Group Processes & Intergroup Relations
GPIR

Article

The straw man effect: Partisan misrepresentation in natural language

Group Processes & Intergroup Relations
1–20

DOI: 10.1177/13684302211014582
journals.sagepub.com/home/gpi


Michael Yeomans

Abstract

Political discourse often seems divided not just by different preferences, but by entirely different representations of the debate. Are partisans able to accurately describe their opponents' position, or do they instead generate unrepresentative "straw man" arguments? In this research we examined an (incentivized) political imitation game by asking partisans on both sides of the U.S. health care debate to describe the most common arguments for and against ObamaCare. We used natural language-processing algorithms to benchmark the biases and blind spots of our participants. Overall, partisans showed a limited ability to simulate their opponents' perspective, or to distinguish genuine from imitation arguments. In general, imitations were less extreme than their genuine counterparts. Individual difference analyses suggest that political sophistication only improves the representations of one's own side but not of an opponent's side, exacerbating the straw man effect. Our findings suggest that false beliefs about partisan opponents may be pervasive.

Keywords

intergroup perception, natural language processing, perspective-taking, political psychology

Paper received 28 April 2020; revised version accepted 13 April 2021.

Introduction

> Condemn me if you will, but condemn
> me by other witnesses than Theodore Roosevelt.
> I was a man of straw; but I have been
> a man of straw long enough. Every man
> who has blood in his body, and who has been
> misrepresented as I have, is forced to fight.
>
> William Howard Taft, 1912

The modern political landscape is unmistakably and increasingly divided along partisan lines (Iyengar & Westwood, 2014; Pew Research Center, 2014). Of course, in any diverse society, the occasional disagreement is inevitable. But it is concerning when the very nature of that disagreement is misrepresented by the parties mired within it. President Taft's experience is shared among all sorts of partisans who have seen their true position reduced to a "weak, defenseless" straw man[1] (Safire, 2008). Why is it that people engage with straw men, rather than with their opponents' true

Harvard Business School, USA

Corresponding author:
Michael Yeomans, Harvard Business School, Harvard University, Boston, MA 02163, USA.
Email: myeomans@hbs.edu

positions? And what is it about straw man arguments that make them unsubstantial?

In its initial construction in philosophy, a "straw man argument" is a rhetorical tactic, deliberately designed to be unpersuasive by a debate opponent (Aikin & Casey, 2011; Bizer et al., 2009; Talisse & Aikin, 2006). But decades of psychological research have shown that the mechanisms of misrepresentation are often far less intentional (Epley & Waytz, 2009; Tajfel, 1982). It can be hard to understand how other people are thinking, so even sincere partisans may inadvertently attribute straw man arguments to their opponents. Here, we hypothesize that straw men are the result of partisans' limited ability to take one another's perspective. We tested our hypothesis by collecting natural language descriptions of partisan positions by people who supported and opposed them. We then quantified the text using machine learning, as well as human judges, to empirically estimate the ways in which opponents consistently misrepresented each other's views. Our analyses document both motivational and cognitive mechanisms that can contribute to the persistence of straw man arguments.

Motivation

Straw men are by no means the only false beliefs to be found in the political world. In theory, the modern world provides an unprecedented amount of information available to citizens about the world and about one another. In practice, many psychological and social forces lead partisans not only to consume and share different kinds of information, but even to be swayed by misinformation that bolsters their partisanship (Allcott & Gentzkow, 2017; Lazer et al., 2018; Scheufele & Krause, 2019). And while some suggest that these forces seem to be accelerating in the modern media environment (Vosoughi et al., 2018), their power is in exploiting the basic mental processes that leave partisans susceptible to misinformation (Pennycook & Rand, 2019). Thus, in an environment where partisan opponents can so often fall prey to false beliefs about the rest of the world, it is perhaps no surprise that they may develop false beliefs about one another.

However, we argue that among various false beliefs, straw man beliefs can have unique and important effects on partisan conflict. Precisely because they are false beliefs about other people, we suggest that they can be especially detrimental to intergroup relations. Effective communication between in-groups and out-groups is a necessary condition for successful interactions (Allport, 1954; Messick & Mackie, 1989; Tajfel, 1982). And the experience of feeling that others are listening is itself a positive experience in an otherwise conflicted environment (Bruneau & Saxe, 2012; Goldstein et al., 2014; Yeomans et al., 2020). Furthermore, even the perception of bias between partisans can lead to a spiral of ill-will and misunderstanding, exacerbated by the force of naive realism (Kennedy & Pronin, 2008). All kind of relationships could suffer when straw man arguments supplant actual understanding.

There are adverse consequences of straw man beliefs to the belief holder as well. Psychologists have shown that, in many domains, a "consider the opposite" strategy can improve judgments by bringing another perspective to mind (Babcock et al., 1997; Herzog & Hertwig, 2009; Lord et al., 1984). But if that opposite perspective is a feeble straw man instead, partisans may not be debiased at all, as John Stuart Mill wrote, "He who only knows his side of the case knows little of that, p. 35" (1869). Straw men may permit partisans to believe they have successfully counter-argued against their opponents, which could entrench their own position even further (McGuire, 1964; Tormala & Petty, 2002).

In this vein, the economist Bryan Caplan has proposed that straw man arguments are a valid signal that the arguer's own position is weak (2011). As a remedy, he proposes the Ideological Turing Test, whereby the strength of an argument can be evaluated by how well the arguer is able to imitate their opponents' true positions. And his suggested paradigm has been explored somewhat in the psychology literature, as two previous papers have collected data from similar imitation games (Dawes et al., 1972; Newman

et al., 2003). However, this previous work is limited in several ways, most notably, sample sizes were small and imitators were not incentivized to provide accurate answers. This meant that participants faced no cost for deliberately misrepresenting their opponents to satisfy other partisan goals in the Talisse and Aikin model (2006) of straw men. Without explicit incentives for accuracy, straw man arguments cannot be considered as a potential indicator of topic knowledge.

More commonly, the accuracy—and inaccuracy—of beliefs about opposing viewpoints has primarily been confined to predicting a person's response to a structured question, such as a Likert scale (e.g., Chambers et al., 2006; Robinson et al., 1995; van Boven et al., 2012) or a question about a numerical fact (e.g., Bullock et al., 2015; Prior et al., 2015). Structured responses are convenient for experimenters, but they can also have unintended consequences for the quality and interpretation of responses (Bauer et al., 2017; Krosnick, 1999). Conceptually, partisans and their opponents may have different construals of how the points on a Likert scale map onto positions in the real world. Furthermore, structured questions provide the range of possible answers to participants, making it easier for them to strategically adjust their answer to suit the goals of the experiment.

Instead, in this research, we elicit beliefs about opposing viewpoints as natural language descriptions. That is, we ask people to elaborate on what they believe to be the arguments that best articulate their opponents' position. By using open-ended responses, we allow partisans to speak for themselves, and put their perspective-taking skills to a sharper test by forcing them to represent their opponents in a high-dimensional space (i.e., text). There is also a pragmatic aspect to studying open-ended text, because most political behavior is encoded in natural language (Grimmer & Stewart, 2013; O'Connor et al., 2011). Language is the primary tool by which groups choose to communicate and by which they understand (and misunderstand) each other in the natural world.

