Cognition 211 (2021) 104623

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Motivated moral judgments about freedom of speech are constrained by a need to maintain consistency

Nikolai Haahjem Eftedal [a,*], Lotte Thomsen [a,b]

[a] Department of Psychology, University of Oslo, Oslo, Norway
[b] Department of Political Science, Aarhus University, Aarhus, Denmark

ARTICLE INFO

Keywords:
Motivated reasoning
Moral judgment
Freedom of speech
Self-deception
Social dominance
Political ideology

ABSTRACT

Speech is a critical means of negotiating political, adaptive interests in human society. Prior research on motivated political cognition has found that support for freedom of speech depends on whether one agrees with its ideological content. However, it remains unclear if people (A) openly hold that some speech should be more free than other speech; or (B) want to feel as if speech content does not affect their judgments. Here, we find support for (B) over (A), using social dominance orientation and political alignment to predict support for speech. Study 1 demonstrates that if people have previously judged restrictions of speech which they oppose, they are less harsh in condemning restrictions of speech which they support, and vice versa. Studies 2 and 3 find that when participants judge two versions of the same scenario, with only the ideological direction of speech being reversed, their answers are strongly affected by the ordering of conditions: While the first judgment is made in accordance with one's political attitudes, the second opposing judgment is made so as to remain consistent with the first. Studies 4 and 5 find that people broadly support the principle of giving both sides of contested issues equal speech rights, also when this is stated abstractly, detached from any specific scenario. In Study 6 we explore the boundaries of our findings, and find that the need to be consistent weakens substantially for speech that is widely seen as too extreme. Together, these results suggest that although people can selectively endorse moral principles depending on their political agenda, many seek to conceal this bias from others, and perhaps also themselves.

1. Introduction

"Goebbels was in favor of free speech for views he liked. So was Stalin. If you're really in favor of free speech, then you're in favor of freedom of speech for precisely the views you despise. Otherwise, you're not in favor of free speech." – Noam Chomsky.

"If we continue to teach about tolerance and intolerance instead of good and evil, we will end up with tolerance of evil" – Dennis Prager.

Someone spreading controversial beliefs or attitudes is likely to face consequences. However, most societies have rules, both written and informal, regulating how harsh the consequences can be. These rules will likely not prevent people from countering a message through civil and honest presentations of arguments and evidence. But they could, to some extent, restrain people from hurting others socially, financially, or physically, or from taking away their platforms. Even people who despise a view that is being spread might object if they feel others are breaking these rules for how to legitimately counter it. We will refer to these rules for how to counter speech as *speech norms.*

The world is currently undergoing a revolution in information technology, with the internet and social media expanding most people's audiences by orders of magnitude. Norms for how information should be spread and restrained are thus increasingly relevant. As highlighted by recent controversies (see e.g. Bey, 2017; TheFire, 2020), there is no clear consensus on what exactly these norms should say. For example, some hold that the removal of certain platforms is legitimate (e.g., Munroe, 2014), while others disagree (e.g. Campbell, 2018). Similarly, attempts at getting people fired for their views are applauded by some (e.g. Bey) and decried by others (e.g., Crook, 2017).

A central question is whether speech content should be morally relevant when making judgments about tactics to counter speech. For instance, does the blameworthiness of sabotaging a speaking event depend on the message being delivered there? The recurring finding from experimental psychological studies examining these kinds of judgments (Brandt, Chambers, Crawford, Wetherell, & Reyna, 2015; Crawford, 2014; Crawford & Pilanski, 2014; Crawford & Xhambazi, 2015; Lindner & Nosek, 2009; White II & Crandall, 2017) is that

* Corresponding author at: Department of Psychology, University of Oslo, Forskningsveien 3A, 0373 Oslo, Norway.
E-mail address: n.h.eftedal@psykologi.uio.no (N.H. Eftedal).

https://doi.org/10.1016/j.cognition.2021.104623
Received 14 February 2020; Received in revised form 1 February 2021; Accepted 3 February 2021
Available online 16 February 2021

participants indeed seem to take speech content into account: They judge an action to be more permissible when used against speech they oppose than when used against speech they support. To illustrate, Crawford (2014) found that conservatives were more inclined to agree that abortion activists should be banned from distributing pamphlets and fliers on a campus when described as pro-choice rather than pro-life, while liberals showed the opposite pattern.

However, it remains unclear from prior studies if.

(A) Participants are explicitly endorsing *content-sensitive speech norms* saying that the blameworthiness of actions to counter speech depends on what that speech is. If so, prior results simply reflect participants' honest opinions that certain views should have less protection than others.

(B) Participants seek to maintain a perception that they support *content-neutral speech norms* saying that actions to counter speech are equally blameworthy, regardless of what is being said. If so, prior results reflect selective, context-specific adjustments of this general level of blame for restricting speech: If a participant supports/opposes the speech in question, they will take the stance that it is *always* wrong/right to counter speech in the specific way they just evaluated.

Here, we seek to disambiguate between these two interpretations. To the extent that prior studies have interpreted results to be due to either (A) or (B), (A) - that people are openly content-sensitive - has mostly been emphasized. And content-sensitive speech norms have certainly existed in many civilizations throughout history (e.g., against blasphemy), both as instantiated through laws and through social sanctions (Graham, 2004; Green & Karolides, 2014). Yet, we favor interpretation B. As argued by e.g. DeScioli and Kurzban (2013), a common feature of moral systems is that rules should be universal and equally applied across cases, with only a minimum of caveats and exceptions. This is a sentiment that has deep evolutionary- and developmental roots (Boehm, 2009; see Thomsen, 2020, for review). At the same time, there is a constant tension in morality between cooperation for the common good and defection to serve other interests (Wilson, 1998), which might have led to an evolutionary arms race in the domain of social negotation and manipulation (Dawkins & Krebs, 1979; Trivers, 1971). Applied to the context of speech protection, the winning strategy could then be to *appear as if* applying rules similarly across all kinds of speech while actually making judgments in a partisan manner, to the extent that this can be plausibly denied. Thus, while most people prefer some views to win out over others (as shown in prior research), we predict that they will still want to see and portray themselves as supporting norms that give similar protections to all (or at least most) views. When plausible deniability is removed, this could then entail making judgments that are fully content-neutral, in line with B, or at least making judgments that are less partisan than they otherwise would have been. Before elaborating on this prediction and how we test it, we will first discuss the value of being able to shape what information other people are, and are not, exposed to.

1.1. Ideology

Views affect how people behave, and which policies and norms they put in place. Whenever people act in ways that make things worse, or fail to act in ways that make things better, they might have acted differently if their views were different. So, from a purely impartial standpoint, there are very good reasons to care what other people think: Changing people's minds can improve the world. And preventing minds from being changed can keep the world from deteriorating. (Note that sorting views into good and bad is different from sorting into true and false; see e.g. Boström, 2011).

Interestingly, people may have quite contrasting opinions on which views should and should not be spread. This could reflect honest disagreements about how to best reach goals we all share. It could also reflect conflicts of interests when our goals are not fully shared (see Alexander, 2018): Altering other people's views can serve selfish and parochial goals as well as altruistic and impartial ones. And the beliefs and values that would make the world better on the whole are not necessarily the same as those that serve your personal preferences: The world is full of trade-offs where improving conditions for some often means making things worse for others.

Societies typically have several ideological groups working to shift trade-offs to their benefit. As implied by the word "ideological", this work involves influencing the flow of ideas (Greene, 2014; Haidt, 2012). Being accepted into such groups has benefits, and so taking part in these actions to spread or restrict views may also be motivated by a desire to signal one's loyalty to the group in question (Kahan, 2015, cf. Greene, 2014).

However, which groups one seeks acceptance from in the first place likely reflects the degree to which their ideological goals overlap with one's own (cf. e.g., Blumer, 1958; LeVine & Campbell, 1972; Tooby & Cosmides, 2010).

There are typically many more conflicts of interests in a society than there are ideological groups. Groups will tend to have multiple agendas at once, and each single member might not care equally about all of these. Social Dominance Theory addresses one of the fundamental, general ideological conflicts, around which groups tend to form and legitimize their agenda, namely that of societal hierarchy versus equality between groups.

1.1.1. Social dominance orientation and legitimizing myths

Social Dominance Theory suggests that the root of many conflicts in moral and ideological matters are the dominance hierarchies that are ubiquitous across human, surplus-producing societies and individual differences in the motives to either sustain or attenuate them (Kleppestø et al., 2019; Sidanius, 1993; Sidanius & Pratto, 2004). Indeed, philosophers from Thucydides and Hobbes to Foucault posit that politics is undergirded by the question of who will dominate whom (McClelland, 1996), and social dominance is represented and motivates affiliation even among infants (Thomas, Thomsen, Lukowski, Abramyan, & Sarnecka, 2018; Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011). The Social Dominance Orientation (SDO) scale (Ho et al., 2012; Ho et al., 2015; Pratto, Sidanius, Stallworth, & Malle, 1994) captures individual differences in preferences for intergroup hierarchies.

Social Dominance Theory further suggests that motives regarding hierarchies - whether one wants to sustain or attenuate them - lead to the production and perpetuation of *legitimizing myths*. Legitimizing myths are narratives and truth-claims that serve to justify one's social and political preferences. High SDO is associated with support for hierarchy-enhancing legitimizing myths (Ho et al., 2012; Lucas & Kteily, 2018; Thomsen et al., 2010) which function to justify hierarchies, be they true or false. Examples include the belief that the current distribution of status and power in society has come about through meritocratic processes, or that negative stereotypes about low status groups are true. Conversely, low SDO predicts support for hierarchy-attenuating legitimizing myths, such as feminism and anti-racism, which discredit the validity of intergroup hierarchies.

The idea that people seek to "legitimize" their preferences is echoed across the literature on motivated reasoning (Ditto et al., 2018; Kunda, 1990; Mercier & Sperber, 2011; Simler & Hanson, 2017; Von Hippel & Trivers, 2011). While one's true motives might be to serve a certain agenda, people generally seek to present themselves as caring mostly about the truth and the common good. Being open about having biased reasons for trying to influence others will typically make one less persuasive. In this sense, one might argue that politics is largely about persuading others that general rules that happen to benefit one's own interests also serve common interests, thus appealing to everyone for support (Petersen, 2015). Trivers further argues that it is in many ways useful to not only deceive others about one's strategic motives, but to keep one's own conscious mind in the dark as well; to be "self-deceived" (Trivers, 2000, 2011). Being self-deceived is less cognitively demanding than being dishonest; it reduces risks of being judged as a liar; and it helps you present a more convincing case for your position.

1.2. Maintaining consistency

A preference for selective protection of certain kinds of speech is perhaps the kind of preference that remains mostly hidden from the conscious mind. It is difficult to legitimize such a policy to an audience that includes people who support viewpoints that would then be protected *less*.

People could then instead portray themselves as supporting largely content-neutral speech norms, where a wide range of views are protected to a similar extent. Such norms are what you would prefer your ideological opposition to support. Content-sensitive norms saying that "bad views should have less protection" are problematic if practiced by people who disagree with you about which views are good and bad. They might use these norms to justify silencing views you support and protecting views you oppose. Endorsing shared principles of equal protection is a way to compromise.

There are reasons to prefer content-neutral speech norms also among your ideological allies, not just your opposition. John Stuart Mill voiced several such reasons in "On Liberty" (Mill, 1859/1966), which have since been echoed by others (Alexander, 2014a; Galef, 2018). For example, Mill (p. 24) argues that silencing wrongful views deprives people of "the clearer perception and livelier impression of truth, produced by its collision with error." Additionally, content-neutral norms could function as a safe-guard against the eventuality that power might at some point fall into the wrong hands (Brennan, 2015), or that suppressed views might "go underground" only to emerge again later in a more destructive form (Pinker, 2007). Nevertheless, it is apparent from the existing experimental record that many people have not been swayed into adopting content-neutral norms from arguments such as these: Otherwise they would not have systematically judged restrictions of speech they oppose to be more permissible than restrictions of speech they support (Brandt et al., 2015; Crawford, 2014; Crawford & Pilanski, 2014; Crawford & Xhambazi, 2015; Lindner & Nosek, 2009; White II & Crandall, 2017).

Importantly, it cannot be conclusively shown from prior between-subject studies that any one specific participant was in fact being content-sensitive in their judgments; this can only be inferred when looking at all responses in conjunction. For each response, participants can claim that they would indeed have responded exactly the same had the ideological content of speech been reversed. This type of psychological wiggle room is common in the real world as well; no two situations are exactly alike, so you can usually point to unique features unrelated to speech content to justify judging in ways that benefit your own position. As long as you can maintain plausible deniability, you can get away with judging selectively while still publicly endorsing content-neutral norms. Such a strategy is only effective, however, so long as its double standard is not laid bare - for others, nor perhaps even yourself (Trivers, 2011).

