Cognition 225 (2022) 105146

Contents lists available at ScienceDirect

Cognition

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

Is political extremism supported by an illusion of understanding?

Steven A. Sloman [*], Marc-Lluis Vives

Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, United States of America

ARTICLE INFO

Keywords:
Explanation
Illusion of explanatory depth
Political psychology
Attitudes
Polarization
Extremism
Moderation
Public policy
Causal models
Mechanism

ABSTRACT

Polarization is rising in most countries in the West. How can we reduce it? One potential strategy is to ask people to explain how a political policy works—how it leads to consequences— because that has been shown to induce a kind of intellectual humility: Explanation causes people to reduce their judgments of understanding of the issues (their "illusion of explanatory depth"). It also reduces confidence in attitudes about the policies; people become less extreme. Some attempts to replicate this reduction of polarization have been unsuccessful. Is the original effect real or is it just a fluke? In this paper, we explore the effect using more timely political issues and compare judgments of issues whose attitudes are grounded in consequentialist reasoning versus protected values. We also investigate the role of social proof. We find that understanding and attitude extremity are reduced after explanation but only for consequentialist issues, not those based on protected values. There was no effect of social proof.

Asking people to explain how a political policy works—how it leads to consequences—induces a kind of intellectual humility (Fernbach, Rogers, Fox, & Sloman, 2013). The process of attempting to explain often results in failure; in general, people are unable to meet their own expectations in giving an account of how policies work. As a result, their judgments of their own understanding of the issues decline (an extension of the illusion of explanatory depth, Rozenblit & Keil, 2002). In Fernbach et al., so did their confidence in their attitudes about the policies; they became less extreme. The original paper showed the effects three times including a study with an incentive compatible measure (people gave money in proportion to the extremity of their attitudes). In short, the studies found that a simple manipulation could demonstrate people's own ignorance to themselves, and use this increase in humility to reduce extremity, a potentially valuable tool in this age of political polarization (Sloman & Fernbach, 2017).

The effect of explanation on humility, on people's sense of understanding, has been replicated several times (Alter, Oppenheimer, & Zemla, 2010; Crawford & Ruscio, 2021; Ebrahimi, Hemmatian, & Purmohammad, 2022; Gaviria et al., 2017; Johnson, Murphy, & Messer, 2016; Vitriol & Marsh, 2018; Voelkel, Brandt, & Colombo, 2018; Zeveney & Marsh, 2016), although the effect on extremity is not so robust. Crawford and Ruscio (2021) reported a failure to replicate our effect on extremity, although Ebrahimi, Hemmatian, & Purmohammad did replicate and Meyers, Turpin, Białek, Fugelsang, and Koehler (2020) obtained mixed results concerning extremity but found that inducing feelings of ignorance through explanation made people more receptive to the expert opinion of economists. Anecdotally, unpublished work of our own and others has mostly replicated the reduction in understanding and only sometimes replicated the effect on extremity.

Why is the effect on extremity not more robust? Sloman and Fernbach (2017) presented pilot data suggesting the effect arises only when issues elicit consequentialist considerations, not when issues are thought about in terms of protected or sacred values. In this paper, we address this possibility and also examine the role of social proof (Cialdini, 1993), whether the effect is influenced by other people's recognition of their illusion of understanding. We also consider Crawford and Ruscio's (2021) failure to replicate Fernbach et al. (2013).

Why did the results of the two studies differ?. Perhaps the findings of Fernbach et al. (2013) were a fluke, although they did find the results three times in the original paper. Perhaps Crawford and Ruscio's (2021) findings were a fluke, although they failed to replicate our results despite three attempts with larger sample sizes than we used. It also might be that we live in a different political climate than we did in 2012, that the effects we reported in 2013 no longer occur in the political domain. Polarization has taken firmer hold in society and it is possible that policy polarization is not so easily reduced. Such an account is obviously difficult to test. Another possibility that is very difficult to gauge empirically is that the environment for testing subjects was different between the two papers' testing periods. Both papers used Amazon's Mechanical Turk to obtain online participants, but the site has

[*] Corresponding author.
E-mail address: steven_sloman@brown.edu (S.A. Sloman).

https://doi.org/10.1016/j.cognition.2022.105146
Received 18 February 2022; Received in revised form 14 April 2022; Accepted 22 April 2022
Available online 6 May 2022

been shown to recently produce relatively poor data quality (Eyal, Rothschild, Gordon, Evernden, & Damer, 2021), and there is now evidence of substantial deception by participants on it (Sharpe Wessling, Huber, & Netzer, 2017).