Theoretical Background

There are many related psychological processes that might bring unconvincing straw men into being. We design our experiments to collect evidence for several competing theories. Our primary hypothesis concerns the role of motivation. We first identify the straw man effect as an honest but insufficient belief, rather than a strategic tactic. One such mechanism is that partisans might pursue in-group affiliation by misrepresenting their opponents as unflattering straw men (Brewer, 1999; Tajfel, 1982). More generally, partisans have been known to distort politically relevant facts to bolster their own preconceptions (Bullock et al., 2015; Kunda, 1990; Prior et al., 2015). And at an even more basic level, perspective-taking is effortful, and partisans may not exert that effort without the right incentives (Epley et al., 2004). This literature would suggest that providing incentives for accuracy would improve the relative accuracy of a partisan's stated beliefs about their opponent's views.

While it is easy to incentivize performance within an experiment, we do not incentivize preparation by asking participants to gather information before the experiment that might help them perform well. Instead, though, we can study the effects of information-gathering indirectly, by comparing individual differences that might moderate the main effect. For example, the literature on intergroup contact might predict that everyday relationships with partisan opponents might attenuate the straw man effect (Allport, 1954; Pettigrew, 1998; Pettigrew & Tropp, 2006). Though the underlying processes of outgroup contact are rich and diverse, here we test a novel consequence, which is that partisans who interact with their opponents regularly may also be better at describing their points of view.

Another important individual difference in this domain is political sophistication, and it is not obvious how that factor might affect partisan misrepresentation. On the one hand, political sophistication entails a larger knowledge base from which to draw, which would improve the representations,

all else equal. However, political sophistication has also been associated with partisan entrenchment (Brandt et al., 2015; Kahan et al., 2012; Palfrey & Poole, 1987; Sidanius, 1984; Taber & Lodge, 2006). In that case, political sophisticates may simply apply their knowledge to buttress their own positions, and the quality of their representations would not improve at all. This hypothesis was tested by collecting several measures of political sophistication from each writer.

Another individual difference we consider is that of ideology. There have been some recent papers suggesting that their ideological foundations naturally lead conservatives to be more close-minded, and be more susceptible to misinformation (Jost et al., 2018; Pennycook et al., 2020). However, other papers have found that some partisan misperceptions seem to be relatively stable across ideologies (e.g., Ditto et al., 2019; Frimer et al., 2017; Toner et al., 2013). This hypothesis was tested by measuring straw man beliefs on both sides of a relatively balanced issue that was mostly split along liberal/conservative lines, and comparing the two groups against one another.

Finally, we used text analysis to better understand the ways in which the imitations are insufficient. One hypothesis suggests that misunderstanding may result from differences in underlying values. That is, if partisans interpret their opponents' policy positions through their own ideological lens, they may generate arguments that do not reflect their opponents' actual priorities (e.g., Crawford et al., 2013; Frimer et al., 2014; Graham et al., 2009; Jost et al., 2008). That is, straw men might seem like good arguments to the writer—and bad arguments to the target—because they appeal to the writer's own values, even if they do not represent the target's actual values.

Another possible account of the straw man effect is that imitators describe a position that is simply more extreme than the position of their average opponent. Previous research has reliably shown this effect on a one-dimensional number scale. Imitators rate their opponents' position as further from their own position (and closer to the far endpoint of the scale) compared to how those opponents actually rate themselves (Chambers et al., 2006; Chambers & Melnyk, 2006; Graham et al., 2012; Robinson et al., 1995; Scherer et al., 2015; Sherman et al., 2003; van Boven et al., 2012; Westfall et al., 2015). However, this relies on the assumption that opposing partisans agree on how the number scale points map onto different versions of an argument. In fact, there is evidence that partisans may even have trouble judging the extremity of their own positions (Fernbach et al., 2013). Thus, it is an open question whether polarization projection is a meaningful contributor to the straw man effect in open-ended text.

Overview of Current Research

We report two studies of the straw man effect on an issue in U.S. politics that was timely, contentious, important, and which involved a variety of reasonable policy beliefs and priorities on both sides of the political spectrum—the Patient Protection and Affordable Care Act (aka "ObamaCare"; McCabe, 2016). Partisans across the spectrum were asked to describe "the reasons someone would give" to either support or oppose ObamaCare. These descriptions were then shown to judges, who were incentivized to guess the authors' true position based on the text. In Study 1, half the writers also received incentives for their ability to convince the judges they were genuine. In Study 2, we collected a larger pool of participants to form a training set for natural language-processing algorithms to perform this detection task. These models provided a performance benchmark for human judges, revealed the distinctive linguistic features of genuine and straw man arguments, and provided tests of the mechanism for the straw man effect. Overall, these results suggest that straw man arguments tend to be hollow and unsubstantial, rather than extreme and inflammatory.

Study 1

We conducted Study 1 in two parts. First, we collected text data from "writers," who described the

arguments made for and against ObamaCare, respectively. Afterwards, we showed those descriptions to "judges," who tried to guess the writers' actual positions based on their descriptions.

Study 1a Methods

We recruited participants in the autumn of 2014 to explain the "the reasons someone would give to explain why they are either in favor of, or opposed to, the Affordable Care Act" (full instructions can be found in Appendix A of the supplemental material). Nine hundred participants enlisted; however, 21 failed the attention check and 40 did not finish, thus, 839 writers were left for the full analyses. This was a within-subjects design, and all writers wrote descriptions for both positions (randomly ordered), producing 1,678 texts in total. All participants were asked to write their descriptions in the first-person form, and were given the following prompt: "This person would answer by writing. . ." All writers were told that their descriptions would be shown to other participants and to try their best to be accurate, as long as they did not look up any outside information. Writers also answered a number of questions about their demographics and political beliefs (see Table 1, and Appendix B in the supplemental material), including their true position on ObamaCare.

Incentive manipulation. Half the writers were randomly assigned to receive incentives based on both their descriptions. Specifically, writers could earn up to 600% of their base payment as a performance bonus, determined by the percentage of judges who later thought their texts were written by a genuine proponent of that position (e.g., if half of judges thought their texts were genuine, they would receive a 300% bonus). Participants had to write out these instructions verbatim, and were reminded of the incentives again as they wrote their descriptions. The other half of participants in the control condition were told about the judges, and to try as hard as they could, but did not receive any performance bonus.

Description cleaning. The collected texts were filtered using two a priori exclusion criteria before being shown to judges. First, we removed all writers whose stated position on ObamaCare was in the middle of the 1 to 7 scale, so that there was no ambiguity about the "ground truth" for each writer. This excluded 305 writers, leaving 1,068 texts from definitive partisans (i.e., 1, 2, 6, or 7).