1.3. Present study

Here, we examine whether people seek to hide their tendency to give more freedom to speech they support (a tendency that has been demonstrated in prior research and which we replicate here). Specifically, we investigate effects of altering the ordering of the speech participants make judgments about. In Study 1, we measure participants' SDO and ask them to respond to scenarios depicting restrictions of speech, where speech content is manipulated to serve goals related to having either high or low SDO. We then analyze whether participants become more supportive of rights for speech they dislike if they have previously judged speech they support, and vice versa. A façade of neutrality is harder to maintain if certain types of speech are systematically given less protection than other types of speech, and so we predict that the bias in free speech judgments shown in the prior literature should be dampened by having first made judgments for the opposite kind of speech.

In Study 2, we manipulate the ideology of speech *within subjects*, such that participants first judge a scenario with one type of speech, and then judge the *exact same* scenario except that the ideology of speech content is reversed. Participants are unaware that the second scenario is coming when responding to the first, and they are unable to alter their responses to the first upon seeing the second. In the terminology introduced by Hsee, Loewenstein, Blount, and Bazerman (1999), the first response is then a *separate evalution* of a single scenario, while the second response is a *joint evaluation* of both scenarios.

Under hypothesis A above, where participants openly endorse that some speech should be less protected than other speech ("content-sensitive speech norms"), the order of presentation should not matter: Participants should simply judge speech suppression differently depending on speech content. On the other hand, if a participant acts in accordance with B, and seeks to maintain the narrative that they are content-neutral in their judgments, then the ordering of scenarios will affect responses to both of them. By having scenarios be exactly matched on everything but the ideology of speech content, participants lose the ability to plausibly claim that content-sensitive judgments are in fact content-neutral. To maintain the perception of content-neutrality, responses to the second scenario must be similar to those for the first. This means that the ordering of conditions will affect both responses, so that either both become more lenient (if the speech in the first scenario was opposed by the participants), or both become more harsh (if the speech in the first scenario was supported). See Fig. 1 for an illustration of these kinds of order effects.

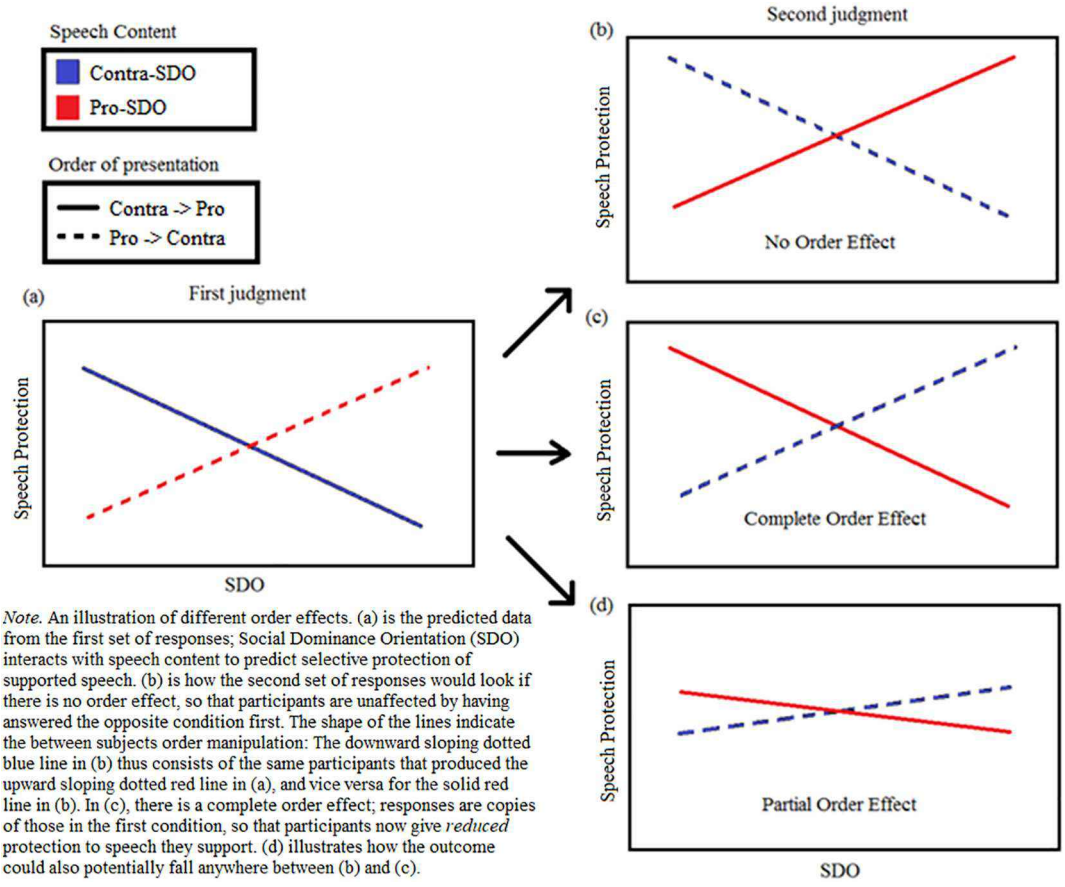
Study 3 is a pre-registered replication of Study 2, which also explores the potential moderating roles of introspectiveness and two measures of moral integrity. Both introspectiveness and integrity could dampen the tendency to shift judgments depending on speech content, if it is indeed the case that this bias is the result of processes that are on the fringes of conscious access (and thus more accessible to the introspective) and which are morally questionable (such that people with more integrity will be less likely to follow them).

In Studies 4 and 5 we investigate the support for content-neutrality versus content-sensitivity when these principles are stated in the abstract, rather than being tied to a specific issue in a particular context. And we also test the prediction that support for content neutral norms is heigthened if they are proposed by one's ideological opposition, which follows from the idea that the reason people seek to appear content-neutral is their hope that their opposition will return the favor.

Finally, in study 6, we explore the potential boundaries of the order effect on judgments. Based on the idea there is an "Overton Window" of acceptable opinions (Lehman, 2012), we predict that participants will no longer strive to seem consistent in their judgments when one of the two relevant viewpoints is widely seen as unacceptable. In that case, the ordering of conditions should no longer have as much influence on responses, and participants should be more open about holding different standards for different viewpoints.

2. Study 1

In Study 1 we use the SDO_7 scale (Ho et al., 2015) to predict participants' judgments of three scenarios describing the use of different tactics for restraining the spread of views: online shaming, selective standards of evidence (Alexander, 2014b), and job-firing. We randomly vary the ideological speech content affected by these tactics *between subjects,* to either support goals related to SDO (Pro-SDO condition) or to oppose them (Contra-SDO condition). In order to investigate our two competing hypotheses for interpretation of prior studies, we analyze the effects of prior exposure to scenarios in the opposite ideological condition to the current one, using a randomization scheme which makes this approach viable. Insofar as participants seek to seem consistent in judgments regardless of speech content, exposure to the opposite condition should dampen the selective protection of favored speech observed in prior studies.



Note. An illustration of different order effects. (a) is the predicted data from the first set of responses; Social Dominance Orientation (SDO) interacts with speech content to predict selective protection of supported speech. (b) is how the second set of responses would look if there is no order effect, so that participants are unaffected by having answered the opposite condition first. The shape of the lines indicate the between subjects order manipulation: The downward sloping dotted blue line in (b) thus consists of the same participants that produced the upward sloping dotted red line in (a), and vice versa for the solid red line in (b). In (c), there is a complete order effect; responses are copies of those in the first condition, so that participants now give *reduced* protection to speech they support. (d) illustrates how the outcome could also potentially fall anywhere between (b) and (c).

Fig. 1. Illustration of order effects.

2.1. Participants

We recruited a sample of 300 participants from Amazon's Mechanical Turk, ensuring ~90% power to detect effects of similar magnitude to those observed in prior studies. All power analyses are done using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007).

We excluded 53 participants who failed more than one of our three attention checks. Among the remaining 247 participants (M_{age} = 37.1, SD_{age} = 11.3, 104 females), there were 69% Caucasians, 18% Asian American, 9% African Americans, 2% Hispanics, 1% Native Americans, and 1% Jews. This project complies with, and was conducted in accordance with, ethical standards as a outlined by the Norwegian Centre for Research Data.

2.2. Procedure

Participants gave an informed consent and basic demographic information before filling out the short 8-item version of the SDO_7 scale (α = 0.93, M = 2.39 out of 7, SD = 1.31). They then responded to three questions for each vignette about the legitimacy of the tactic used to counter speech. Vignettes were presented in random order. The speech in the vignettes supported goals associated with either high- or low SDO, depending on the experimental condition. Except for minor adjustments, the three vignettes match the corresponding scenarios in Studies 2 and 3, which are displayed in Box 1 (labeled "Job security", "Academic freedom", and "Online hate"; for the verbatim texts used in Study 1, see the Online Supplement).

For each vignette, participants' responses to the associated statements were averaged (with reversed scoring of reversed items) to form speech protection measures serving as the dependent variables for our analyses (α's > 0.68).

For 166 of the 247 participants (200 of the original 300), we randomized the experimental condition separately for each vignette, so that they were likely to receive some vignettes from each condition. For the remaining 81 participants (100 out of original 300), we rather randomized condition for all vignettes simultaneously, so that all vignettes were in the same condition. Following this randomization scheme, approximately half of responses to each individual vignette would be preceded by responses to another vignette in the other experimental condition (i.e., framed in the opposite ideological direction).

2.3. Results

All analyses and figures across all four studies were done in the statistical computing environment R (R Core Team, 2020). The regression models to investigate selective support of free speech can be seen in Table 1, and the data are visualized in Fig. 2. The significant interactions between SDO and Condition in all three vignettes are consistent with prior work, suggesting that participants indeed support free speech norms selectively, depending on their attitudes to the speech content: They condemn restriction of speech more for content which they support and less for content with which they disagree (see Table 1).

The three-way interaction between SDO, Speech Content and Exposure to Both Conditions (EBC) indicates whether this selective protection-effect changes when the respondent faces an inconsistency issue because s/he has prior exposure to vignettes framed in the opposite ideological direction. Combining all three vignettes in a mixed effects model, with random coefficients to account for dependencies between multiple observations from the same subject and from the same vignette, the three-way interaction between SDO, Speech Content and EBC is significant (see Table 2, and Fig. 2b). The negative coefficient is about half in magnitude but with opposite sign to the interaction between SDO

Table 1
The coefficients from the regression models of SDO, Condition, and their interaction on speech protection, across the three scenarios in Study 1.

	Online Shaming			*Academic Freedom*			*Job Security*		
	β	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$
(Intercept)		3.70	[3.53, 3.86]		5.39	[5.22, 5.56]		4.96	[4.70, 5.21]
SDO	0.11	0.10	[−0.07, 0.28]	−0.22*	−0.16	[−0.29, −0.02]	−0.08	−0.09	[−0.28, 0.10]
SpeechContent	−0.02	−0.02	[−0.19, 0.14]	0.06	0.05	[−0.12, 0.23]	−0.04	−0.05	[−0.30, 0.20]
EBC	0.01	0.01	[−0.23, 0.25]	0.04	0.07	[−0.16, 0.30]	−0.04	−0.10	[−0.46, 0.25]
SDO*SpeechContent	0.21*	0.19	[0.02, 0.37]	0.19*	0.13	[0.00, 0.27]	0.27**	0.29	[0.10, 0.48]
SDO*EBC	−0.14	−0.18	[−0.42, 0.06]	−0.05	−0.05	[−0.22, 0.13]	−0.08	−0.12	[−0.38, 0.15]
SpeechContent*EBC	0.05	0.07	[−0.17, 0.31]	−0.09	−0.12	[−0.35, 0.12]	−0.03	−0.05	[−0.41, 0.30]
SDO*SpeechContent*EBC	−0.09	−0.12	[−0.36, 0.12]	−0.10	−0.09	[−0.27, 0.08]	−0.11	−0.16	[−0.43, 0.10]

Note. SDO is a participant's mean centered score on the Social Dominance Orientation scale. SpeechContent is contrast coded as 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. EBC indicates whether a participant has been exposed to both experimental conditions at the time of responding, coded as 1 when this is the case and 0 when it is not. Due to how we randomized conditions (described in Methods), this was the case for approximately half of the responses to each vignette. Confidence intervals refer to the unstandardized betas. $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

and Speech Content. This demonstrates that the motivated, selective protection of speech with which one agrees is quite substantially diminished by prior exposure to speech framed in the opposite ideological direction about a different case: If people have previously evaluated a case of restricting speech which they support, they will more strongly condemn a case of restricting speech which they oppose. Conversely, if people have previously evaluated a case of restricting speech which they oppose, they will condemn other cases of restricting speech which they support less strongly.

3. Study 2

In Study 1, participants can still plausibly deny any influence of speech content on judgments, even if they are exposed to both ideological flavors of speech across the different vignettes: The scenarios have other differences besides the speech content, which could justify differences in judgments. To directly block such avenues for plausible deniability, in Study 2 we manipulate speech *within* subjects, so that all participants respond to both versions of each vignette.