1. Consequentialism versus protected values

One interpretation of the original illusion of explanatory depth is that it results from subjects having different interpretations of the question about their degree of understanding before and after they engage in explanation. In the domain of simple artifacts studied by Rozenblit and Keil (2002), the act of attempting to explain could have changed subjects' framing of the understanding question from one about intuitive knowledge to one about articulable, deliberative knowledge. Similarly, the illusion in the policy domain could reflect a reframing from a protected values perspective to a consequentialist one. That is, subjects may be in the habit of simplifying policy issues by thinking about them in terms of the protected values they embody (Baron & Spranca, 1997; Tetlock, 2003) and they might answer the initial question about their understanding from this perspective. For example, positive attitudes regarding sanctions on Iran may reflect a protected value about American dominance—and negative attitudes one about the value of pluralism—rather than a difficult assessment of the actual outcomes that the policy would lead to. But the requirement to explain would force subjects to try to work out those consequences; in other words, explanation might induce a consequentialist perspective. Understanding the consequences of a policy requires a causal analysis that projects into the future and is thus harder than understanding the protected values associated with a policy. Hence, understanding judgments would be lower following an attempt to explain than prior to one.

One implication of this account is that the illusion should not occur for issues that do not lend themselves to a consequentialist perspective, issues that elicit a protected values perspective. Sloman and Fernbach (2017) report pilot data consistent with this implication. The data suggest that the illusion of explanatory depth did not occur with policies that induced a strong sacred values orientation (e.g., assisted suicide, whether doctors should be able to give individuals experiencing extreme suffering assistance and approval to commit suicide). Perhaps, over time, people's attitudes toward policies become entrenched as sacred or protected values such that Crawford and Ruscio's (2021) subjects did not show the effect because they were more likely to have a values-based orientation than the Fernbach et al. (2013) subjects. More generally, we address the question here whether the type of issue makes a difference to the effect of explanation on understanding and extremism.

2. Social proof

Ever since the classic work of Cialdini (1993), the influence of learning about others' actions on an individual's actions has been recognized. We are more likely to use less electricity if our bill tells us how little our neighbors use (Hunt Allcott, 2011) or to be more selfish if our in-group is (Vives, Cikara, & FeldmanHall, 2021). Social proof of this kind might also influence the illusion of explanatory depth: People might be more willing to admit to themselves and others that they do not understand as well as they thought they had upon learning that others have also experienced the phenomenon, and puncturing their illusion might in turn reduce the extremity of their attitude. We include a manipulation in our experiment to evaluate this possibility.

3. Experiment

We report an experiment designed to address the reasons for Crawford and Ruscio's (2021) failure to replicate Fernbach et al. (2013) focusing on the possibility that the effect occurs only for issues deemed consequentialist. It also examines the role of social proof on the effect.

The experiment uses essentially the same methodology as Fernbach et al.'s Experiment 2 except that it employs more up-to-date political policies, varies whether the issue elicits a protected values versus consequentialist frame, and varies the presence of a social proof induction. It also differs in that we obtained initial ratings of four policies rather than six. To ensure that the experiment does not suffer from the bot or professional survey taker issues that have been identified with Mechanical Turk, we used a different platform to find internet subjects, Prolific.

As in the previous work, we compare the effect of causal explanation to the effect of asking people to enumerate the reasons why they hold the policy attitude they do. Unlike our explanation findings, there is evidence that when people think about why they hold a position their attitudes sometimes become more extreme (Hirt & Markman, 1995; Ross, Lepper, Strack, & Steinmetz, 1977; Tesser, 1978). Reason generation encourages people to access a supportive rationale, thereby supporting their commitment to a prior position. Crawford and Ruscio (2021) found such a positive effect of reason generation though Fernbach et al. (2013) found little effect. In contrast, asking for a mechanistic explanation forces subjects to confront their lack of understanding, thereby decreasing their commitment.

To vary social proof, we either did or did not present subjects with three mock statements from others on social media (a pseudo-Twitter feed) that supported subjects' subjective experience of failing to be able to explain a policy they thought they had understood (see Fig. 1) immediately after attempting to explain or generate reasons.