With these remaining texts, two research assistants (blind to condition) independently flagged writers who clearly did not attempt to follow the instructions, with disagreements resolved through deliberation. Participants in the control condition were more likely to be flagged when they were describing a position they did not support (M = 14.9%, 95% CI [10.6%, 19.2%]) than when describing the position they did support (M = 4.2%, 95% CI [1.8%, 6.6%]); $\chi^2(1) = 16.12$, $p < .001$. Reassuringly, when incentives were offered, writers were not significantly more likely to be flagged for their imitations (M = 6.3%, 95% CI [3.4%, 9.1%]) than for their genuine responses (M = 4.4%, 95% CI [2.0%, 6.9%]); $\chi^2(1) = 0.6$, $p = .445$. In total, 79 such responses were flagged and removed, and in cases where only one of a writer's responses was flagged, their nonflagged response was also removed. This left 464 writers and 928 texts in the data set. The research assistants also cleaned the text by correcting any third-person to first-person form, as well as spellchecking.

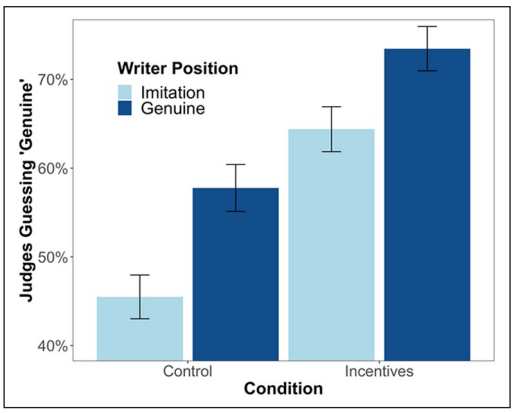
Study 1b Methods

Nine hundred forty-one participants were recruited to be judges in the autumn of 2014; 72 participants failed the attention check and another 16 did not finish the task, leaving 853 participants in the full analysis. They first read the exact instructions the writers were given, and then judged 24 different texts—a block of 12 that described supporters, and a block of 12 that described opponents. For each text, they made a binary choice ("Was this writer actually an ObamaCare supporter or an ObamaCare opponent?") and expressed their confidence in that choice on a scale from 50 (*pure guess*) to 100 (*absolutely certain*). All judges were incentivized as well:

each judge was given a bonus of up to 200% of their base pay, depending on how accurate they were (e.g., if they were correct on half of their judgments, their bonus was 100% of their base pay).

Judges were assigned to texts randomly, with six of each appearing in every block (shuffling together incentivized and control writers). Our design ensured that every text was labeled by at least seven supporter judges and seven opponent judges, with an average of 9.8 judges per supporter text and 12.6 judges per opponent text (opponents were oversampled because a slight majority of writers were ObamaCare supporters). Finally, judges all completed the same demographic questions as the writers (see Appendix B in the supplemental material), with an additional survey about news media consumption (see Appendix C in the supplemental material).

Figure 1. Average guesses of judges across all descriptions in study 1, divided by writer position and writer condition.



Note. Error bars indicate 95% confidence intervals for the group mean.

Study 1 Results

Our analyses (in this study and throughout) collapse across block order, which did not affect the main results. Additionally, the results are collapsed across target positions (i.e., support vs. oppose), so that all descriptions were labeled as either "genuine" or an "imitation."

The effects of writer position and writer incentives on the judges' guesses are plotted in Figure 1. These results confirm that judges had some insight into the writers' true identity—genuine writers were more likely to be labeled as genuine (M = 66.1%, 95% CI [64.2%, 68.1%]) than imitators (M = 55.6%, 95% CI [53.6%, 57.5%]); $t(926) = 7.61$, $p < .001$. This difference in believability is empirical confirmation of the straw man effect in text—partisans failed to adequately simulate their opponents' arguments.

The effect of incentives is also clear: writers who were incentivized to write genuine-sounding descriptions were more convincing (M = 68.9%, 95% CI [67.1%, 70.8%]) than those with no incentives (M = 51.6%, 95% CI [49.7%, 53.5%]); $t(926) = 12.9$, $p < .001$. However, we did not find a significant interaction between writer position and writer incentives in a regression, $t(924) = 1.2$, $p = .217$. Furthermore, the simple effect of writer position was still significant even among all incentivized writers, $t(492) = 5.0$, $p < .001$. Incentives had, at best, a modest impact on the size of the straw man effect, and the effect clearly persisted even when writers had good reason to do better.

If the judges' labels are taken at face value, their average accuracy was 55.3%, better than chance (50%); one-sample t test: $t(20471) = 15.2$, $p < .001$, but also much lower than their average confidence (M = 77.3%); two-sample t test: $t(20471) = 696$, $p < .001$. Judges were also more accurate for nonincentivized writers (M = 56.2%, 95% CI [55.2%, 57.2%]) than for incentivized writers (M = 54.5%, 95% CI [53.6%, 55.4%]); $\chi^2(1) = 5.71$, $p = .017$. We also calculated wisdom-of-crowds averages, using two different strategies to combine all judgments for each text into a single combined judgment. First, we tried a simple "majority vote" aggregation rule, where each judgment for a text was given equal weight in an average, and the final prediction was determined by whether this average was above or below the median of all texts for this position. These simple averages produced somewhat more

accurate predictions than the individual judges (M = 58.2%, 95% CI [55.0%, 61.4%]); $\chi^2(1) = 31.9$, $p < .001$. We also tried a more sophisticated aggregation rule that weighted judges based on their expressed confidence, again using the median split threshold to generate a single prediction from each average. These weighted judgments were slightly more accurate (M = 59.1%, 95% CI [55.9%, 62.2%]). However, even these confidence-weighted crowd judgments were highly correlated with the word count of the responses, $r = .659$, $t(926) = 26.7$, $p < .001$. This suggests that much of the human judges' accuracy lay in their ability to evaluate the writer's effort level.

Finally, we wanted to test for a directional difference in the accuracy of partisan imitations. Here, we used the confidence-weighted crowd judgments, the most accurate approach tested so far, to evaluate the relative accuracy of the tests written. Our analysis showed that supporters (M = 55.3%, 95% CI [52.9%, 57.7%]) and opponents (M = 54.6%, 95% CI [51.3%, 57.9%]) wrote imitations that were rated as equally genuine; two-sample t test: $t(462) = 0.3$, $p = .732$. The same pattern held when comparing their genuine descriptions (supporters: M = 65.6%, 95% CI [63.0%, 68.2%]; opponents: M = 65.9%, 95% CI [62.9%, 69.0%]); two-sample t test: $t(462) = 0.2$, $p = .851$. These results suggest that, in this setting, the straw man effect was equally present on both sides of the debate.

Study 1 Discussion

This study demonstrated the straw man effect using explicit and objective criteria. Descriptions of opponents' positions were readily distinguished from descriptions by genuine position holders by outside observers. We found that incentives improved the accuracy of positions written by opponents. However, we also found that incentives had close to the same effect on genuine description writers. This result is difficult to explain with a purely motivational account of the straw man effect. If partisans really construed this task as an opportunity to cheerlead for their side, we would have expected them to be just as emphatic for their own side even without incentives. Incentives did level the rate at which supporting and opposing writers were flagged as insincere by research assistants. However, among the remaining descriptions, partisan misrepresentation by opponents was still common. Even when partisans tried to accurately represent their opponents, in many cases they still failed to do so.