We randomized the order of presentation *between* subjects: Participants either first make judgments about four vignettes with speech they support and then the same vignettes with speech they oppose, or the other way around. Importantly, participants are unaware that the second set of vignettes is coming when answering the first set, and they are also unable to go back to change the answers to the first vignettes after having seen the second set.

If participants openly support content-sensitive speech norms, then the order of presentation should not matter: Participants should simply judge differently depending on speech content in each condition, since they are open about thinking that the extent to which an action is a blameworthy violation of speech rights depends on the content of the speech that is counteracted.

If participants are instead motivated to maintain content-neutrality in their judgments, as we predict, then they should be motivated to *answer similarly* to their replies to the first vignettes when answering their ideological opposites the second time. For instance, if a participant first judges that it is blameworthy to get someone fired for having a view she supports. She might feel compelled to then say that it is equally blameworthy to get someone fired for the exact opposite view, in order to maintain consistency. But if ordering was switched, so that the first judgment was about the firing of someone with views she opposed, then both judgments would have been more lenient.

3.1. Participants

As we expected at least as large effects in Study 2 as those in Study 1, we once again recruited 300 participants from MTurk. Since the distribution of SDO scores among our sample in Study 1 had a clear overrepresentation of low scores, Study 2 recruited participants that were 50–50 Democrats and Republicans, all Americans, by utilizing MTurks filtering function. As SDO is robustly correlated with political affiliation (Ho et al., 2012; Ho et al., 2015), this should lead to more variability in SDO scores, increasing our ability to detect interaction effects involving SDO.

We excluded participants on a per-vignette basis if they did not pass an attention check for each of the vignettes. We also excluded participants who did not clear at least 3 of the 4 attention checks from all analyses. This yielded a mean sample size of 244 participants per analysis.

Among the 262 participants not excluded from the study (M_{age} = 40.0, SD_{age} = 12.1, 141 females), there were 84% Caucasians, 5% Asians, 6% African Americans, and 4% Hispanics.

3.2. Procedure

Participants gave an informed consent and basic demographic information before filling out the short 8-item version of the SDO_7 scale (α = 0.93, M = 2.64 ± SD of 1.50). They then made moral judgments about four different free speech vignettes (in random order) where the speech being suppressed was either pro- or contra goals associated with high SDO.

The four vignettes used are shown in Box 1. They resemble those in Study 1, with the addition of a fourth scenario about a public speaking event being sabotaged ("No-platforming"). Participants's average responses across the four judgments for each vignette were the dependent variables in analyses. Responses were always scored such that higher scores indicate more support for the speaker. Cronbach's α was at >0.76 for three of the four vignettes. The exception was the *Academic Freedom* vignette, with α = 0.57. We report analyses where these items are nonetheless combined into a single scale, since the pattern of results is robust across all four single items.

After responding to the first set of vignettes, participants were given attention checks and then asked to respond to the scenarios once again, but now with alterations. They indicated the extent to which they felt the changes to the text were relevant to the associated judgments (on a 5-point likert from "not at all" (1) to "to a large extent" (5)), before they rated agreement to the judgments once again with the new scenario in mind. They had no option to go back and view or alter previous responses at this stage. The ordering of vignettes was the same as for the first set.

3.3. Results

When considering only responses to the first set of vignettes, there are significantly positive coefficients for the interactions between SDO and Condition (p's < 0.001), for all four vignettes. This replicates the

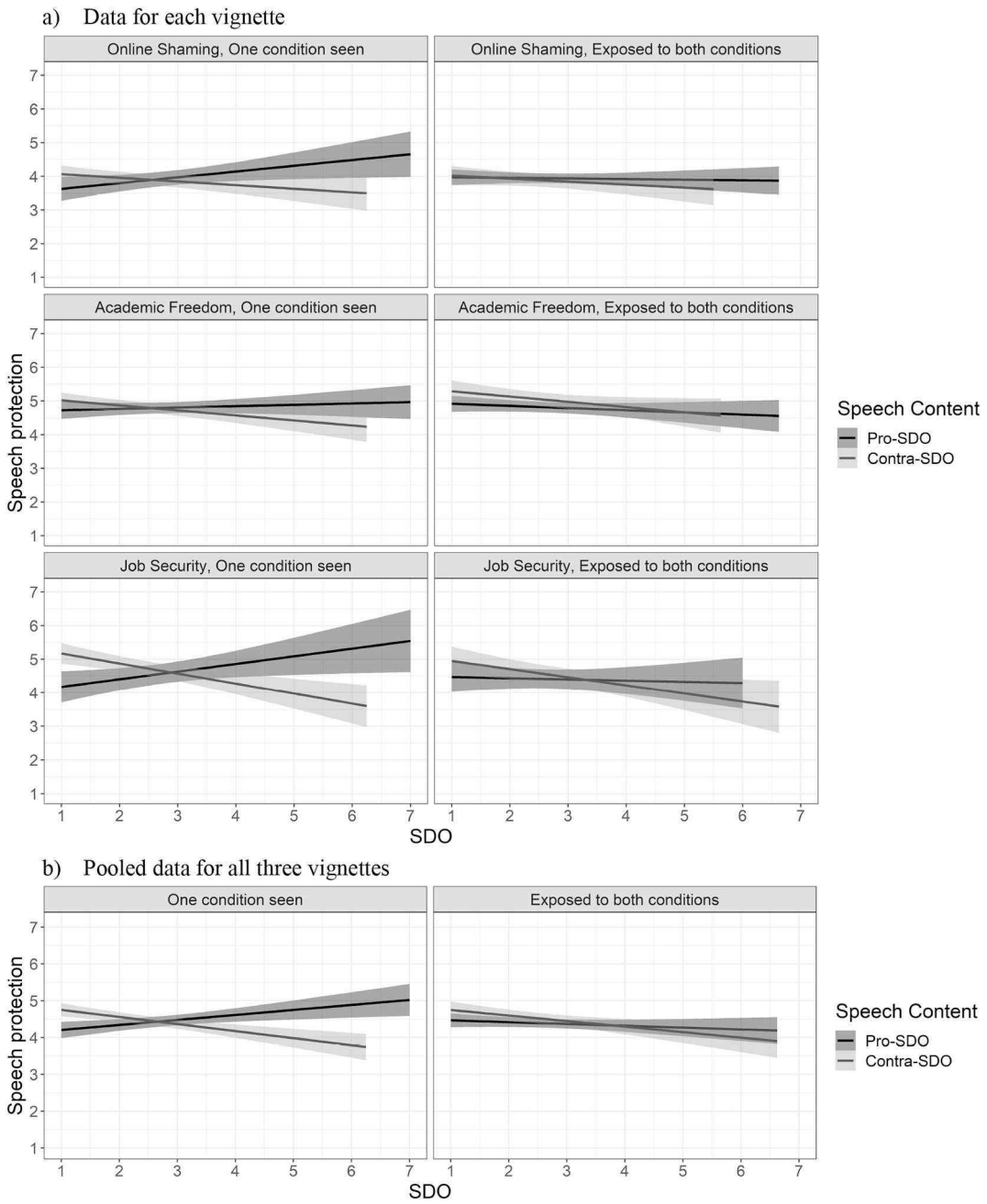


Fig. 2. Interaction plots, showing the three-way interactions between SDO, speech content and exposure to both conditions on speech protection, across each vignette and for pooled data of all three vignettes, in Study 1. SDO is Social Dominance Orientation. Speech protection is a participant's mean response to the three questions for each vignette about the blameworthiness of counteracting speech. a) shows data for each vignette separately, and b) shows the pooled data from all three vignettes. The left column contains only responses from participants who have only been exposed to one of the two speech content conditions, while the right column is from participants who have already responded to another vignette with the opposite kind of speech content to the current one. The shaded areas around the regression lines indicate 95% confidence intervals. The darkness of shading around the regression lines represents the speech condition (bright gray is contra-SDO speech, dark gray is pro-SDO speech).

prior finding that people favor increased protection for views they support (See Table 3 and the first column of Fig. 3 for results and visualizations). Low SDO predicts increased protection of the speech in the Contra-SDO condition and vice versa for high SDO.

For the *Job Security* vignette only, there is a negative main effect of Condition ($p < .01$), meaning that the posting of photoshopped images of Hillary Clinton was tolerated less than photoshopped images of Trump, when averaged across all participants. For the *No-platforming* vignette only, there is a positive main effect of SDO ($p < .001$), indicating that the average protection of speech across both conditions was higher among those with high SDO here.

The second set of responses is towards the same vignettes and statements as in the first set, except that all the judgments have been preceded by making the same judgments in the opposite ideological condition. As predicted, when running the same analysis as above on these responses (see Table 4 and the second column of Fig. 3), the interaction effects are significant (p's < 0.05) but in the *opposite direction* in all but one of the vignettes (*No-Platforming*, $p = .87$).

To investigate order effects more directly, we ran analyses with all responses included, and with an added variable called Order indicating

Table 2
Fixed effects from a mixed effects regression model of SDO, Condition, EBC, and their interactions on judgments, on pooled data for the three vignettes in Study 1. The model has random effects accounting for dependencies between multiple responses across subjects and vignettes.

	Pooled data		
	β	*b*	$CI_{95\%}$
(Intercept)		4.66	[3.66, 5.66]
SDO	−0.08	−0.08	[−0.16, 0.01]
SpeechContent	−0.00	−0.00	[−0.12, 0.11]
EBC	0.01	0.03	[−0.13, 0.20]
SDO*SpeechContent	0.21***	0.21	[0.12, 0.29]
SDO*EBC	−0.05	−0.06	[−0.19, 0.06]
SpeechContent*EBC	−0.02	−0.03	[−0.19, 0.13]
SDO*SpeechContent*EBC	−0.11**	−0.15	[−0.27, −0.03]

Note. SDO is a participant's mean centered score on the Social Dominance Orientation scale. SpeechContent is contrast coded as 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. EBC indicates whether a participant has been exposed to both experimental conditions at the time of responding, coded as 1 when this is the case and 0 when it is not. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001. Degrees of freedom for significance tests in all our mixed models are calculated using the Kenward-Roger approximation (Kenward & Roger, 1997).

whether a response is a participant's first or second response for that vignette (see Table 5).

The three-way interaction between SDO, SpeechContent and Order describes how the selective protection-effect (i.e. the interaction between SDO and SpeechContent on judgments) changes when a response is second in the order rather than first. Mirroring the findings from the analyses on the second set of responses only, there are significant three-way interactions for all the vignettes ($p < .001$, except for the *No-platforming vignette* which has $p = .016$).

Next, directly assessing participants' explicit support for content-sensitive speech norms, we considered their responses to being asked directly if the change in speech content made a difference to their judgments. The mean response across scenarios was 1.94 on a scale from 1 to 5 (SD = 1.09), meaning that participants largely reported that the changes in speech content were *not* morally relevant to them, indicating explicit support for content-neutral speech norms. Consistent with how the order effect was larger for the *No-platforming* scenario, the mean response here was significantly higher than for the three others, at 2.14 ($p < .01$).

In summary, when looking only at responses to the second, reversed set of scenarios, the effects of SDO on free speech judgments appear to be opposite of what one would expect if this was the first set. There is now increased protection of the speech that participants *oppose*, and decreased protection of the speech they *support,* demonstrating that many participants comply more with content-neutral speech norms when their practice of content-sensitive norms would otherwise be obvious. However, the ideologically consistent effects of SDO on judgments in the first set of scenarios again demonstrates that this is not the case if selective judgments may go undetected, even though participants largely deny that speech content matters for their endorsements of speech rights.

4. Study 3

Study 3 was a pre-registered replication of Study 2. Additionally, we explored if the results of Study 2 were robust to the following set of potential moderators related to honesty and self-awareness: Introspectiveness (Hansell & Mechanic, 1985), the Honesty/Humility dimension of the HEXACO-PI (Ashton & Lee, 2009), and the number of heads reported out of five coin-flips when there is a bonus-payment for each head. For specific details on how we predicted these variables to moderate order effects, and for analyses investigating these predictions, see the Online Supplement.

4.1. Participants

To gain more accurate estimates of the effects in Study 2, we now increased the sample size from 300 to 400. Once again, an MTurk sample was specified to be 50–50 American democrats and republicans. Exclusion-criteria also mirrored those in Study 2. This yielded a mean sample size of 344 participants per analysis.

Among the 351 participants not excluded from the study (M_{age} = 40.1, SD_{age} = 11.6, 148 females), there were 81% Caucasians, 4% Asians, 6% African Americans, 6% Hispanics, and 3% other.

4.2. Procedure

The procedure from Study 2 was repeated, except that now the Introspectiveness- and the Honesty/Humility scales were filled out following the SDO_7 scale. Additionally, after the subjects had responded to all the vignettes, they were asked to flip a coin five times and report the number of heads. They were informed that they would receive 10 cents for each head they reported.