4. Methods

4.1. Participants and design

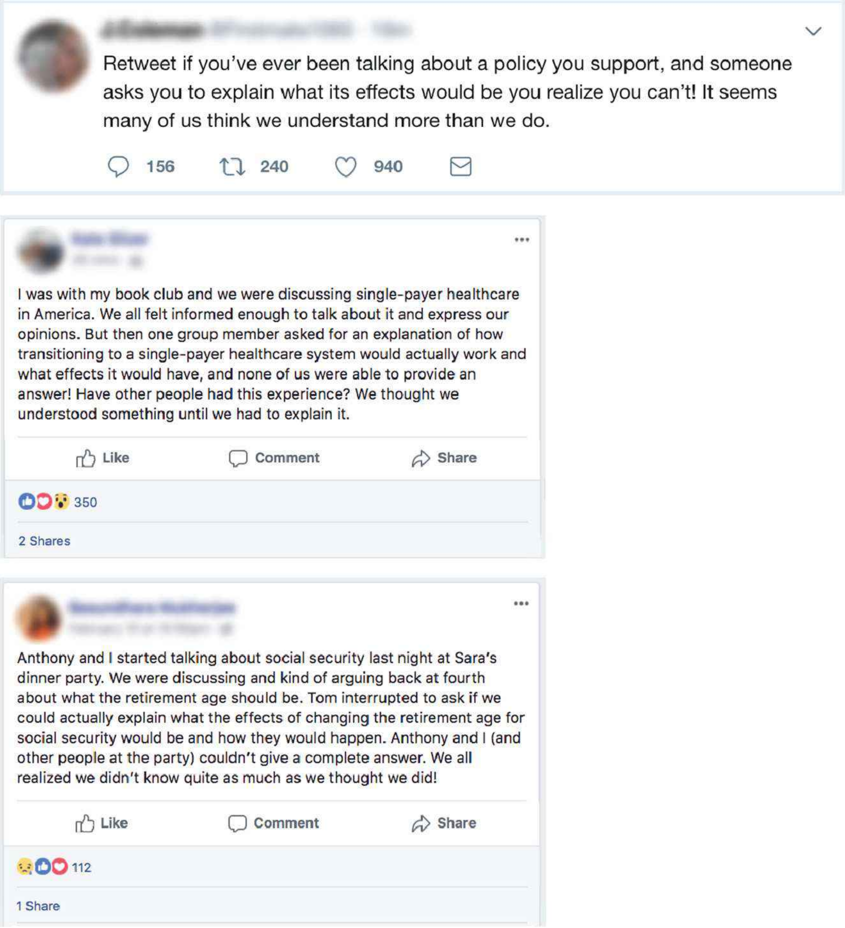
We recruited 739 residents of the United States from the Prolific internet subject platform in exchange for a small payment. Participants were 51% female, 46% male, and 3% other, with an average age of 33.5. Sixty-four per cent identified as Democrats, 17% Republicans, and 19% independents.

4.2. Materials and procedure

Participants were asked to state their position on four political policies on a 7-point scale with endpoints "strongly agree" and "strongly disagree." We selected four timely political issues. Two of the policies were consequentialist: (1) The U.S. government should implement a nationwide cancellation of federal student loans for all borrowers. (2) Governments should institute a limit on the market pricing of prescription medications. The other two policies were based on protected values: (3) Governments should implement a policy that widens and more strictly enforces the background check qualifications necessary to legally purchase a firearm. (4) The qualifications necessary to obtain legal citizenship in the U.S. should be stricter.

All participants were next trained to use a rating scale to quantify their level of understanding. These instructions were modeled on Rozenblit and Keil (2002), but rather than describing different levels of understanding for an object (in their case, a crossbow), we described different levels of understanding for a political issue. After reading these instructions participants were asked to judge their level of understanding of the four policies (e.g. "How well do you understand the impact of the U.S. government implementing a nationwide cancellation of federal student loans for all borrowers?") using a 7-point Likert scale.

After judging their understanding of all four policies, participants proceeded to a new screen on which they were asked to provide either a mechanistic explanation or reasons for one of the policies. These instructions were also adapted from Rozenblit and Keil (2002) and an example is shown in the SOM. Participants were then asked to re-rate their understanding on the same scale as before and rate their position. After completing these questions participants repeated the process for a second issue. The policies were blocked such that participants



Retweet if you've ever been talking about a policy you support, and someone asks you to explain what its effects would be you realize you can't! It seems many of us think we understand more than we do.