The human judges were able to modestly detect some signal from the written texts, in line with previous research on deception detection (Bond & DePaulo, 2006). But it is hard to interpret their performance because the judges' ability to detect imitations is directly opposed to the writers' ability to generate imitations. Imagine, for example, that new samples were collected, and the judges were unable to discriminate between imitation and genuine texts. This result would be expected if the new writers were much better imitators than the originals, or if the new judges were much worse detectors than the originals. By that same logic, it is hard to evaluate the accuracy of the writers' imitations, because we do not know the accuracy of the judges. To estimate the writers' performance, we needed an objective empirical benchmark of imitation fidelity that could determine the ceiling for how well the judges could have done. In the next study, we introduce new empirical approaches to resolve this dilemma.

Study 2

While Study 1 relied on human judges to judge the quality of imitations, in this study we present results from a different type of analysis: natural language processing (Hirschberg & Manning, 2015; Jurafsky & Martin, 2017). Here we use two kinds of language models. More traditionally, we apply dictionaries that tally classes of words to form conceptual features that have domain-general scoring rules. Additionally, we apply machine learning algorithms to learn distinctive features directly from the text and produce a domain-specific scoring rule.

These language models complement our results from human judges in three ways. First, they provide a benchmark for the judges' accuracy—were the imitations really that good, or were judges missing important features in the text? Second, this benchmark can be compared to covariates as a measure of individual differences in imitation skill. Third, language models are more interpretable than human judges. While both humans and models give holistic scores, models also identify which features of the text are most useful for producing accurate scores.

Study 2 Methods

Sample. We added to the pool of description writers in Study 1 by recruiting new participants in the autumn of 2014, in the lead-up to the midterm elections. For all Study 2 analyses, we combined these new writers with the 834 texts written by the 417 incentivized writers from Study 1 (discarding the texts from the 422 unincentivized writers). The procedure was exactly the same as in Study 1, with two exceptions: all writers were incentivized (instead of only half) and a survey about news media consumption (see Appendix C in the supplemental material) was added to the set of demographic questions (see Appendix B in the supplemental material). This similarity allowed us to pool the data into a larger sample for text analysis.

One thousand five hundred and sixty-five participants completed the new writing study. As in Study 1, research assistants read the texts and flagged writers who did not attempt to follow the instructions (and corrected spelling, and third-person to first-person form). They flagged 119 descriptions and, after removing both texts from those writers, this left 1,459 new writers for a total of 2,918 new texts. When the writers from Study 1 were added, this created an analysis data set for Study 2 containing 3,700 texts from 1,850 incentivized, unflagged writers. This sample included 402 people who were truly neutral on the issue (i.e., 4 on a 7-point scale), so we excluded them from all analyses. However, in a change from Study 1, we did include moderate writers (i.e., 3 or 5 on a 7-point scale) for our analyses of the text. By including these moderates, we increased our sample size even further and also had more variation in the writers' position extremity, which factored into our following analyses.

The writers' incentive scheme in the new study was identical to that in Study 1a. We determined the incentives using a sample of 1,616 human judges to rate the descriptions in Study 2. Their protocol was identical to that of Study 1b, including the judges' incentives. Though we do not discuss the judges in detail here, we report their aggregate results as a benchmark.

Individual differences. Domain knowledge was measured in two ways. First, participants self-reported their subjective knowledge about the politics of ObamaCare. Second, participants completed a short factual quiz to evaluate their objective knowledge. Participants also reported their education (highest degree completed) and their partisanship (official party registration). In Study 2 ($N = 1{,}459$) they also reported their regular news media sources, which were compiled into two measures: total news media, as the total number of news sources, and news media bias, calculated based on the balance of left- and right-leaning sources (according to the Pew Research Center [2014]).

Participants self-reported their outgroup social contact as the percentage of ObamaCare supporters in their social circle (family, friends, coworkers, neighbors, etc.). Where possible, participants' IP address was taken from their survey response ($N = 1{,}627$) and matched to the Partisan Voting Index (PVI) of their congressional district (Cook & Wasserman, 2014). These three covariates—media bias, social contact, and PVI—were all reverse-scored in half the sample so that higher values always implied more exposure to outgroup members. Finally, we also measured participants' position extremity by calculating the distance of their position from the middle of the position scale, from 1 (i.e., 3 or 5 on the scale) to 3 (i.e., 1 or 7 on the scale).

Domain-general language models. The dictionary set most familiar to psychologists is the *Linguistic Inquiry and Word Count* (LIWC; Pennebaker et al.,

2007), which uses predefined dictionaries of common words to parse the text into 64 different features that represent basic word categories (e.g., self-references, negations, food words) that can themselves be combined to define linguistic markers of specific psychological constructs.

The most conceptually relevant set of dictionaries we tested was a recently updated version of the *Moral Foundations Dictionary* (Frimer et al., 2019; see also Graham et al., 2009). This work offers 10 categories of morally charged words that are theorized to reflect the values and intuitions that differ between liberals and conservatives. We also tested composite features intended to capture deception (Newman et al., 2003) and integrative complexity (Slatcher et al., 2007; Tetlock, 1983). Finally, the LIWC counts positive and negative emotion words, and they can be added to calculate the average emotionality per word, while the difference of the two is a good measure for sentiment analysis (e.g., Waytz et al., 2014).

Domain-specific language models. We also applied basic natural language-processing tools to empirically learn a scoring rule for the texts (Hirschberg & Manning, 2015; Jurafsky & Martin, 2017). First, we extracted common 1- to 3-word phrases ("ngrams") from the data (Benoit et al., 2018). Punctuation, upper case, numbers, and stop words were removed. All but a few politically relevant words (e.g., "conservative," "healthcare") were stemmed (Porter, 1980), and the rarest ngrams (i.e., appearing in less than 1% of the descriptions) were dropped.

In addition to the ngram features, we also compared the results of a similar model that uses a feature set consisting of all the dictionary features. We also tested a feature set generated from word vectors, which uses large corpora to learn a low-dimensional representation of meaning in natural language (Mikolov et al., 2017). Finally, we included a model that uses a feature set combining all the ngrams, word vectors, and dictionaries together. All models also included the word count.