4.3. Results

For all four vignettes, the two-way interactions between SDO and SpeechContent and also the three-way interactions between SDO, SpeechContent, and Order were all significant at $p < .001$, in the predicted directions (see Table 6 and Fig. 4). The effect sizes were mostly similar to those in Study 2. Neither introspection, trait-honesty, nor cheating for an economic bonus were found to moderate these effects. This suggests either that the processes producing these biases operate too far outside consciousness to be accessible through introspection, or that participants are uninterested in overriding them when they gain access to them (regardless of their levels of honesty/integrity). For further analyses and discussion of Study 3, see the Online Supplement.

Table 3
First set of responses: The effects of SDO, Condition, and their interaction on speech protection for the first sets of responses to each of the four vignettes in Study 2.

	Online Shaming			*Academic Freedom*			*Job Security*			*No-Platforming*		
	β	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$
(Intercept)		4.06	[3.92, 4.20]		4.72	[4.61, 4.83]		4.68	[4.50, 4.86]		4.03	[3.88, 4.19]
SDO	0.02	0.02	[−0.07, 0.11]	−0.03	−0.02	[−0.09, 0.06]	−0.08	−0.09	[−0.21, 0.03]	0.21***	0.21	[0.10, 0.31]
SpeechContent	0.09	0.11	[−0.03, 0.24]	0.02	0.02	[−0.09, 0.13]	−0.15**	−0.25	[−0.43, −0.07]	−0.11	−0.15	[−0.31, 0.00]
SDO*SpeechContent	0.30***	0.25	[0.16, 0.34]	0.31***	0.22	[0.14, 0.29]	0.27***	0.29	[0.17, 0.41]	0.24***	0.23	[0.13, 0.34]

Note. SDO is a participant's score on the Social Dominance Orientation scale, mean centered. SpeechContent is coded 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. Confidence intervals refer to the unstandardized betas. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

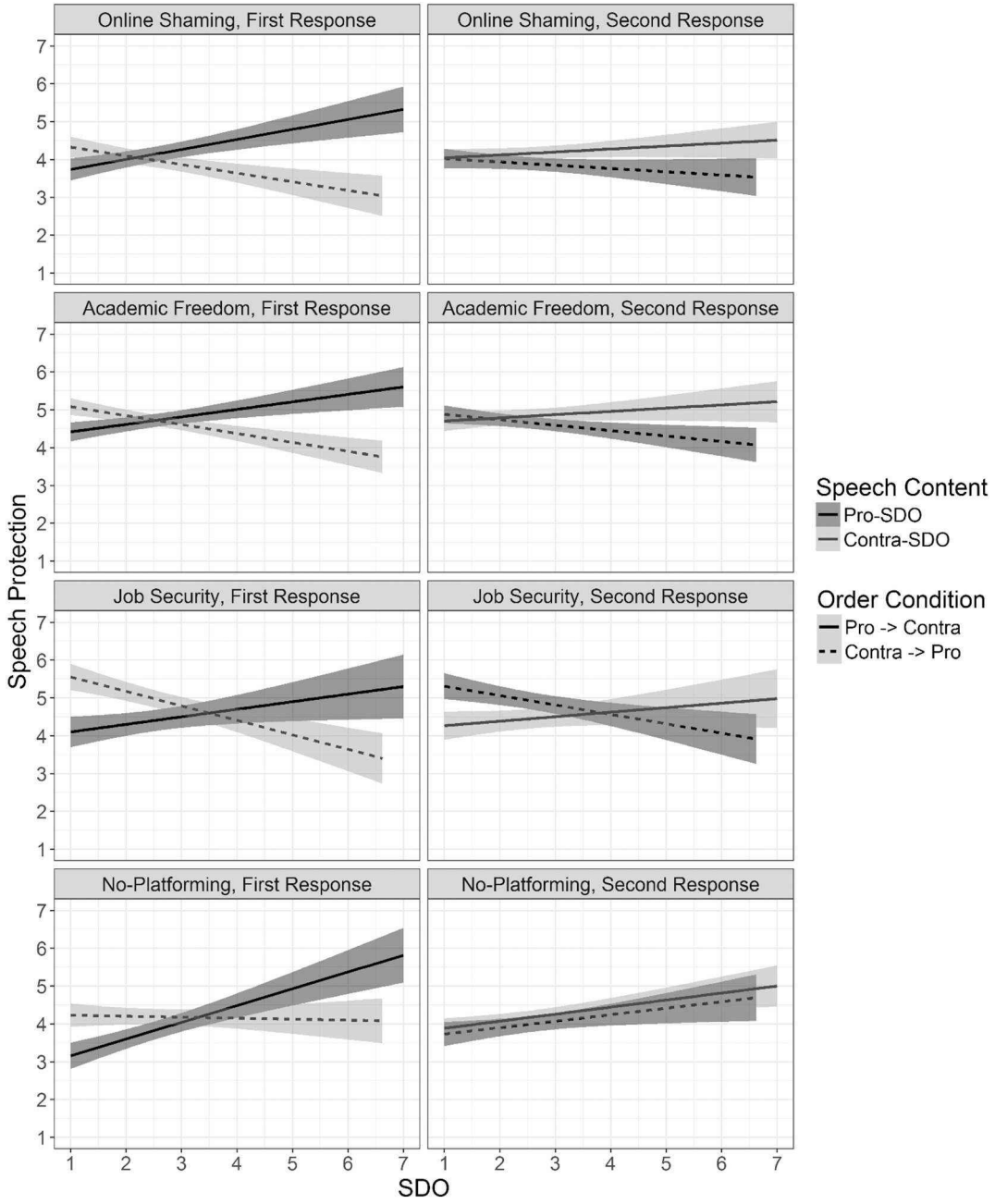


Fig. 3. Interaction plots, showing the three-way interactions between SDO, speech content and presentation order on speech protection, across all four vignettes in Study 2. SDO is Social Dominance Orientation. Speech protection is a participant's mean response to the four questions for each vignette about the blameworthiness of counteracting speech. There is one row for each vignette. The left column contains only responses from the first showing of a vignette, the right column contains only responses from the second showing. The darkness of shading around the regression lines represents the speech condition (bright gray is contra-SDO speech, dark gray is pro-SDO speech). The shaded areas around the regression lines indicate 95% confidence intervals. To accentuate which lines in the first- and second columns "consist of the same participants", the shapes of the regression lines indicate the between subject order manipulation: dotted lines are from the participants who got the Contra-SDO condition first, solid lines are those who got the Pro-SDO condition first.

5. Study 4

Up to this point, preferences for speech norms have only been assessed in relation to specific scenarios, rather than in general terms. Evaluating abstract moral principles can be challenging; Having them applied to specific scenarios makes things more concrete and understandable, and it might give participants better access to their attitudes. Still, we feel it is worth investigating if the expressed support for content-neutral speech norms in specific scenarios carries over to support for the more general principle that the content of speech should not affect judgments of its suppression.

Furthermore, we investigate the prediction that participants are particularly supportive of content-neutral speech norms in the contexts where they are most useful. That is, when they are proposed by the ideological opposition, and contrasted with the kind of content-sensitivity that selectively suppresses the views you support. After all, the benefits to oneself from having others give equal protections to all speech are largest when these others would have otherwise gone after views that you support. To explore this, we ask participants to take sides in a hypothetical discussion about free speech between two friends, one pro- and one against content-sensitive speech norms (see Box 2 for the full text). We experimentally manipulate the ideological commitments

Table 4
Second set of responses: The effects of SDO, Condition, and their interaction on speech protection for the second sets of responses to each of the four vignettes in Study 2.

	Online Shaming			*Academic Freedom*			*Job Security*			*No-Platforming*		
	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$
(Intercept)		4.03	[3.91, 4.14]		4.74	[4.62, 4.86]		4.68	[4.51, 4.85]		4.10	[3.96, 4.24]
SDO	−0.01	−0.00	[−0.08, 0.07]	−0.04	−0.03	[−0.11, 0.05]	−0.06	−0.06	[−0.18, 0.05]	0.21***	0.18	[0.09, 0.27]
SpeechContent	−0.14*	−0.14	[−0.26, −0.03]	−0.09	−0.10	[−0.02, 0.22]	0.15*	0.22	[0.05, 0.39]	−0.07	−0.09	[−0.23, 0.05]
SDO*SpeechContent	−0.12*	−0.08	[−0.16, −0.00]	−0.16**	−0.11	[−0.19, −0.04]	−0.18**	−0.18	[−0.30, −0.07]	−0.00	−0.01	[−0.10, 0.09]

Note. SDO is a participant's score on the Social Dominance Orientation scale, mean centered. SpeechContent is coded 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001.

Table 5
Fixed effects from a mixed effects regression model predicting speech protection from SDO, SpeechContent, Order, and their interactions across the four vignettes in Study 2. The model also includes random intercepts for each subject.

	Online Shaming			*Acaddemic Freedom*			*Job Security*			*No-Platforming*		
	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$	β	*b*	$CI_{95\%}$
(Intercept)		4.07	[3.92, 4.23]		4.72	[4.60, 4.84]		4.70	[4.52, 4.88]		4.07	[3.91, 4.23]
SDO	0.01	0.01	[−0.09, 0.11]	0.00	0.00	[−0.08, 0.08]	−0.08	−0.09	[−0.21, 0.03]	0.28***	0.25	[0.15, 0.36]
SpeechContent	0.09	0.11	[−0.05, 0.26]	0.01	0.01	[−0.10, 0.13]	−0.16**	−0.26	[−0.44, −0.08]	−0.12*	−0.17	[−0.33, −0.01]
Order	−0.02	−0.04	[−0.17, 0.09]	0.02	0.05	[−0.05, 0.14]	−0.00	−0.00	[−0.11, 0.10]	0.03	0.09	[−0.05, 0.24]
SDO*SpeechContent	0.34***	0.26	[0.16, 0.36]	0.37***	0.25	[0.17, 0.33]	0.27***	0.29	[0.17, 0.41]	0.26***	0.23	[0.13, 0.34]
SDO*Order	−0.02	−0.02	[−0.10, 0.07]	−0.01	−0.01	[−0.07, 0.05]	0.02	0.02	[−0.04, 0.09]	−0.02	−0.02	[−0.12, 0.08]
SpeechContent*Order	−0.14	−0.23	[−0.51, 0.05]	−0.06	−0.10	[−0.31, 0.12]	0.22**	0.49	[0.15, 0.84]	0.03	0.06	[−0.22, 0.35]
SDO*SpeechContent*Order	−0.31***	−0.34	[−0.52, −0.16]	−0.39***	−0.37	[−0.51, −0.23]	−0.31***	−0.47	[−0.70, −0.24]	−0.18*	−0.23	[−0.42, −0.04]

Note. SDO is a participant's score on the Social Dominance Orientation scale, mean centered. SpeechContent is coded 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. Order is coded 0 for the first time responding, 1 for the second. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001.

of *both* discussants simultaneously, so that they are either both liberals or both conservatives. We predict that the level of support for content-neutral speech norms will be higher, and that support for content-sensitive speech norms will be lower, when they are proposed by someone who does not share the participant's own ideological commitments.

To address whether participants are open about supporting content-neutral norms more when they are voiced by the opposition, we again implement an order manipulation; Participants respond to both conditions of the vignette, in randomized order.

5.1. Participants

We recruited from Amazon's Mechanical Turk a sample of 400 people, split 50–50 between Republicans and Democrats living in the US, using MTurk's filtering functionality.

Out of the 406 people who completed our survey, 33 failed the attention check, and one did not respond to our predictor variable, leaving a sample of 372 participants for our analyses (M_{age} = 39.1, SD_{age} = 12.2, 213 females). This sample consisted of 82% Caucasians, 9% African Americans, 4% Asians, and 3% Hispanics.

5.2. Procedure

Participants gave an informed consent and basic demographic information, and they indicated their political position on economic- and social issues using two visual analog scales ranging from "very liberal" to "very conservative". The scores on these two scales (r = 0.81) were combined to form an overall index of political orientation functioning as our predictor variable.

Participants then read and responded to the vignette. The names of the two discussants in the vignette, Riley and Casey, were chosen to be as uninformative as possible with regards to both gender and political affiliation: People named Riley or Casey in the US are about equally likely to be male or female (Flowers, 2015), and they are also about equally likely to have voted democratic or republican in the 2016 election (Clarity Campaign Labs, 2018).

The responses to the questions were combined to create a single scale indicating support for content-neutral speech norms (α = 0.91). Higher scores on this scale then imply agreement with Casey, who advocated content-neutral norms, and disagreement with Riley, who took the opposing position.

After responding to the vignette the first time, participants were asked to read the vignette once again, this time in the other experimental condition, and they once again rated their agreement to the same six statements about the vignette. They were also asked to explicitly

Table 6
Fixed effects from a mixed effect regression models predicting speech protection from SDO, SpeechContent, Order, and their interactions across the four vignettes in Study 3. The models also include random intercepts for each subject.