I was with my book club and we were discussing single-payer healthcare in America. We all felt informed enough to talk about it and express our opinions. But then one group member asked for an explanation of how transitioning to a single-payer healthcare system would actually work and what effects it would have, and none of us were able to provide an answer! Have other people had this experience? We thought we understood something until we had to explain it.

Anthony and I started talking about social security last night at Sara's dinner party. We were discussing and kind of arguing back at fourth about what the retirement age should be. Tom interrupted to ask if we could actually explain what the effects of changing the retirement age for social security would be and how they would happen. Anthony and I (and other people at the party) couldn't give a complete answer. We all realized we didn't know quite as much as we thought we did!

Fig. 1. Social validation through social media.

explained both a consequentialist issue and a protected values one.

Half the subjects in the explanation condition and half in the reasons condition were also shown social proof. Immediately after their attempt to explain they were shown the images in Fig. 1. Finally, we asked several demographic questions: gender, level of education, political party affiliation, ideology (liberal to conservative).

4.3. Manipulation check

To validate our main manipulation, 32 additional participants were asked to evaluate to what degree their attitudes on each policy were based on sacred-values or consequentialist reasons on a scale from 1 (based on sacred values) to 5 (based on consequentialist reasons). One subject failed to provide a judgment for the student loan issue. Corroborating our manipulation, participants judged consequentialist issues as more consequentialist (average = 3.19, sd = 1.43) than sacred values issues (average = 2.20, sd = 1.27; t (125) = 4.10, *p*-value <0.001).

5. Results

*Data can be found at*https://osf.io/n6bf7/?view_only=7aac39a8f8684d9697f8131dad41a2d0

Mechanistic explanations reduce extremity attitudes for consequentialist issues. First, we examined the *mechanism* condition because it was the main focus of Crawford and Ruscio (2021). See Fig. 2. Judgments of understanding were submitted to a repeated measures ANOVA with before and after mechanistic explanation and consequentialist versus protected values policies as within-subject variables. Like previous studies, participants decreased their sense of understanding after attempting to provide a mechanistic explanation, $F\ (1, 304) = 71.52$, $p < 0.001$, *partial* $\eta^2 = 0.19$. This main effect was further qualified by an interaction with consequentialist versus protected value issues such that the reduction of understanding was larger for protected value policies than for consequentialist policies, $F\ (1, 304) = 7.40$, $p = 0.006$, *partial* $\eta^2 = 0.02$. This effect was probably a consequence of higher understanding of protected value than consequentialist issues before providing an explanation, $F\ (1, 304) = 4.74$, $p = 0.03$, *partial* $\eta^2 = 0.003$, leaving more room for correcting their rating of understanding after trying to provide a mechanistic explanation.

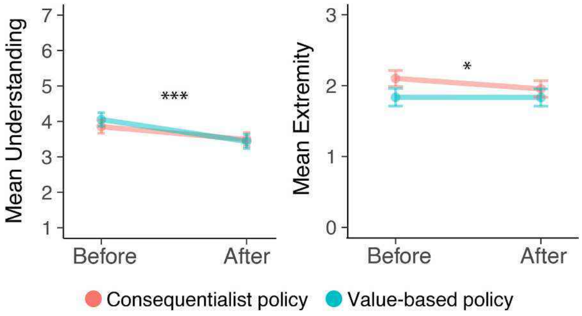


Fig. 2. Results of Experiment with 95% confidence intervals. Left figure shows the reduction in understanding by type of policy before and after attempting to provide a mechanistic explanation. Right figure shows the significant reduction in extremity for consequentialist policies, in contrast with no-change for value-based policies. *** $p < 0.001$, * $p < 0.05$.

The same repeated measures ANOVA was run for attitudinal extremity. The main effect of before versus after mechanistic explanation was right on the boundary of significance, F (1, 304) = 3.85, $p = 0.05$, partial $\eta^2 = 0.01$. Importantly, it was qualified by a significant interaction with the type of policy, F (1, 304) = 3.85, $p = 0.02$, partial $\eta^2 = 0.01$. Planned contrasts revealed that, while there was no significant change in extremity for protected values policies (t (304) = 0.06, $p = 0.95$, M = 0.003), extremity for consequentialist policies showed a significant decrease (t (304) = 3.03, $p = 0.002$, M = 0.15, see Fig. 2). Overall, attempting to provide mechanistic explanations caused people to reduce their attitude extremity when policies relied on consequentialist reasoning, but not when policies were based on protected values.