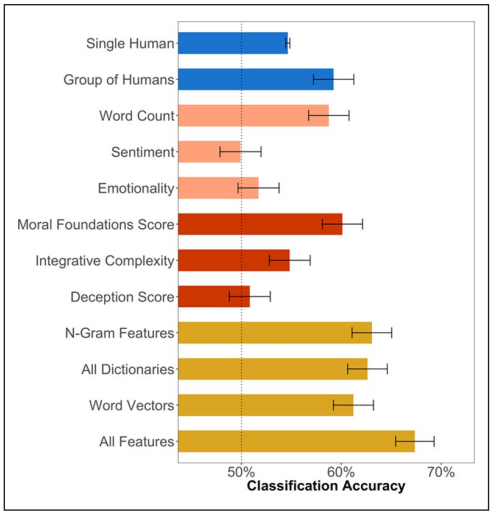
The descriptions were separated by target position (support vs. oppose) and, in each half, these features were fed into a simple machine learning model. The dependent variable of the model was the binary variable indicating that the description was genuine or an imitation. The model we chose was a logistic LASSO regression (Hastie et al., 2009; Tibshirani, 1996), which chooses a prediction model empirically from among many models using the same feature set by comparing them all across an inner 20-fold cross-validation loop. To estimate out-of-sample prediction accuracy, we used a second 20-fold nested outer cross-validation loop (Stone, 1974; Varma & Simon, 2006). Assignment to outer folds was stratified to balance the numbers of imitations in each fold. The output of the model for each new text was continuous (from 0 to 1); the predicted probability that each new text was from the genuine group as well as all predictions were averaged over five randomly seeded runs of the entire procedure to produce more consistent predictions.

Study 2 Results

Classification accuracy. Accuracy was calculated as the average rate at which the predicted label matched the writer's true position (binary). Of course, the language models' predictions were continuous (as were the averages of humans). Furthermore, our sample was slightly imbalanced, with slightly more pro-ObamaCare participants than anti-ObamaCare ones, so we did not want to use the same threshold for each position to binarize the continuous predictions. Instead, we generated binary predictions separately for each position by splitting the distribution to match the base rate of genuine writers in the sample from that position. Additionally, all of the results we report in what follows are unweighted averages, assigning equal weight to each observation from both positions (further, reweighing observations to equilibrate the influence of the two positions does not substantively change our results).

The top-line accuracy of the different language models is plotted in Figure 2. For comparison with the human judges, this figure counts accuracy only among nonmoderates (i.e., 1, 2, 6, or 7 on the position scale), although the comparisons across

Figure 2. Performance of human judges and language-processing models in Study 2.



Note. Accuracy for the three machine learning models was estimated using nested cross-validation. The X axis represents binary classification accuracy (error bars indicate 95% confidence intervals).

language models are substantively similar if we include moderates (i.e., 3 or 5 on the scale) in the sample as well. The word count of the description proved to be an effective rule of thumb (M = 58.8%, 95% CI [57.0%, 60.6%]), though this does not imply that imitators would do better by simply writing more, if they had nothing left to say. Individual human judges (M = 55.2%, 95% CI [54.8%, 55.6%]) could not match the length rule, though equal-weighted averages of human judges did about as well (M = 59.2%, 95% CI [57.4%, 61.1%]), as did confidence-weighted averages (M = 60.1%, 95% CI [58.3%, 61.9%]).

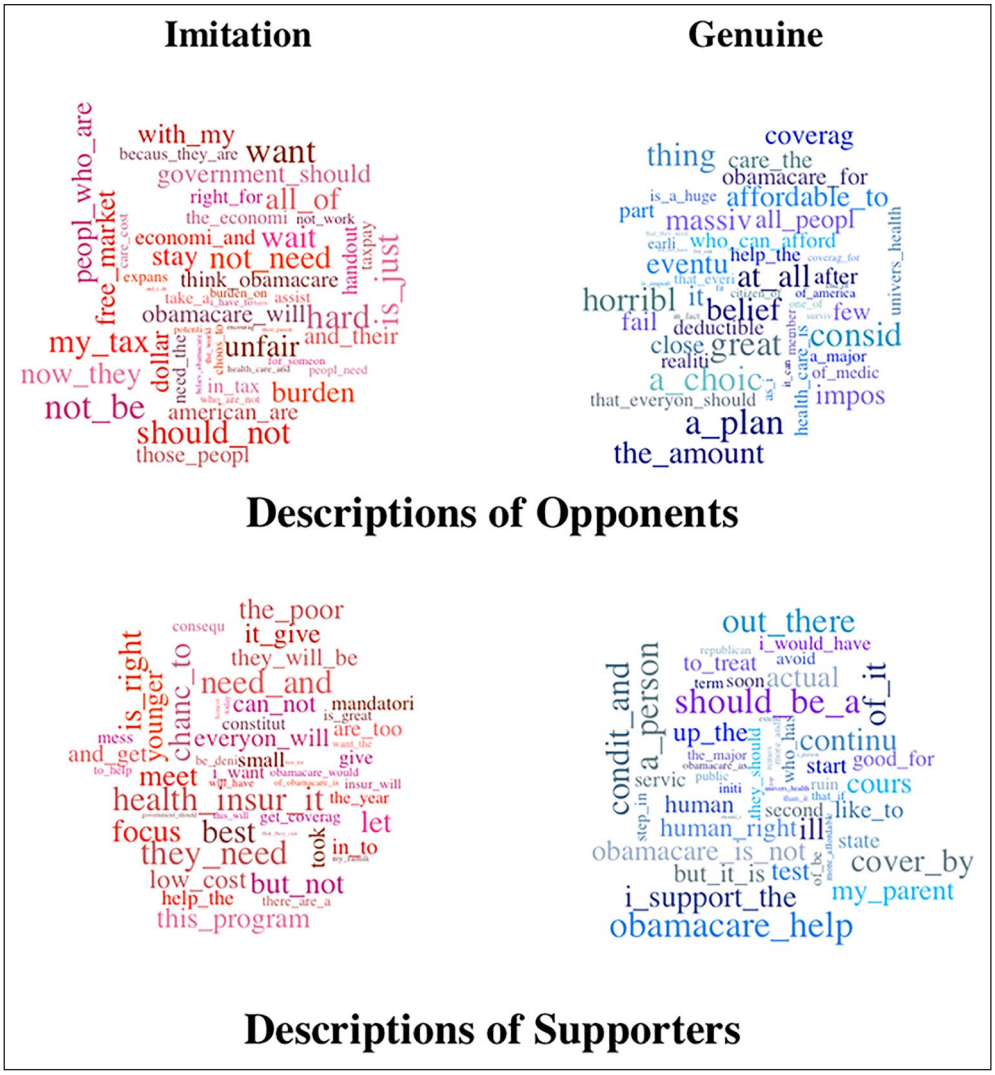
Using these confidence-weighted averages, we also replicated the findings regarding the directional effects of accuracy from Study 1. That is, we found that imitations from supporters were rated as genuine (M = 60.2%, 95% CI [58.8%, 61.6%]) at about the same rate as those from opponents (M = 60.4%, 95% CI [58.6%, 62.1%]); $t(1382) = 0.0$, $p = .997$. And we also found no differences in the ratings for genuine descriptions from supporters (M = 69.3%, 95% CI [67.8%, 70.8%]) and opponents (M = 68.5%, 95% CI [66.9%, 70.1%]); two-sample t test: $t(1382) = 0.7$, $p = .506$.