	Online Shaming			*Academic Freedom*			*Job Security*			*No-Platforming*		
	β	b	$CI_{95\%}$	β	b	$CI_{95\%}$	β	b	$CI_{95\%}$	β	b	$CI_{95\%}$
(Intercept)		4.07	[3.93, 4.20]		3.52	[3.40, 3.63]		4.32	[4.17, 4.47]		3.65	[3.52, 4.26]
SDO	−0.04	−0.03	[−0.12, 0.06]	0.04	0.03	[−0.05, −0.02]	−0.01	−0.01	[−0.11, 0.10]	0.18***	0.16	[0.07, 0.26]
SpeechContent	0.01	0.01	[−0.12, 0.15]	0.06	0.07	[−0.19, 0.23]	−0.15**	−0.22	[−0.37, −0.07]	−0.17**	−0.22	[−0.37, −0.09]
Order	0.03	0.06	[−0.06, 0.18]	0.01	−0.02	[−0.10, 0.30]	−0.00	−0.00	[−0.08, 0.08]	0.06*	0.16	[0.03, 0.28]
SDO*SpeechContent	0.31***	0.24	[0.16, 0.33]	0.34***	0.27	[0.18, 0.35]	0.27***	0.28	[0.18, 0.39]	0.35***	0.32	[0.22, 0.41]
SDO*Order	0.02	0.02	[−0.06, 0.10]	−0.01	−0.01	[−0.07, 0.04]	−0.02	−0.03	[−0.08, 0.03]	−0.04	−0.05	[−0.13, 0.04]
SpeechContent*Order	−0.08	−0.14	[−0.38, 0.10]	−0.08	0.13	[−0.10, 0.35]	0.17*	0.36	[0.07, 0.65]	0.07	0.13	[−0.11, 0.37]
SDO*SpeechContent*Order	−0.26***	−0.29	[−0.45, −0.13]	−0.34***	−0.38	[−0.53, −0.22]	−0.32***	−0.48	[−0.68, −0.28]	−0.26***	−0.33	[−0.50, −0.17]

Note. SDO is a participant's score on the Social Dominance Orientation scale, mean centered. SpeechContent is coded 1 for the Pro-SDO condition and −1 for the Contra-SDO condition. Order is coded 0 for the first time responding, 1 for the second. Confidence intervals refer to the unstandardized betas. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

indicate the extent to which the changes to the text would impact their responses.

5.3. Results

Consistent with studies 1 and 2, our sample sided quite clearly with the discussant arguing for content-neutral speech norms, with a mean agreement here on the first reading of the vignette at 5.1 on a 7-point scale. This was significantly higher than the midpoint of 4.0 ($t(371) = 14.8$, $p < .001$).

In our model predicting support for content-neutral speech norms (see Table 7 and Fig. 5), the coefficient for the interaction was significantly negative ($p < .001$), suggesting that left-leaning participants were relatively more inclined to support such norms when both discussants are conservatives, and that right-leaning participants had the opposite tendency. The positive main effect of political orientation ($p < .001$) suggests that, averaging across the conditions, conservative participants were more strongly supportive of content-neutral norms than liberals were.

In the models investigating order effects (Table 8) we find, as in study 2, that the coefficient for the three-way interaction is of opposite sign and of larger magnitude than the coefficient for the interaction between political orientation and Condition ($p < .001$), indicating that order of presentation has a substantial effect on responses here as well.

In summary, while participants generally and clearly favor content-neutral norms, their levels of support nevertheless systematically depend on the context. Specifically, they support content-neutral speech norms more when they are proposed by someone from their political opposition, and contrasted with content-sensitive norms favoring the opposition. This conforms to how content-neutral norms are particularly useful when the alternative is suppression of one's own favored views.

We once again find a significant order effect, meaning that the selective endorsement of content-neutral speech norms is significantly attenuated when this selectivity would be laid bare by answers to the second, ideologically-reversed scenario. This again supports the idea that people are motivated to feel morally consistent.

6. Study 5

In Study 4, we identified two contextual factors that influenced our participants' level of support for content-neutral speech norms. Specifically, support for content-neutrality increasd when (1) the person proposing them was from one's political opposition, and (2) the content-sensitive norms that content-neutrality was contrasted with were slanted in favor of the opposition. However, these two factors were systematically confounded, so we could not determine their relative importance.

To disambiguate between these two factors identified in Study 4, in Study 5 we use a vignette with three discussants rather than two. One discussant is *always* a liberal, and argues for liberal content-sensitivity (i.e. that it is more blameworthy to suppress the speech of liberals than conservatives), and another discussant is *always* conservative, and argues for conservative content-sensitivity. The third discussant always advocates content-neutrality, but we experimentally manipulate whether they are a conservative, a liberal, or if they are politically neutral.

With this design, content-neutrality is always contrasted with both the liberal and conservative kinds of content-sensitivity. If the preference for content-neutrality is higher when proposed by one's opposition, then this effect should show up in this design as well. If this preference is rather an effect of what content-neutrality is contrasted with, then the political affiliation of the person proposing content-neutrality should have no effect here, since both kinds of content-sensitivity are always present as contrasts.

6.1. Participants

We recruited from Amazon's Mechanical Turk a sample of 400 people, split 50–50 between Republicans and Democrats in the US, using MTurk's filtering functionality.

Out of the 400 people who completed our survey, 47 failed our attention check, leaving a sample of 353 participants for our analyses ($M_{age} = 40.4$, $SD_{age} = 12.3$, 154 females). This sample consisted of 77% Caucasians, 9% Asians, 6% African Americans, 5% Hispanics, and 3% Native Americans.

6.2. Procedure

Participants gave an informed consent and basic demographic information, and they indicated their political position on economic- and

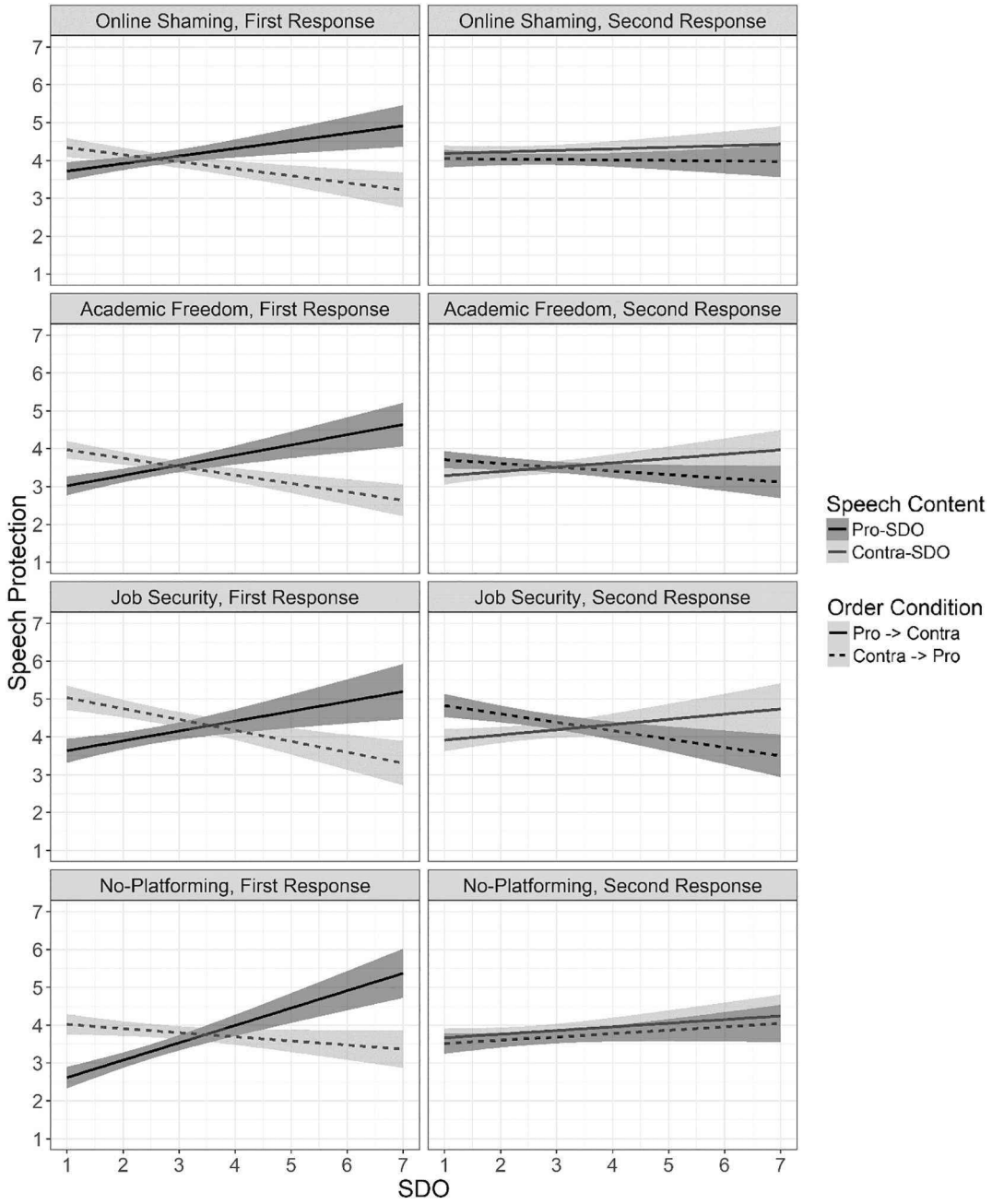


Fig. 4. Interaction plots, showing the three-way interactions between SDO, speech content and presentation order on speech protection, across all four vignettes in Study 3. SDO is Social Dominance Orientation. Speech protection is a participant's mean response to the four questions for each vignette about the blameworthiness of counteracting speech. There is one row for each vignette. The left column contains only responses from the first showing of a vignette, the right column contains only responses from the second showing. The darkness of shading around the regression lines represents the speech condition (bright gray is contra-SDO speech, dark gray is pro-SDO speech). The shaded areas around the regression lines indicate 95% confidence intervals. To accentuate which lines in the first- and second columns "consist of the same participants", the shapes of the regression lines indicate the between subject order manipulation: dotted lines are from the participants who got the Contra-SDO condition first, solid lines are those who got the Pro-SDO condition first.

social issues using two visual analog scales ranging from "very liberal" to "very conservative". The scores on these two scales ($r = 0.83$) were combined to form an overall index of political orientation functioning as our predictor variable.

Participants then read and responded to the vignette. As with the first two discussants, who are again named Riley and Casey, the third discussant, Kim, also has a name that is uninformative with regards to both gender and political affiliation. The name Kim is used for both males and females in the US, and people with this name are about equally likely to have voted democratic or republican in the 2016 election (Clarity Campaign Labs, 2018).

6.3. Results

As in Study 4, participants once again largely preferred content-neutral over content-sensitive speech norms. The mean level of agreement with the discussant supporting content-neutrality in the vignette was at 4.96 on a 7-point scale, which is significantly higher than the mid-point of 4 ($t(352) = 11.3$, $p < .001$). Support for content neutrality was also significantly higher than the mean support for both conservative

Table 7
The effects of Political Orientation, Condition, and their interaction on levels of expressed support for content-neutral speech norms in Study 4.

	Support for content-neutral norms		
	β	*b*	$CI_{95\%}$
(Intercept)		5.18	[5.05, 5.32]
Pol_Orient	0.22***	0.09	[0.05, 0.13]
Condition	0.03	0.05	[−0.09, 0.19]
Pol_Orient*Condition	−0.21***	−0.09	[−0.13, −0.05]

Note. Pol_Orient is political affiliation from −5 ("very liberal") to 5 ("very conservative"). Condition is coded 1 for the Conservative condition, where both discussants are conservatives, and −1 for the Liberal condition, where they are both liberals. Confidence intervals refer to the unstandardized betas. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

content sensitivity (M = 2.21, t(352) = 26.0, p < .001) and liberal content sensitivity (M = 2.56, t(352) = 23.0, p < .001). As one might expect, support for both of the two kinds of content sensitivity depended significantly on political orientation, such that conservatives were more inclined to support conservative content sensitivity and less inclined to support liberal content sensitivity, and vice versa for liberals (all p < .001).

Contrary to the idea that support for content-neutrality is stronger when it is proposed by one's political opposition, there were no significant effects of our experimental manipulation on responses. Participants were equally inclined to support content-neutral norms whether they were proposed by a political ally, opponent, or neutral, as indicated by the lack of significant interaction effects between political orientation and the experimental manipulations in our main analysis of interest (detailed in Table 9). This then suggests that participants prefer content-neutral speech norms more strongly when they are contrasted with content-sensitive norms favoring one's opposition. Thus, when the threat to silence both conservative and liberal speech is held constant (so that both liberal and conservative participants will always face the threat that their speech might be silenced) the affiliation of the person proposing content neutral speech norms no longer affects how much people agree with them.

Table 8
Fixed effects from a mixed effects regression models predicting expressed support for content-sensitive and –neutral speech speech norms from political orientation, Condition, Order, and their interactions, in Study 4. The model also includes random intercepts for each participant.