We correlated the magnitude of change in understanding before and after providing a mechanistic explanation for consequentialist policies with the magnitude of change in attitude extremity. Replicating Fernbach et al.'s (2013) results, we found a significant positive correlation: the larger the reduction in reported understanding, the larger the reduction in attitude extremity: r (303) = 0.26, $p < 0.001$.

Providing reasons for or against policies does not reduce understanding or attitude extremity. We next compared the magnitude of change in understanding within consequentialist policies across the between-subjects mechanistic explanation versus reasons manipulation. Again, we found a main effect of before versus after, F (1, 737) = 8.62, $p = 0.003$, *partial* $\eta^2 = 0.01$. Crucially, replicating our previous phenomenon, this main effect was further qualified by a significant interaction with the type of text generated, F (1, 737) = 23.01, $p < 0.001$, *partial* $\eta^2 = 0.03$. Planned comparisons revealed that, while participants reduced their sense of understanding after attempting to explain how the policy worked, (t (304) = 4.75, $p < 0.001$, $M = 0.36$), their sense of understanding did not change after listing reasons for or against the policy (t (433) = −1.52, $p = 0.13$, $M = -0.09$).

The same repeated-measures ANOVA was conducted for attitude extremity. Results revealed a main effect of before and after, F (1, 737) = 7.48, $p = 0.006$, *partial* $\eta^2 = 0.01$, that was qualified by an interaction with the type of explanation, F (1, 737) = 4.52, $p = 0.03$, *partial* $\eta^2 = 0.006$. Replicating our previous work, planned comparisons demonstrated that the reduction of attitude extremity occurred only after participants attempted to provide a mechanistic explanation of how the policy worked (t (304) = 3.03, $p = 0.002$, $M = 0.15$), not reasons (t (433) = 0.49, $p = 0.62$, $M = 0.02$). Even though providing reasons did not change people's attitudes, we still observed a significant correlation between changes in understanding and attitude change, r(432) = 0.23, $p < 0.001$. This correlation probably captures a general tendency to report different attitudes when changing one's mind about one's level of understanding.

Social Proof does not modulate the effect of explanation on attitude extremity. Finally, we examined the between-subjects social proof manipulation. We conducted a repeated measures ANOVA for consequentialist policies when participants had to give mechanistic explanations including social proof together with the before and after dependent variable. Neither the understanding nor the attitude ratings were affected by the social proof manipulation (all ps > 0.1.

6. Discussion

This study replicates the original Fernbach et al. (2013) experiments showing a reduction of both the sense of understanding and attitude extremity of policies by virtue of an attempt to explain how the policy works. Like Fernbach et al., the effects were absent when subjects generated reasons rather than an explanation. Unlike causal explanations, reasons can draw on values, hearsay, and general principles that do not require much knowledge. We did observe a necessary condition for the effects. The effect of explanation was only reliable for issues whose attitudes were grounded in consequentialist reasoning—in the potential outcomes of the policies—and not for issues that elicited more protected values. To our surprise, we did not find an effect of social proof.

Why did these results differ from those of Crawford and Ruscio (2021)? The fact that we replicated Fernbach et al. (2013) using a larger sample size reinforces the reliability of those results. But we suspect Crawford and Ruscio's results are also credible given that they failed to replicate Fernbach et al. three times and they also used a larger sample size.

We attribute the different results to three differences between our study and theirs: First, like Fernbach et al. (2013) and unlike Crawford and Ruscio (2021), our study used policies under current debate and only timely issues are likely to be perceived through a partisan lens. And issues that are seen as partisan are likely to be judged more extremely and thus leave more room for extremism to be reduced.

Second, the effect of using more timely issues is that people are more likely to bring partisan frames to bear in their initial framing of the issues. They are more likely to be aware of their party's position, providing an initial framing and sense of familiarity. This initial framing and sense of familiarity could be the source of the extremism of attitudes ("if those I trust and respect have an opinion on this issue, then I do too"). The subsequent attempt to explain necessarily elicits an attempt to consider consequences and how they come about and this consequentialist framing can then be revealing of what the explainer does not know. But when issues are not timely, there is no such initial sense of familiarity and thus less of an illusion for the explanation to puncture. Of course, when people reject the consequentialist framing, as they may well do for issues like abortion whose attitudes are generally based on sacred values, then the explanation has no effect because people do not assign the explanation relevance. They are satisfied by the knowledge and attitude housed in their protected values. This would explain why we only found the effect for issues that elicited consequentialist and not protected value frames. Unpublished work in one of our labs has replicated the findings distinguishing consequentialist versus protected values issues, but only in a between-subjects design. Within-subject designs that ask people to explain both sorts of issues may inhibit subjects' ability to keep the two types of considerations separate.