Language model accuracy. Some of the individual dictionary models performed well. The *Moral Foundations Dictionary* (Frimer et al., 2019) performed quite well, as people who genuinely believed in their position used more words from the corresponding moral intuitions dictionaries (M = 60.1%, 95% CI [58.1%, 62.1%]). We present the scores for all five dictionaries separately in Appendix D of the supplemental material for each task, and found the results are primarily driven by the balance of care-related words. Integrative complexity was also higher for genuine writers than for imitators (M = 54.8%, 95% CI [52.8%, 56.9%]; Slatcher et al., 2007). However, several scales did not have any correlation with the writers' positions. Average emotionality (M = 51.7%, 95% CI [49.7%, 53.8%]), was not a significant predictor of writer position, nor was sentiment (M = 49.9%, 95% CI [47.9%, 52.0%]). The LIWC deception construct also did not differentiate genuine and imitation partisans (M = 50.8%, 95% CI [48.8%, 52.9%]; Newman et al., 2003).

None of the domain-general models could match the performance of the domain-specific machine learning models, however. The model using only ngram features performed about as well (M = 63.1%, 95% CI [61.1%, 65.0%]) as the model that combined all the dictionary scales (M = 62.6%, 95% CI [60.6%, 64.6%]) and the model that used word vector features (M = 61.2%, 95% CI [59.2%, 63.2%]). However, the most accurate model combined all the ngram and dictionary features in a supervised learning model (M = 67.4%, 95% CI [65.4%, 69.3%]). We use predictions from the combined features as our model of description quality in what follows.

Model content. In addition to accuracy estimates, the natural language descriptions offer a rich representation of how the contents of imitation writers' beliefs differ from the beliefs of genuine writers. To focus on interpretability, we trained a new classification algorithm with only one cross-validation loop and using only the ngram features.

Figure 3. Word clouds indicating the 50 most influential (i.e., distinctive and common) words in each direction for each description type, as determined by a LASSO algorithm.



By multiplying the coefficients of the LASSO algorithm by the frequency of each ngram in the model, we got a rough sense of which words and phrases had the largest influence on the algorithm's decisions (i.e., were common and distinctive). We then used that influence metric to select the top 50 most influential ngrams for predicting writers' true positions, in both directions and for each of the two description tasks. On Figure 3, we group each of these ngram lists into four distinct word clouds, with font size proportional to the influence of the ngram.

Individual differences. Having developed a model of description quality that is more valid than human raters, we now wanted to compare this metric to different author characteristics. To do this, we conducted a series of standardized

Table 1. Demographics of participants in the current research

Population	*n*	Suports ObamaCare	Party member	Age	Male	Bachelor's degree
Study 1: Writers	839	58.8%	36.9%	34.7 (11.2)	49.5%	60.2%
Study 1: Raters	853	57.3%	46.5%	35.5 (11.8)	49.4%	54.6%
Study 2: Writers	1,565	57.2%	37.3%	34.7 (11.7)	43.5%	59.2%
Study 2: Raters	1,616	58.9%	43.6%	33.5 (11.1)	48.9%	54.3%
2012 U.S. voters	**2 x 108**	**40–45%**	**63%**	**55.7 (19.6)**	**46.6%**	**40.7%**

Note. All were recruited online through Amazon's Mechanical Turk for a "U.S. political survey." Standard deviations are in parentheses, where applicable. The bottom row shows voter statistics culled from census.gov and from Real Clear Politics.

Table 2. Demographic covariates of the straw man effect, as judged by the language processing model in Study 2

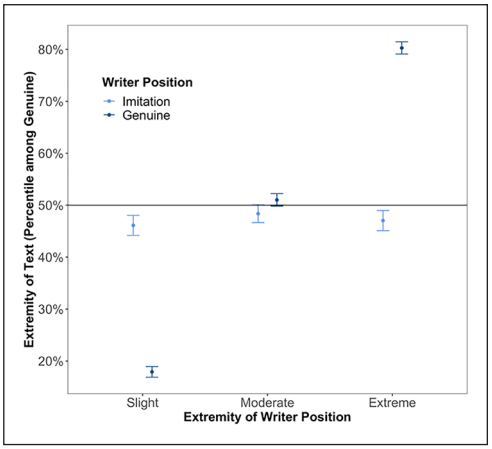
Covariate	Genuine writers	Imitator writers	Position interaction	Extremity control
Subjective knowledge	**0.048 (0.024)***	−0.027 (0.024)	**−0.075 (0.034)***	**−0.081 (0.035)***
Objective knowledge	**0.087 (0.023)*****	−0.016 (0.023)	**−0.104 (0.033)****	**−0.105 (0.033)*****
Highest education	**0.094 (0.023)*****	−0.026 (0.023)	**−0.119 (0.033)*****	**−0.12 (0.033)*****
News media total	**0.048 (0.026)**	0.003 (0.026)	−0.045 (0.037)	−0.046 (0.037)
News media bias	0.004 (0.026)	0.01 (0.026)	0.006 (0.037)	0.004 (0.037)
Outgroup social contact	**−0.053 (0.024)***	−0.017 (0.024)	0.035 (0.034)	0.041 (0.036)
Outgroup PVI	**0.052 (0.024)***	−0.015 (0.024)	**−0.067 (0.034)***	**−0.068 (0.034)***
Party registration	**0.055 (0.023)***	**−0.114 (0.023)*****	**−0.17 (0.032)*****	**−0.179 (0.033)*****
Position extremity	0.026 (0.029)	0.028 (0.029)	0.002 (0.041)	–

Note. Each covariate was tested in a multiple standardized regression to estimate the simple effect on the quality of genuine descriptions (Column 1), imitating descriptions (Column 2), and the interaction with the writer's position (Column 3) to test whether the effects in the first two columns are significantly different from one another. We also report this same interaction test from a model that controls for position extremity, to isolate the role of polarization from demographics that might be correlated (Column 4). All cells report the standardized regression coefficient from the model (standard errors in parentheses) and boldface indicates results significant at the p = .05 level.
PVI = Partisan Voting Index.
*$p < .05$. **$p < .01$. ***$p < .001$.

regressions. In each regression, the outcome was the model's predicted probability that the writer was genuine, and the predictors were the covariate of interest, the writer's position (genuine/imitator), and the interaction between the two. The full covariate analyses are reported in Table 2.

Our primary measure of political sophistication was participants' responses on a factual quiz. Participants who scored high on the quiz were better able to describe their own genuine position, $r = .093$, $t(1647) = 3.8$, $p < .001$. But, surprisingly, this factor had almost no effect on the quality of their imitations, $r = -.017$, $t(1647) = 0.7$, $p = .485$, and in a multiple regression, the interaction term confirmed the two effects are significantly different from one another, $\beta = .104$, $SE = 0.033$, $t(3294) = 3.2$, $p = .002$. This same interaction holds if sophistication is defined by self-reported domain knowledge, $\beta = .075$, $SE = 0.034$, $t(3294) = 2.2$, $p = .029$, or by proxy in relation to their educational level, $\beta = .119$, $SE = 0.033$, $t(3294) = 3.6$, $p < .001$. These results are also robust when we add controls for position extremity in the regression model (both the simple effect and the interaction with writer position). In these models, the original interaction terms were still significant, which indicates that these results are identifying an effect of sophistication that is independent of polarization.

