability that many participants regularly eat meat. Disagreement with the moral claim in question

may be a factor here, though cognitive dissonance and other factors that attenuate (or reverse) perceptions of positive moral character, empathy, and social desirability may also play a role.

These results illustrate that the effects in Study 1 are perhaps weaker than what we'd find if we provided greater context. In this case, that context consisted of the expression of various concrete moral stances. Such findings tended to be stronger than the no statement condition, and the same may hold if participants are given other information the target or population they are asked to judge. Future studies could explore perceptions of metaethical stances and character traits under a variety of circumstances where participants are given more context or richer details. This could include testing a variety of other moral claims, providing more detailed more claims, or providing more information about the person making the claim. It could also involve more detailed vignettes that provide context to the moral judgment itself, rather than a sparse and general assertion about some general category of moral transgression.

In addition, future studies could explore judgments about those making a variety of moral claims other than claims about the wrongness of an action. This could include judgments that an action is morally good, morally permissible, morally required, and so on. Variation in language, and the employment of thick moral concepts (e.g., "courageous" or "generous") could likewise yield fascinating insights into people's perception of the metaethical standards of other people. Novel insights may also be obtained by evaluating members of other cultures, people in the past, what people will think in the future, or what nonhuman civilizations or artificial intelligences might think. Given the almost total absence of previous research on perceived metaethical stances, there are undoubtedly many ways to expand on the present findings.

**S6.3 Additional Commentary**

*S6.3.1 Inter-domain comparisons and the disagreement paradigm*

In the main text, I point out that most studies that have made cross-domain comparisons have employed the disagreement paradigm. Why has this occurred? I suspect that researchers want to maintain balanced stimuli, but rely exclusively on using the same phrasing in their stimuli across domains, so as to minimize differences in interpretation due to the use of different wording, sentence structure, and so on. For instance, consistency in stimuli may prompt us to present word-for-word descriptions of a disagreement between Alex and Sam, and simply swap out a bracketed portion of the text that references murder for a statement about the age of the earth. The presumption is that if the wording of the stimuli is otherwise identical, it will be interpreted in the same way. Yet this is not necessarily true. What researchers fail to appreciate is that meaning is not determined solely by maintaining as much of the syntactic structure and terminology as possible. People interpret language *holistically*, so *any* change in the content of the stimuli, such as a shift from a disagreement about a moral issue to a scientific or historical issue, could alter how people interpret the rest of the text extraneous to the altered content of the sentence. In other words, if participants are presented with the following stimuli:

Alex and Sam disagree about [murder].

Alex and Sam disagree about [the age of the earth].

…there is no guarantee that participants would reliably interpret "disagree" in the same way, because changing the content in brackets can change how they interpret *the rest of the sentence*, including what it means for Alex and Sam to "disagree," along with inferences about why they might disagree and the nature of the disagreement. To provide one illustration of how the meaning of "disagree" can vary in this way, consider the following statements:

Alex and Sam disagree about [what time the movie starts].

Alex and Sam disagree about [what to have for dinner].

In this case, the first disagreement concerns a factual dispute about a descriptive matter. However, the second does not. It's not likely that Alex and Sam think there is a "correct" answer regarding what to have for dinner. Rather, their "disagreement" is a conflict in goals. As such, it is not a dispute about what's true or false, but a coordination problem. The meaning of "disagreement" changed based on the content of what they're disagreeing about, despite holding the rest of the structure and wording of the sentences constant. This illustrates how the meaning of a word isn't fixed and rigid, but is determined by the rest of the context in which it's embedded. When researchers design stimuli, they often pay little or no attention to this fact, and simply cut out part of a sentence and replace it with other stimuli, then presume that because the wording was held constant, that participants will interpret the statements in the same way. This is not true, and could be easily refuted with a proper study to illustrate the point. This creates a dilemma for researchers: if they want to ensure cross-domain consistency, preserving wording may be insufficient, since interpretation could vary. But any change in wording in an attempt to hold interpretation constant may require changes that create imbalances in the stimuli. I see no easy solution to this. The holistic nature of language may serve as a practically insurmountable barrier to certain aspects of conventional survey design.

### S6.3.2 Problems with characterizing the "factual" domain

This, too, is not without shortcoming. Naturalists and some antirealists may regard moral claims as descriptive or reducible to descriptive claims as well. It also won't be called the "physical" or "material" domain. While it does typically include claims that could be understood in this way, e.g., claims about science or history, it could also include claims about mathematics. I have no satisfactory account of this domain. I think it would be best to regard it not as a domain for which there is some principled distinction between it and other domains, but is rather distinguished as a kind of loose family resemblance of claims that fall outside the scope of other, more well-defined domains.

### S6.3.3 The proper negation of denying a first-order normative moral claim

In other words, the proper negation of the claim that "murder is impermissible" isn't "murder is permissible," it's that "it's not the case that murder is impermissible." And the claim that it's not the case that murder is impermissible is *not* the logical equivalent of the claim that murder is permissible: you could believe that murder is neither permissible nor impermissible. For comparison, I don't think the number "7" tastes bad. This does not commit me to believing it tastes good.

### S6.3.4 Antirealism and error theory

It also doesn't require a commitment to error theory. Error theorists are often restricted in the sorts of objections they can raise to realists, since they concede that normative moral discourse conforms with the realist's use of it. As such, they lack a non-revisionary way of asserting that actions can be right or wrong, or good or bad. Yet antirealists are not obligated to view language this way. Such positions tacitly operate within the presuppositions of contemporary analytic philosophy, which conventionally abide by questionable assumptions about how language and meaning work. For instance, they may operate under the presumption that normative terms like "should" and "good," have rigid and fixed meanings, such that there is a single correct analysis of their metaethical implications. An antirealist could reject this, and maintain that they are not making any mistakes or speaking in any sort of meaningfully "revisionary" way when they say that some actions are bad or wrong. Doing so does not *require* them to endorse any particular semantic thesis about the meaning of ordinary moral language.