	Support for content-neutral norms		
	β	*b*	$CI_{95\%}$
(Intercept)		5.18	[5.05, 5.32]
Pol_Orient	0.23***	0.09	[0.05, 0.13]
Condition	0.03	0.05	[−0.09, 0.19]
Order	0.02	0.06	[−0.01, 0.12]
Pol_Orient*Condition	−0.22***	−0.09	[−0.13, −0.05]
Pol_Orient *Order	−0.04*	−0.02	[−0.04, −0.00]
Condition*Order	−0.05	−0.10	[−0.37, 0.16]
Pol_Orient*Condition*Order	0.23***	0.13	[0.05, 0.21]

Note. Pol_Orient is political orientation measured from −5 ("very liberal") to 5 ("very conservative"). Condition is coded 1 for the Conservative condition, where both discussants are conservatives, and −1 for the Liberal condition, where both are liberals. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001. Degrees of freedom for significance tests are calculated using the Kenward-Roger approximation.

7. Study 6

While we have now shown that many people seem to have a need to appear quite consistent in using free speech principles, we have not established whether this need to be consistent holds for all kinds of debates. All the views expressed in our hypothetical scenarios thus far are currently quite well-supported in the US, where our data was collected. This was deliberate, as we needed substantial amounts of participants to be both for and against all the different views presented for our analyses to be informative. A drawback to this, however, is that all of the views in our scenarios can be argued to be inside the so-called "Overton Window", containing all the views within the realm of acceptable discourse (Lehman, 2012). Therefore, we cannot generalize our findings outside of this window, to fringe views widely seen as truly toxic and destructive. It might well be the case that many people only strive to remain consistent about free speech for viewpoints that are at least somewhat within the norm, and that they openly support giving reduced protection for speech that they see as beyond the pale. Here, we explore whether there is such an Overton Window of acceptable discourse, marking the boundaries of the effects we have found thus far.

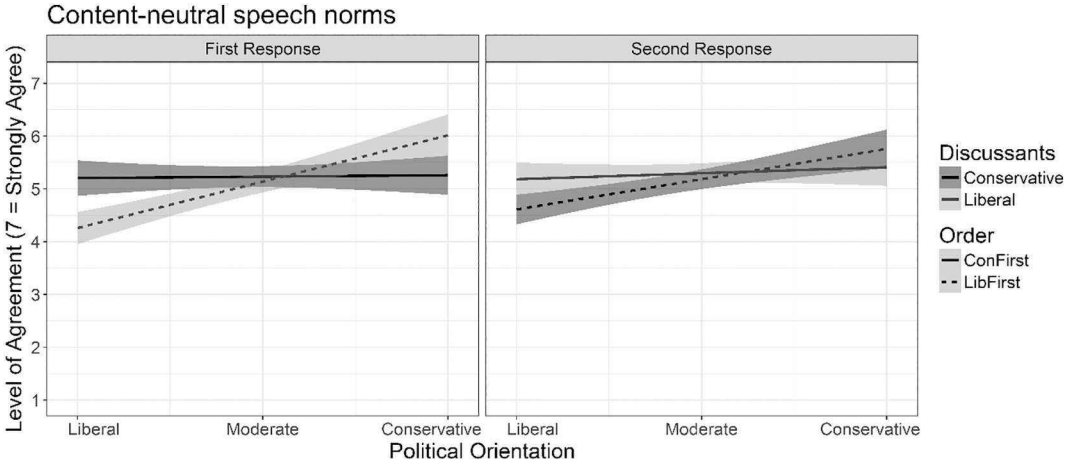


Fig. 5. Interaction plots showing the three-way interactions between Political Orientation, Condition and Order condition on Level of Agreement, across all four vignettes in Study 4. Political Orientation is a participant's position from very liberal (−5) to very conservative (5). Level of Agreement is a participant's mean response to the six statements indexing support for content-neutral speech norms. The left column contains only responses from the first showing of a vignette, the right column contains only responses from the second showing. The shade around regression lines represents the condition (bright gray is when discussants are liberals (Liberal Condition), dark gray is when they are conservatives (Conservative Condition)). To accentuate which lines in the first- and second columns consist of the same participants, the shapes of the regression lines indicate the between subject order manipulation: dotted lines are from the participants who got the liberal condition first ("LibFirst"), solid lines are those who got the Conservative Condition first ("ConFirst").

Table 9
The effects of Political Orientation, Condition, and their interactions on levels of expressed support for content-neutral speech norms in Study 5.

	Support for content-neutral norms		
	β	*b*	$CI_{95\%}$
(Intercept)		4.89	[4.61, 5.17]
Pol_Orient	0.11	0.06	[−0.04,0.15]
Condition_C	−0.05	−0.17	[−0.56,0.23]
Condition_L	−0.02	−0.06	[−0.46,0.33]
Pol_Orient*Condition_C	−0.08	−0.07	[−0.21,0.06]
Pol_Orient*Condition_L	−0.02	−0.01	[−0.14,0.11]

Note. Pol_Orient is political affiliation from −5 ("very liberal") to 5 ("very conservative"). Condition_C and Condition_L are dummy variables indicating the experimental condition. Condition_C is coded as 1 for the condition where content-neutral norms are proposed by a conservative and 0 otherwise, while Condition_L is coded as 1 for the condition where content-neutral norms are proposed by a liberal and 0 otherwise. The third condition, where content-neutral norms are proposed by a neutral, is then represented when both these variables are at 0. Confidence intervals refer to the unstandardized betas. $*p \leq .05$, $**p \leq .01$, $***p \leq .001$.

To do this, we again use our order effect design, where participants first respond to one side of each debate, and then make a new response for the other side. In addition to looking at debates where the US population is split at around 50–50, we also included cases where one side has much more support than the other, and where the less-supported view may be seen as morally reprehensible by the majority of the population.

We included views that represent debates where public opinion is split quite evenly between the two sides, and where support for one or the other side can be predicted from SDO and political orientation: abortion, gun control, the superiority of Western culture, and that Donald Trump is a good president. Both sides of each of these issues have at least 35% support in the US at the time of writing (Fahmy, 2020; Gallup, 2020a; Pew Research, 2011; and Gallup, 2020, respectively).

We also included claims which all have less than 10% of the US population willing to say that they support them (Imhoff & Jahnke, 2018; Inglehart et al., 2014; Krysan & Moberg, 2016; Newport, 2012; Oliphant, 2017; Rhode, 2016). We included two claims that, although uncommon, still reflect ideological points of view on which people may reasonably differ, namely moral veganism and that the US should have completely open borders for anyone in the world. We also included 4 claims which were purposefully chosen to be inflammatory, in that their spread could result in serious harm to other human beings who are in a weak position to defend themselves: support for pedophilia, stoning of women, rape, and old school racism that Blacks lack inborn ability to learn.

We hypothesized that inconsistent speech support would be stronger for cases where people have a strong opinion on the issue, and where this opinion is also very popular, so that they would pay less of a cost from alienating the people whose opinions they would restrict. If any potential Overton window effects depend not only on the relative level of support for a view, but also on its moral characteristics, then equal protection for the the latter four claims (racism, rape, stoning and pedophilia) should also be lower than for the first two (open borders and veganism).

Another question that is still left open is whether our finding of a need for consistency is driven by an internal motivation to feel consistent, as we have hypothesized. It is also possible that participants, even though they know they are anonymous, feel like their responses have an audience, and that they are mostly driven by an external motivation to appear consistent before this audience. To investigate the prediction from this second hypothesis that the need to be consistent will increase if the feeling of being observed increases, we here experimentally manipulate this feeling.

7.1. Participants

We recruited a sample of 400 participants from Amazon's Mechanical Turk, which again was split 50–50 between American republicans and democrats, using MTurk's filtering functionality.

Out of the 401 people who completed our survey, 61 failed our attention check, leaving a sample of 340 participants for our analyses ($M_{age} = 40.4$, $SD_{age} = 12.3$, 152 females). This sample consisted of 77% Caucasians, 9% Asians, 6% African Americans, 5% Hispanics, and 3% Native Americans.

7.2. Procedure

Participants gave an informed consent and basic demographic information, and they responded to the SDO_7 scale ($\alpha = 0.89$, $M = 2.78 \pm$ SD of 1.39). To manipulate participants' feeling of being observed, we then randomized whether we told them that "when responding to the rest of the questions, please be aware that there is a possibility that we would like to contact you later on with follow-up questions about your answers", or that "when responding to the rest of the questions, please be aware that this survey is completely anonymous". We also included a manipulation check towards the end of the survey, asking if participants felt anonymous/observed, on a 7-point scale.

In the main part of the survey, participants were presented with a series of views. They were asked to "Suppose that each view is put forward by a person with a loyal following of over 50.000 people on social media."

For each of the views we presented, we asked participants what they thought were the most severe kinds of measures that were morally permissible for opponents of the view to use in response. The scale ranged from 0 to 10, where 0 meant no measures were permissible, low numbers meant only mild measures (reasoned arguments, etc.), and higher numbers meant increasingly stronger measures, from shaming and ridicule all the way up to imprisonment and violence.

We made a total of ten pairs of views, where the views in each pair were negations of each other. For views that represent debates where public opinion is split quite evenly between two sides, and where support for one or the other side can be predicted from SDO, we included: *Abortion should be [legal/illegal] in most cases*; *Gun control laws in the US [should/should not] become more strict*; *Western culture [is/is not] superior to most other cultures*; *Donald Trump [is/is not] a good president.*

For political/ideological views which have less than 10% of the US population willing to say that they support them, we included open borders and meat consumption (*The United States should have a policy of completely open borders, so that anyone in the world who wants it can get citizenship* and *Consuming meat and other products from animals is evil*). For claims that imply harm to people who are in a weak position defend themselves we included racism, rape, stoning and pedophilia (*A large part of the reason why there is racial inequality in the US is that blacks have less inborn ability to learn; Rape should not be considered a serious crime; Stoning women for adultery is a good practice that should be used in the US; Sexual relationships between adults and young children can sometimes be a good thing*). The claims were all paired up with their exact negations (e.g. "The US **should not** have completely open borders", "Rape **should** be considered a serious crime", and so on).

We once again used the order effect design from earlier, so that we first presented one view from each of the ten pairs, and then presented the second set of ten views after that. For the debates that the US population is divided about 50–50 on, we randomized whether we showed the high-SDO or the low-SDO side first. Independently of this, for the debates with >90% support for one of the sides, we randomized whether participants judged all the popular or all the unpopular views first.

At the end of the survey, we asked participants to indicate their own position on each of the ten issues, on a 6-point scale with no midpoint.

7.3. Results

7.3.1. Descriptives

The mean response to our 11-point scale for the appropriate severity of measures to combat speech was at 3.39 (SD = 3.27) across all the twenty different viewpoints in the study, indicating that participants were generally opposed to harsh measures like censorship and violence for most views.

For the issues where both sides have substantial support, many of our participants once again showed a tendency to give similar levels of protection for speech representing either side. The mean absolute differences on the scale between the two sides here were at 1.60 (SD = 2.17), 1.84 (SD = 2.48), 2.60 (SD = 2.70), 2.15 (SD = 2.27), for the issues of gun control, Donald Trump, abortion, and western cultural superiority, respectively (all significantly different from zero, at $p < .001$). The percentages of participants giving the same level of protection for speech for both sides for each of these issues were at 64% for gun control, 58% for Trump, 51% for western culture, and at 45% for abortion.

Among the six issues where >90% of the US population are on the same side, the views about open borders and meat consumption were treated differently by participants than the highly inflammatory views. For open borders, most participants gave similar levels of protection to both sides of the issue with the mean absolute differences in responses to the two sides being 1.65 (SD = 2.34) for open borders, with 49% of participants giving equal scores. The same was the case for meat consumption with a mean difference of 1.45 (SD = 2.06) and 65% equal scores.

Things were quite different for the four issues purposefully chosen to be provocative. The mean absolute differences in support for restricting the two sides to each issue were here at 5.54 (SD = 3.40) for rape, 5.57 (SD = 3.49) for stoning, 6.37 (SD = 3.42) for pedophilia, and 3.51 (SD = 3.05) for racism. All these four difference scores are significantly larger than those for the other six, less inflammatory, issues (all $p < .001$). The percentages giving equal protection to both sides were at 13% for both rape and stoning, 9% for pedophilia, and 21% for racial inequality.

7.3.2. Predicting inconsistency in speech protection

Next, we investigated which features of an issue determine whether or not someone will seek to apply consistent standards for speech from both sides of it. Our measure of inconsistency is the difference in speech restrictions supported by a participant to each side of an issue.

Recall our hypothesis that inconsistent speech support would be stronger for cases where the participant has a strong opinion on the issue, and where this opinion is also very popular, so that you pay less of a cost from alienating the people whose opinions you restrict. This then translates into an interaction between our measure of opinions and a variable indicating whether the issue in question is one where a large majority is on the same side.

We also use this model to investigate whether the need to be consistent increases when participants feel as if they are responding before an audience, as influenced by our publicness manipulation.