Figure 4. Predicted position extremity for imitation and genuine writers from Study 2, as labeled by the polarization detection model.



Note. The Y axis represents average extremity, expressed as each text's percentile rank within the distribution of genuine writers (error bars indicate 95% confidence intervals).

Polarization projection. One plausible account of the straw man effect is that imitators describe a position that is more extreme than the position of their average opponent. Previous research has reliably shown similar effects on a one-dimensional number scale; however, measuring extremity with a number scale relies on the assumption that opposing partisans agree on what are the moderate and the extreme versions of each argument. Furthermore, Table 2 suggests that extreme writers were more successful at writing genuine descriptions than moderates. To resolve this apparent discrepancy, we addressed this question directly by training a supervised model to detect position extremity.

Using exactly the same text processing tool and LASSO algorithm as in Study 2, we trained a model to learn the differences—among genuine writers—between extreme and moderate supporters (imitations were held out entirely). The model's predictions of extremity within the genuine writers correlated well with their actual polarization, $r = .320$, $t(1647) = 13.7$, $p < .001$, which provided internal validation for the model.

Finally, we applied the polarization detector to the imitations from Study 2. Every imitation was assigned a polarization score, and these polarization scores were entered into a regression as the dependent variable (standardized and adjusted for the difference in means between tasks), with the writer's position (binary: support/oppose) as the independent variable. The average polarization scores for imitations and genuine descriptions are plotted in Figure 4. Overall, our analysis estimated that imitator descriptions were actually less extreme ($M = -0.006$, 95% CI [−0.010, −0.003]), on average, than genuine descriptions ($M = 0.006$, 95% CI [0.002, 0.011]); $t(3296) = 4.4$, $p < .001$. That is, their language was more similar to that of moderates than that of extremists. We also tested for polarization projection and found no significant effect of own extremity on imitation extremity, $\beta = .003$, $SE = 0.002$, $t(3294) = 1.5$, $p = .139$. These results suggest that in open-ended text and in this context, misrepresentation was not primarily driven by exaggerated extremity of opponents' positions.

General Discussion

The evidence presented here confirms that, in open-ended text, partisans did not represent their opponents' arguments well. However, in contrast to the colloquial understanding of straw man arguments, our results suggest that this kind of partisan disagreement is often not a deliberate tactic. The straw man effect was robust to incentives for accuracy, implying that partisans were often unable—not just unwilling—to take their opponents' perspective. Furthermore, our results suggest that partisans are not particularly accurate at detecting imitations of genuine positions either. Instead, we found that machine learning algorithms could detect imitations with substantially higher accuracy than human judges.

Theoretical Implications

While other recent research has shown that incentives can reduce or even extinguish partisan

gaps on factual questions (Bullock et al., 2015; Prior et al., 2015), these results showed a small reduction. However, that earlier work relied on questions that asked participants to guess the correct number on a scale (e.g., percentage of GDP growth under Obama). In these cases, the question itself provides the range of possible answers, and it is easy for partisans to deliberately adjust their answer to suit their goals. However, in open-ended tasks, the range of possible responses is very wide, and it is more difficult for partisans to intuit the correct adjustment from the question. To be sure, incentives do not guarantee that the responses were valid measures of partisans' actual mental representations of their opponents. For example, incentives may induce them to recite familiar stereotypes rather than their true beliefs about their opponents. But these stereotypes might still serve as signal of meta-knowledge about the contours of a debate. Regardless, incentives are necessary to distinguish the mechanism here from the Talisse and Aikin model (2006) of straw men as deliberate distortions for partisan gain. And our results suggest that at least some straw man arguments persist even when partisans have sincere intentions.

Our results also suggest that the straw man effect is exacerbated by political sophistication. Participants with more political knowledge wrote better descriptions of their own position, but that knowledge was of no help when describing their opponents'. The natural language-processing models also suggested that the language of imitators was more similar to that of genuine moderates than to that of genuine extremists, even though previous research has shown that partisans believe that their opponents hold extreme positions. This perhaps is related to why participants were overconfident in their ability to distinguish imitations from genuine arguments.

These results suggest that perceived polarization might be a natural consequence of asymmetric expertise, whereby partisans gather evidence to buttress their own preferred conclusions. Also known as the rationalizing voter theory (Lodge & Taber, 2013), this is supported by mechanisms at many cognitive levels (e.g., Frenda et al., 2012; Kahan, 2015; Lord et al., 1979; Robinson et al., 1995; Toner et al., 2013), and is a compelling explanation for why, in this research, partisans who could so faithfully defend their own position were at a loss when asked to describe their opponents' point of view. This lack of insight is typical in social judgment (Dunning et al., 2003; Nisbett & Wilson, 1977; Pronin et al., 2002), but it poses particular difficulties for intergroup research because the very processes that divide partisans may also distort their construal of others' positions.

Limitations and Future Research

In this research, we did not find an intervention to reduce partisan misrepresentation. Study 1 showed that incentives were, at best, a weak moderator of the straw man effect. However, this was a short-term intervention and could not be effective if partisans simply lacked the knowledge base to accurately take their opponents' perspective. To make an analogy to memory, our incentives could plausibly affect participants' biases in recall, but it would have been too late to make any impact on their biases during encoding. This suggests some skepticism is warranted for the potential of other short-term interventions (though see Saguy & Kteily, 2011; Stern & Kleiman, 2015). In addition, any potential encoding biases may not be eliminated by self-directed information search, since our results suggested that politically sophisticated partisans were no more accurate than political naives in their imitations (Keltner & Robinson, 1993; Lord et al., 1984; Thompson & Hastie, 1990). We also found essentially no effect of intergroup contact on accuracy for opponents, which agrees with a recent review suggesting that the effects of contact are more varied and context-dependent that is often acknowledged (Paluck et al., 2018). Our analysis of the extremity of partisans' imitations may even provide a mechanism for a potential backfire effect of

intergroup contact (similar to Bail et al., 2018). Specifically, partisans' imitations tended to be more moderate than the actual positions of their opponents—perhaps if they learned how extreme their opponents' positions tended to be, this would have other negative consequences for intergroup harmony and understanding.

Another limitation in this research is our focus on a single topic. This focus facilitated a rich language model, but it is important to consider whether the topic of debate might have moderated our results. We focused on a high-stakes political topic that featured a wide range of competing evidence as well as genuine differences in preferences and values, so that many texts could be collected from enthusiastic partisans on both sides. However, some topics face a clearer divide where the preponderance of evidence stands against one side—for example, antivaccination debates, global warming denialism, or other conspiracy theories (Hornsey et al., 2018; Rutjens et al., 2018; Stoknes, 2015). In these cases, it is possible that the straw man effect would be asymmetrical, as beliefs based on false premises may also disproportionately reinforce themselves with false beliefs about the opposing arguments. Additionally, many topics of disagreement engage less partisan vigor, where personal preferences are more respected. In cases where partisans are not trying to win a public debate, people may be more genuinely curious about one another, lessening the effect of the rationalizing voter mechanism. Future work could pack these mechanisms across a range of topics and domains.