The fact that we replicated the original Fernbach et al. (2013) study suggests that the reason for Crawford and Ruscio's (2021) failure to replicate is not because of the current highly polarized political climate. However, unlike the previous studies that obtained their subjects from Amazon's Mechanical Turk, we used a different internet subject market, Prolific. So it is possible that one reason for Crawford and Ruscio's failure to replicate is the relatively poor data quality currently on Mechanical Turk (Eyal et al., 2021) and a possible increase in the number of deceptive participants on Mechanical Turk between 2013 and 2018 (Sharpe Wessling et al., 2017).

In sum, the failure to replicate the effect of explanation certainly suggests that the effect is not as robust as we once thought. However, our replication of it along with others in the literature prevent us from giving up on the idea that mechanistic explanation leads to more moderate positions by forcing people to confront their ignorance. But it does seem that we should only expect this effect with issues that are grounded in an evaluation of outcomes (consequentialist issues), not issues whose attitude is derived from protected values. It is possible that the frame applied to an issue is not always fixed but sensitive to contextual variables. Indeed, when frames are not fixed, the request for a causal explanation may have the effect of changing a frame from one about protected values to consequentialist. As we have seen, such a shift could have the benefit of reducing polarization.

Author's note

We would like to thank Phil Fernbach and Babak Hemmatian for comments, Rachael Perrotta for research assistance, Todd Rogers for suggesting some of the issues we used, Caroline Turner for developing the social proof manipulation, and Louis van Camp for help with collecting data and data analysis.

Author contributions

Steven Sloman oversaw the design and implementation of the experiment and wrote the first draft.

Marc-LLuis Vives analyzed the data and commented on the draft. The work was funded by a grant from Brown University.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*(1), 1–16.

Cialdini, R. B. (1993). *The psychology of persuasion*. New York: Quill/William Morrow.

Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman's (2013) Findings. *Psychological Science, 32*(4), 611–621.

Ebrahimi, A., Hemmatian, B., & Purmohammad, M. (2022). The Illusion of Explanatory Depth Differentially Moderates Attitudes Towards National and International Issues. *Proceedings of the Cognitive Science Society*. Cognitive Science Society.

Eyal, P., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20. https://doi.org/10.3758/s13428-021-01694-3

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24*(6), 939–946.

Gaviria, C., et al. (2017). "If it matters, I can explain it": Social desirability of knowledge increases the illusion of explanatory depth. In *Paper presented at the Annual Conference of the Cognitive Science Society*.

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*, 1069–1086.

Hunt Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics, 95*(9–10), 1082–1095.

Johnson, D. R., Murphy, M. P., & Messer, R. M. (2016). Reflecting on explanatory ability: A mechanism for detecting gaps in causal knowledge. *Journal of Experimental Psychology: General, 145*, 573–588.

Meyers, E. A., Turpin, M. H., Białek, M., Fugelsang, J. A., & Koehler, D. J. (2020). Inducing feelings of ignorance makes people more receptive to expert (economist) opinion. *Judgment and Decision making, 15*(6), 909–925.

Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 35*, 817–829.

Rozenblit, L., & Keil, F. C. (2002). The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth. *Cognitive Science, 26*, 521–562.

Sharpe Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research, 44*(1), 211–230.

Sloman, S. A., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. New York: Riverhead Press.

Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.)*, Vol. 11. Advances in experimental social psychology* (pp. 289–338). New York: Academic Press.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320–324.

Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology, 48*(7), 955–969.

Vives, M. L., Cikara, M., & FeldmanHall, O. (2021). Following your group or your morals? The in-group promotes immoral behavior while the out-group buffers against it. *Social Psychological and Personality Science, 13*(1), 139–149.

Voelkel, J. G., Brandt, M. J., & Colombo, M. (2018). I know that I know nothing: Can puncturing the illusion of explanatory depth overcome the relationship between attitudinal dissimilarity and prejudice? *Comprehensive Results in Social Psychology, 3* (1), 56–78. https://doi.org/10.1080/23743603.2018.1464881

Zeveney, A., & Marsh, J. (2016). The illusion of explanatory depth in a misunderstood field: The IOED in mental disorders. In *Paper presented at the Annual Conference of the Cognitive Science Society*.