The results from this analysis are in Table 10. As predicted, we find significant main effects of both opinion strength and the Consensus variable, suggesting that inconsistency becomes more likely when participants have strong opinions and when the issue is one where there is large consensus in the population favoring one of the sides. And we also find a significant interaction between opinions and Consensus, so that opinion strength has even stronger effects on the level of inconsistency for issues where there is also consensus.

With respect to the effects of making speech judgments before an audience, our anonymity/publicness manipulation significantly affected whether participants felt anonymous or that they were being watched, on our 7-point manipulation check (M(public) = 3.51, M(anonymous) = 3.13, $t(338) = 2.90$, $p < .01$). However, we found no evidence that this manipulated feeling of anonymity/publicness has any effect on inconsistent speech support across the ten issues, whether as main effects or as interactions. This was the case no matter if we used experimental condition (anonymous/public) or the continuous measured feeling of being anonymous or watched (i.e. the manipulation check) as predictor variables.

Table 10
Fixed effects from a mixed effect regression model predicting inconsistency in speech protection, as measured by the difference in the support for harsh measures against speech supporting two sides of the same issue, in Study 6. In addition to the predictors in the table, the model also includes random intercepts for each subject and for each issue.

	Inconsistency		
	β	*b*	$CI_{95\%}$
(Intercept)		0.30	[−0.49, 1.10]
Opinion	0.21***	0.54	[0.43, 0.65]
Consensus	0.24**	1.11	[0.34, 1.89]
Public	0.02	0.09	[−0.19, 0.37]
Opinion*Consensus	0.12***	0.33	[0.21, 0.44]
Opinion*Public	0.01	0.01	[−0.09, 0.12]
Consensus*Public	−0.02	−0.09	[−0.30, 0.11]
Opinion*Consensus*Public	0.01	0.02	[−0.09, 0.12]

Note. Consensus is contrast coded as 1 for the issues where the US population is largely in agreement, and −1 for the issues where there are >35% on both sides. The dependent variable, Inconsistency, is coded such that positive values represent inconsistency in favor of views that are the popular position for the Consensus issues and agreement with the pro-SDO position for the non-Consensus issues, and vice versa for negative values. Opinion indicates a participant's level of agreement with the popular position for the Consensus issues and agreement with the pro-SDO position for the non-Consensus issues, coded as −2.5 for strongly disagree, and increasing in steps of 1 up to 2.5 for strongly agree. Public is contrast coded as 1 for the participants made to feel as if they would be observed, and −1 for the others. Confidence intervals refer to the unstandardized betas. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

7.3.3. Order effects

Table 11 shows analyses of order effects for the four issues where each side has >35% support, and attitudes were predicted using SDO. Here, the results largely follow the pattern from our previous results, such that the partisan bias in responses to the first condition is largely mitigated as participants seek to be consistent when answering for the second time. The three-way interaction indicating this effect is only significant for two of the four issues, however ($p < .10$ for all four). As opposed to the previous studies, we also found a consistent main effect of SpeechContent on responses, such that the contra-SDO speech receives more protection, even when controlling for the participant's SDO. This tendency in responses is also largely reduced on the second response, however, thus representing another kind of order effect, indicated by the interactions between SpeechContent and Order.

With the exceptions of open borders and meat consumption, the unpopular opinions in our sample were so unpopular that over 90% of participants indicated strong opposition to them. Thus, for these four issues (concerning rape, pedophilia, racism, and stoning), the measured levels of opposition/agreement were unusable as predictors in regression analyses. We could still explore order effects, however, since SpeechContent is now almost perfectly correlated with attitudes towards that speech. An order effect would then be represented by a significant interaction between SpeechContent and Order, so that the effect of speech content on responses is weaker on the second response than the first. In our data (Table 12), we found no evidence of any such interactions, suggesting that the desire to be consistent in speech judgments does not extend to issues were one side is both highly unpopular and highly provocative. The effect of SpeechContent on responses was significant at $p < 10^{-16}$ for all four of these issues.

For the issues of meat consumption and open borders there were

Table 11
Fixed effects from mixed effects regression models predicting support for harsh measures to counter speech, for the four debates where both sides have more than 35% support in the US, in Study 6. Predictors are SDO, Speech, Order, and their interactions. The models also include random intercepts for each subject.

	Abortion			*Gun control*			*Trump*			*Western Culture*		
	β	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$
(Intercept)		3.15	[2.88, 3.42]		3.00	[2.75, 3.63]		2.89	[2.65, 3.12]		3.08	[2.85, 3.31]
SDO	0.05	0.10	[−0.11, 0.30]	0.14*	0.25	[0.06, 0.44]	−0.04	−0.08	[−0.26, 0.10]	0.07	0.11	[−0.06, 0.29]
SpeechContent	−0.23***	−0.55	[−0.82, −0.28]	−0.17**	−0.44	[−0.70, −0.18]	−0.24***	−0.56	[−0.80, −0.33]	−0.23***	−0.51	[−0.73, −0.28]
Order	−0.01	−0.05	[−0.43, 0.33]	0.01	−0.06	[−0.43, 0.31]	−0.06	−0.27	[−0.60, 0.06]	−0.04	−0.18	[−0.50, 0.15]
SDO*SpeechContent	0.20***	0.37	[0.16, 0.57]	0.15**	0.28	[0.08, 0.47]	0.22***	0.39	[0.21, 0.35]	0.23***	0.39	[0.21, 0.56]
SDO*Order	−0.04	−0.09	[−0.38, 0.20]	−0.05	−0.14	[−0.41, 0.13]	0.03	0.06	[−0.19, 0.32]	−0.06	−0.14	[−0.39, 0.11]
SpeechContent*Order	0.16**	0.56	[0.18, 0.94]	0.19***	0.66	[0.29, 1.03]	0.17***	0.57	[0.25, 0.91]	0.10†	0.30	[−0.02, 0.63]
SDO*SpeechContent*Order	−0.15*	−0.38	[−0.67, −0.09]	−0.09†	−0.23	[−0.51, 0.04]	−0.14**	−0.36	[−0.61, −0.10]	−0.09†	−0.22	[−0.47, 0.02]

Note. As the focus of investigation is now the topics of discussion rather than the method used to restrict speech, the analyses are here differentiated according to topic. SpeechContent is coded 1 for the condition with Pro-SDO speech and −1 for the condition with Contra-SDO speech. Order is coded 0 for the first time responding, 1 for the second. Confidence intervals refer to the unstandardized betas. †p ≤ .10, *p ≤ .05, **p ≤ .01, ***p ≤ .001.

Table 12
Fixed effects from mixed effects regression models predicting support for harsh measures to counter speech, in Study 6, for the remaining four of the issues where more than 90% of the US population supports the same side. Topic headers correspond to the unpopular position for each issue. Predictors are Speech, Order, and the interaction between them. The models also include random intercepts for each subject.

	Race & IQ			*Sexual Assault*			*Pedophilia*			*Stoning*		
	β	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$
(Intercept)		3.84	[3.57, 4.10]		4.16	[3.86, 4.47]		4.73	[4.42, 5.04]		4.46	[4.14, 4.77]
SpeechContent	−0.44***	−1.35	[−1.61, −1.09]	−0.58***	−2.20	[−2.51, −1.90]	−0.63***	−2.53	[−2.84, −2.22]	−0.56***	−2.17	[−2.48, −1.86]
Order	−0.04	−0.23	[−0.60, 0.14]	0.01	0.04	[−0.39, 0.47]	−0.06	−0.47	[−0.90, −0.03]	−0.01	−0.10	[−0.54, 0.34]
SpeechContent*Order	−0.02	−0.11	[−0.48, 0.27]	0.02	0.09	[−0.34, 0.52]	0.02	0.11	[−0.33, 0.54]	−0.01	−0.05	[−0.49, 0.40]

Note. SpeechContent represents the condition: it is coded 1 for speech supporting the popular position with >90% support in the US, and −1 for speech opposing it. Order is coded 0 for the first time responding, 1 for the second. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001.

Table 13
Fixed effects from mixed effects regression models predicting support for harsh measures to counter speech, in Study 6, for two of the debates where >90% of the US population supports the same side (that borders should not be completely open, and that eating meat is not evil, respectively). In addition to the predictors in the table, the models also include random intercepts for each subject.

	Open Borders			*Meat Consumption*		
	β	*b*	$CI_{95\%}$	*β*	*b*	$CI_{95\%}$
(Intercept)		2.86	[2.52, 3.22]		3.27	[3.40, 3.63]
Opinion	0.09	0.11	[−0.10, 0.32]	0.30***	0.57	[−0.05, −0.02]
SpeechContent	−0.20*	−0.40	[−0.75, −0.05]	−0.07	−0.17	[−0.19, 0.23]
Order	0.04	0.16	[−0.34, 0.65]	0.07	0.34	[−0.10, 0.30]
Opinion*SpeechContent	0.01	0.01	[−0.20, 0.21]	0.07	0.10	[0.18, 0.35]
Opinion*Order	−0.00	−0.00	[−0.30, 0.29]	−0.02	−0.04	[−0.07, 0.04]
SpeechContent*Order	0.20*	0.58	[0.09, 1.08]	0.18	0.58	[−0.10, 0.35]
Opinion*SpeechContent*Order	0.18	0.31	[0.01, 0.60]	0.15	0.28	[−0.53, −0.22]

Note. Opinion indicates a participant's level of agreement with the unpopular position, coded as −2.5 for strongly disagree, and increasing in steps of 1 up to 2.5 for strongly agree. SpeechContent represents the condition: it is coded 1 for the speech supporting the popular position with >90% support in the US, and −1 for speech opposing it. Order is coded 0 for the first time responding, 1 for the second. Confidence intervals refer to the unstandardized betas. *p ≤ .05, **p ≤ .01, ***p ≤ .001.

more than 20% of our participants indicating support for the unpopular position (suggesting that MTurkers are not fully representative of the US population in this respect), so we here ran models using these opinion measures as predictors (Table 13). For meat consumption there were no significant biases in speech protection to begin with, so there was no room for any order effects to mitigate these biases. However, we found (an unpredicted) main effect of Opinion, such that people who are more inclined to agree that eating meat is evil allow for stronger measures to combat speech both for and against their own position. For open borders, there was a main effect of SpeechContent, such that participants generally allowed harsher measures against proponents than against opponents of open borders. This tendency was cancelled out for the second responses, as indicated by the interaction between SpeechContent and Order.

8. Discussion

The present series of studies qualifies, and consolidates, emerging findings that people are selectively protective of views they support when making free speech judgments. Between the hypotheses that people either (A) operate with explicitly different standards for different speech (*content-sensitive speech norms*), or (B) seek to maintain a façade of operating with a single standard for all speech (*content-neutral speech norms*) while selectively adjusting this standard to suit the context, the present investigation lends partial support to (B). When making judgments for different kinds of speech in the same context, several participants adjust their judgments to maintain complete content-neutrality, and still more participants responded in ways that were at least less content-sensitive than they otherwise would have been. This fits well with the more general notion that people tactically adjust their moral judgments (Bartels, 2008), as well as the notion that people lack awareness of the flexibility of their moral convictions and rather feel like steadfast believers in the values and norms that happen to currently be helpful (Ditto et al., 2018; Trivers, 2000).

In Study 1 we showed that when judging a series of scenarios raising free speech issues, the tendency to selectively protect supported speech was dampened when participants had been previously exposed to other scenarios with speech of the opposite ideological flavor, suggesting that participants seek to maintain consistent standards for all speech.

Studies 2 and 3 further corroborated this finding by investigating *order effects*: when participants responded to both conditions of our design sequentially, without knowing that the second condition was coming when answering the first, and without being able to go back to alter prior responses, the order of presentation was found to have a large impact on responses. To seem consistent, participants gave similar responses to the second version of each scenario as those they gave for the first. And since their first responses were biased in favor of speech they supported, their second responses then produced almost the same level of bias in favor of speech they opposed.

It is plausible that attribute evaluability effects, as described by Hsee et al. (1999), have influenced our results here. As per their evaluability hypothesis, when switching from separate to joint evaluations as we do in our design, aspects that are not easy to evaluate, such as the long term impacts of speech on society, will be given less weight. This could lead to responses that are more similar to each other than they would be in separate evaluations.

Studies 4 and 5 found that people also largely support the more abstract principle that one should generally be content-neutral in free speech judgments, in addition to supporting it for both sides within specific debates. Study 4 also found that support for content-neutral norms is further heightened when they are proposed by someone from one's political opposition, and also contrasted with the particular brand of content-sensitivity that favors one's opposition. Study 5 disambiguated this finding, and suggested that it was the contrasting with opposition-slanted content-sensitivity, rather than the political affiliation of the person proposing content-neutrality, that was the deciding factor influencing the increase in support for content-neutrality.

Study 6 explored the boundaries of the need to maintain consistency in judgments, by investigating the presence of an Overton Window of acceptable discourse. Here, we found that the tendency to make similar judgments for both sides of an issue can disappear completely, under certain conditions. Specifically, people are generally much less concerned with giving both sides equal speech rights when they strongly dislike one side and this side is also condemned by a large majority of society. This could reflect how it is less costly to support restrictions on despised views when there is only a small minority that would object to this.

If our results are generalized to other domains of moral and social judgment, they suggest that it is normal to have at least two sets of norms and values: One set is the "official" one, which we present to others; the other set reflects our true preferences, which we keep hidden (perhaps even from ourselves). Our publicly endorsed norms and values will typically represent a compromise between the conflicting interests of many parties. For example, content-neutral speech norms can be seen to represent a compromise between speech norms that are explicitly content-sensitive in either one or the other ideological direction. These kinds of norms can be seen as our offer in a negotiation with parties who have opposing interests to our own; if these parties are willing to adjust, then we are willing to return the favor, and settle on a set of rules everyone can live with. When there are situations giving room for plausible deniability, however, such that we no longer need to compromise, and can instead safely work towards our true goals, we take advantage of this. This way, we are at the same time able to hold the ideological opposition accountable whenever they detectably transgress on shared norms, since we can claim that we are keeping our end of the bargain, yet we are also able to transgress on these very norms ourselves whenever we can do so undetectably.

A potential ambiguity with our series of experiments concerns whether participants were truly self-deceived about their moral consistency or if they rather were conscious about the selective nature of their judgments. In principle, it is possible that participants at all times had conscious access to their "true judgments" about the issues in our scenarios, and that they purposely adjusted away from these true judgments to suit the context. While our studies were all completely anonymous, it is still possible that participants had the feeling of responding before observers who were evaluating them, and thus felt the need to be dishonest. This interpretation is rendered less likely, however, by our finding in Study 6 that participants manipulated to feel more like they were observed did not have a stronger tendency to be consistent in their judgments. This indicates that insofar as people perform moral consistency before an audience, this audience has no different effect when it includes only themselves or others. We think the most parsimonious interpretations of these results is that most of our participants who judged consistently did believe that speech content was irrelevant to their judgments, even as they unconsciously took it into account. Such a conceptualization of moral judgment processes as being largely hidden from our conscious minds has broadening support (Haidt, 2012; Mercier & Sperber, 2011; Trivers, 2011).

While we have showed that there is a limit to people's need to be consistent in judgments about free speech, we also find that this limit is located at different places for different people. Even for mainstream issues there were many who were explicitly inconsistent. In other words, rather than talking about a single Overton Window for all of society, it is perhaps more accurate to talk of separate Overton Windows for separate people, that only partly overlap. This could reflect divisions in society. Some values are sacred almost everywhere (Tetlock, 2003), but other values are only sacred to some and not to others.

Following from this, it can be argued that as divisions between mainstream factions increase, as is arguably the case in the US in recent years (Garimella & Weber, 2017), explicitly content-sensitive speech norms in support of coalition-specific sacred values might rise in popularity among all camps, as the hope of sustaining cooperation

across ideological divides decreases. An optimistic take-away from the present results is then that among Americans polarization seems to not (yet?) have reached the point were most people have completely given up on having a shared set of norms.

9. Conclusion

While judgments of actions to restrict speech systematically depend on whether one supports its ideological content, we have shown that people seek to correct these biases whenever they may be revealed, so as to indicate that this is not the case for them: In these cases, many people adjust their judgments and extend similar protections to speech they support and oppose. The exception is for speech that is widely considered to be highly damaging among most political and cultural groups: People are generally more open about supporting restrictions for that kind of speech.

To the extent that people truly wish to follow the precept of not taking speech content into account when making judgments about speech rights, our study suggests that it is not sufficient to just go with one's intuitions. The processes producing our intuitions seem to be more tactical and devious than we are able to tell from introspection. A good approach could then be to imagine oneself as blind to the specifics of each case: how would one feel about speech being counteracted in a certain way if it was unknown whether this speech affirmed or contradicted your own views on the matter?

Author contributions

N.H.E & L.T. designed research; N.H.E collected and analyzed data; N.H.E. & L.T. wrote and edited the paper.

Declaration of Competing Interest

None.

Acknowledgements

We thank our two anonymous reviewers for their insightful suggestions on a previous version of the manuscript. The research was funded by grants 0602-01839B and 231157/F10 from the Danish- and Norwegian Research Councils, respectively (to L.T.).

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104623.

References

Gallup. (2020a). Guns. Retrieved from https://news.gallup.com/poll/1645/guns.aspx.
Alexander, S. (2014a). In Favor of Niceness, Community, and Civilization. Retrieved from http://slatestarcodex.com/2014/02/23/in-favor-of-niceness-community-and-civilization/.
Alexander, S. (2014b). Beware Isolated Demands For Rigor. Retrieved from http://slatestarcodex.com/2014/08/14/beware-isolated-demands-for-rigor/.
Alexander, S. (2018). Conflict vs. mistake. Retrieved from https://slatestarcodex.com/2018/01/24/conflict-vs-mistake/.
Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345.
Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*(2), 381–417.
Bey, A. (2017). Does your right to free speech extend to the workplace?. Retrived from https://www.mediabistro.com/be-inspired/right-free-speech-workplace/.
Blumer, H. (1958). Race prejudice as a sense of group position. *Pacific sociological review, 1*(1), 3–7.
Boehm, C. (2009). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.
Boström, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy, 10*, 44–79.
Brandt, M. J., Chambers, J. R., Crawford, J. T., Wetherell, G., & Reyna, C. (2015). Bounded openness: The effect of openness to experience on intolerance is moderated by target group conventionality. *Journal of Personality and Social Psychology, 109*(3), 549.
Brennan, O. (2015). The Enemy Control Ray. Retrieved from https://thingofthings.wordpress.com/2015/05/15/the-enemy-control-ray/.
Campbell, B. (2018). The free speech crisis on campus is worse than people think. Retrived from https://quillette.com/2018/11/14/the-free-speech-crisis-on-campus-is-worse-than-people-think/.
Core Team, R. (2020). R: A language and environment for statistical computing. In *R foundation for statistical computing*. Vienna, Austria. URL https://www.R-project.org/.
Crawford, J. T. (2014). Ideological symmetries and asymmetries in political intolerance and prejudice toward political activist groups. *Journal of Experimental Social Psychology, 55*, 284–298.
Crawford, J. T., & Pilanski, J. M. (2014). Political intolerance, right and left. *Political Psychology, 35*(6), 841–851.
Crawford, J. T., & Xhambazi, E. (2015). Predicting political biases against the Occupy Wall street and tea party movements. *Political Psychology, 36*(1), 111–121.
Crook, C. (2017). Google moves into the business of thought control. Retrieved from https://www.bloomberg.com/opinion/articles/2017-08-14/google-moves-into-the-business-of-thought-control.
Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London Series B. Biological Sciences, 205*(1161), 489–511.
DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*(2), 477.
Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., … Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science, 1745691617746796*.
Fahmy, D. (2020). 8 key findings about Catholics and abortion. Retrieved from https://www.pewresearch.org/fact-tank/2020/10/20/8-key-findings-about-catholics-and-abortion/.
Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.
Flowers, A. (2015). The Most Common Unisex Names In America: Is Yours One Of Them?. Retrieved from https://fivethirtyeight.com/features/there-are-922-unisex-names-in-america-is-yours-one-of-them/.
Galef, J. (2018). Insightful articles on free speech & social justice. Retrieved from https://juliagalef.com/2018/01/10/insightful-articles-on-free-speech-social-justice/.
Gallup. (2020). Presidential approval ratings – Donald Trump. Retrieved from https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx.
Garimella, V. R. K., & Weber, I. (2017). A long-term analysis of polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 528–531.
Graham, P. (2004). What you can't say. Retrieved from http://www.paulgraham.com/say.html.
Green, J., & Karolides, N. J. (2014). *Encyclopedia of censorship*. Infobase Publishing.
Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion* (Vintage).
Hansell, S., & Mechanic, D. (1985). Introspectiveness and adolescent symptom reporting. *Journal of Human Stress, 11*(4), 165–176.
Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., … Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO$_7$ scale. *Journal of Personality and Social Psychology, 109*(6), 1003.
Ho, A. K., Sidanius, J., Pratto, F., Levin, S., Thomsen, L., Kteily, N., & Sheehy-Skeffington, J. (2012). Social dominance orientation: Revisiting the structure and function of a variable predicting social and political attitudes. *Personality and Social Psychology Bulletin, 38*(5), 583–606.
Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576.
Imhoff, R., & Jahnke, S. (2018). Determinants of punitive attitudes toward people with pedophilia: Dissecting effects of the label and intentionality ascriptions. *Archives of Sexual Behavior, 47*(2), 353–361.
Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2014). *World values survey: Round six-country-pooled datafile version* (p. 12). Madrid: JD Systems Institute.
Kahan, D. M. (2015). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (pp. 1–16).
Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.
Kleppestø, T. H., Czajkowski, N. O., Vassend, O., Røysamb, E., Eftedal, N. H., Sheehy-Skeffington, J., … Thomsen, L. (2019). Correlations between social dominance orientation and political attitudes reflect common genetic underpinnings. *Proceedings of the National Academy of Sciences, 116*(36), 17741–17746.
Krysan, M., & Moberg, S. (2016). *A portrait of African American and White racial attitudes*. University of Illinois, Institute of Government and Public Affairs.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480.
Clarity Campaign Labs. (2018). Partisan Name Calculator. Retrieved from https://www.claritycampaigns.com/names.
Lehman, J. (2012). A Brief Explanation of the Overton Window. Retrieved from http://www.mackinac.org/OvertonWindow.

LeVine, R. A., & Campbell, D. T. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior*.
Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology, 30*(1), 67–92.
Lucas, B. J., & Kteily, N. S. (2018). (Anti-) egalitarianism differentially predicts empathy for members of advantaged versus disadvantaged groups. *Journal of Personality and Social Psychology, 114*(5), 665.
McClelland, J. (1996). The place of Elias Canetti's crowds and power in the history of Western social and political thought. *Thesis eleven, 45*(1), 16–27.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.
Mill, J. S. (1859/1966). On liberty. In *In A selection of his works (pp. 1–147)*. London: Palgrave.
Munroe, R. (2014). Free Speech. Retrieved from https://xkcd.com/1357/.
Newport, F. (2012). In U.S., 5% consider themselves vegetarians. Retrieved from https://news.gallup.com/poll/156215/consider-themselves-vegetarians.aspx.
Oliphant, J. B. (2017). Women and men in both parties say sexual harassment allegations reflect 'widespread problems in society'. Retrieved from https://www.pewresearch.org/fact-tank/2017/12/07/americans-views-of-sexual-harassment-allegations/.
Petersen, M. B. (2015). Evolutionary political psychology: On the origin and structure of heuristics and biases in politics. *Political Psychology, 36*, 45–78.
Pinker, S. (2007). In defense of dangerous ideas. *Chicago Sun-Times, 15*.
Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*(4), 741.
Research, P. (2011). The American-Western European Values Gap. Retrieved from https://www.pewresearch.org/global/2011/11/17/the-american-western-european-values-gap/.
Rhode, D. L. (2016). *Adultery: Infidelity and the law*. Cambridge, MA: Harvard University Press.
Sidanius, J. (1993). The psychology of group conflict and the dynamics of oppression: A social dominance perspective. In S. Iyengar, & W. J. McGuire (Eds.), *Duke studies in political psychology. Explorations in political psychology* (pp. 183–219). Durham, NC, US: Duke University Press.
Sidanius, J., & Pratto, F. (2004). *Social dominance theory: A new synthesis*.
Simler, K., & Hanson, R. (2017). *The elephant in the brain: Hidden motives in everyday life*. Oxford University Press.
Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320–324.
TheFire (2020). Disinvitation Database. Retrieved from https://www.thefire.org/resources/disinvitation-database/.
Thomas, A. J., Thomsen, L., Lukowski, A. F., Abramyan, M., & Sarnecka, B. W. (2018). Toddlers prefer those who win, but not when they win by force. *Nature Human Behavior, 2*(9), 662–669.
Thomsen, L. (2020). The developmental origins of social hierarchy: How infants and young children mentally represent and respond to power and status. *Current Opinion in Psychology, 33*, 201–208.
Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science, 331*(6016), 477–480.
Thomsen, L., Green, E. G., Ho, A. K., Levin, S., van Laar, C., Sinclair, S., & Sidanius, J. (2010). Wolves in sheep's clothing: SDO asymmetrically predicts perceived ethnic victimization among White and Latino students across three years. *Personality and Social Psychology Bulletin, 36*(2), 225–238.
Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. *Human morality and sociality: Evolutionary and comparative perspectives*, 91–234.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences, 907*(1), 114–131.
Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life* (Basic).
Von Hippel, W., & Trivers, R. (2011). Reflections on self-deception. *Behavioral and Brain Sciences, 34*(1), 41–56.
White, M. H., II, & Crandall, C. S. (2017). Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology, 113*(3), 413.
Wilson, E. O. (1998). The biological basis of morality. *The Atlantic Monthly, 281*(4), 53–70.