The current research also demonstrates how machine learning can be applied to develop psychological theory. Initially, it was difficult to interpret the judges' low accuracy in Study 1—were partisans bad at being judges, or good at being imitators? The results of Study 2 indicated the former was true, as the language model made it clear that human judges had room for improvement. This result confirmed our central hypothesis that the writers, too, were often failing at their task of recreating their opponents' perspectives. Similar methods may be useful in other interpersonal judgment tasks—mind perception is hard, and inaccuracy is ubiquitous (Epley & Waytz, 2009). But perspective-taking is often studied in cases where the perspective taker has all the needed information available. In the world outside of the lab, however, social interactions are filled with ambiguity, and the blame for inaccuracy is perhaps better apportioned across both the mind perceiver and the mind perceived. To distinguish the two, researchers must estimate how much information is actually available, and our research demonstrates an empirical framework for that process. The accuracy of social perceptions is an oft-debated topic (e.g., Judd & Park, 1993; Jussim & Zanna, 2005; Zaki & Ochsner, 2011). Our research demonstrates how natural language can be used to study interpersonal accuracy in other domains where data are rich, but misunderstanding is common.

Conclusion

This research sought to understand how the basic processes of political stereotyping applied to unstructured text—could partisans articulate the reasoning of people who disagree with them? The results suggest that partisans not only do not know what the other side is thinking, but that they lack the ability to judge why their imitations are insufficient. Partisans' misunderstanding was exposed in unstructured responses, but that understanding was modelled by language-processing algorithms that distilled the linguistic and demographic profiles of the straw man effect. The results from these three studies support our hypothesis that the straw man effect is born not out of cynicism, but out of ignorance.

Data availability

For each study, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. The exact data and code from each study are available as online supporting information at https://bit.ly/2ZnMoBq

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Michael Yeomans https://orcid.org/0000-0001-5651-5087

Supplemental material

Supplemental material for this article is available online.

Note

1. Though this idiom is unfortunately gendered, we retain the original wording for clarity.

References

Aikin, S. F., & Casey, J. (2011). Straw men, weak men, and hollow men. *Argumentation*, *25*, 87–105. https://doi.org/10.1007/s10503-010-9199-y

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–236. https://doi.org/10.1257/jep.31.2.211

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Babcock, L., Loewenstein, G., & Issacharoff, S. (1997). Creating convergence: Debiasing biased litigants. *Law & Social Inquiry*, *22*, 913–925. https://doi.org/10.1111/j.1747-4469.1997.tb01092.x

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Fallin Hunzaker, M., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the USA*, *115*, 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left–right scale a valid measure of ideology? *Political Behavior*, *39*, 553–583. https://doi.org/10.1007/s11109-016-9368-2

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*, Article 774. https://doi.org/10.21105/joss.00774

Bizer, G. Y., Kozak, S. M., & Holterman, L. A. (2009). The persuasiveness of the straw man rhetorical technique. *Social Influence*, *4*, 216–230. https://doi.org/10.1080/15534510802598152

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Brandt, M. J., Evans, A. M., & Crawford, J. T. (2015). The unthinking or confident extremist? Political extremists are more likely than moderates to reject experimenter-generated anchors. *Psychological Science*, *26*, 189–202. https://doi.org/10.1177/0956797614559730

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*, 429–444. https://doi.org/10.1111/0022-4537.00126

Bruneau, E. G., & Saxe, R. (2012). The power of being heard: The benefits of "perspective-giving" in the context of intergroup conflict. *Journal of Experimental Social Psychology*, *48*, 855–866. https://doi.org/10.1016/j.jesp.2012.02.017

Bullock, J. G., Gerber, A. S., Hill, S. J., & Huber, G. A. (2015). Partisan bias in factual beliefs about politics. *Quarterly Journal of Political Science*, *10*, 519–578. https://doi.org/10.1561/100.00014074

Caplan, B. (2011). *The Ideological Turing Test*. https://www.econlib.org/archives/2011/06/the_ideological.html

Chambers, J. R., Baron, R. S., & Inman, M. L. (2006). Misperceptions in intergroup conflict: Disagreeing about what we disagree about. *Psychological Science*, *17*, 38–45. https://doi.org/10.1111/j.1467-9280.2005.01662

Chambers, J. R., & Melnyk, D. (2006). Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Personality and Social Psychology Bulletin*, *32*, 1295–1311. https://doi.org/10.1177/0146167206289979

Cook, C. E., & Wasserman, D. (2014). Recalibrating ratings for a new normal. *PS: Political Science & Politics*, *47*, 304–308. https://doi.org/10.1017/S1049096514000079

Crawford, J. T., Modri, S. A., & Motyl, M. (2013). Bleeding-heart liberals and hard-hearted conservatives: Subtle political dehumanization through differential attributions of human nature and human uniqueness traits. *Journal of Social and Political Psychology*, *1*, 86–104. https://doi.org/10.5964/jspp.v1i1.184

Dawes, R. M., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its

implications for intergroup communication and the indirect assessment of attitude. *Journal of Personality and Social Psychology*, *21*, 281–295. https://doi.org/10.1037/h0032322

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*, 273–291. https://doi.org/10.1177/1745691617746796

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83–87. https://doi.org/10.1111/1467-8721.01235

Epley, N., Keysar, B., van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*, 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Epley, N., & Waytz, A. (2009). Mind perception. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The handbook of social psychology* (5th ed., pp. 498–541). Oxford University Press.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*, 939–946. https://doi.org/10.1177/0956797612464058

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2012). False memories of fabricated political events. *Journal of Experimental Social Psychology*, *49*, 280–286. https://doi.org/10.1016/j.jesp.2012.10.013

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). *Moral foundations dictionary for linguistic analyses 2.0*. OSF. https://doi.org/10.17605/OSF.IO/EZN37

Frimer, J. A., Gaucher, D., & Schaefer, N. K. (2014). Political conservatives' affinity for obedience to authority is loyal, not blind. *Personality and Social Psychology Bulletin*, *40*, 1205–1214. https://doi.org/10.1177/0146167214538672

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1–12. https://doi.org/10.1016/j.jesp.2017.04.003

Goldstein, N. J., Vezich, I. S., & Shapiro, J. R. (2014). Perceived perspective taking: When others walk in our shoes. *Journal of Personality and Social Psychology*, *106*, 941–960. https://doi.org/10.1037/a0036395

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLoS ONE*, *7*, Article e50092. https://doi.org/10.1371/journal.pone.0050092

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297. https://doi.org/10.1093/pan/mps028

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*, 261–266. https://doi.org/10.1126/science.aaa8685

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology*, *37*, 307–315. https://doi.org/10.1037/hea0000586

Iyengar, S., & Westwood, S. J. (2014). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*, 690–707. https://doi.org/10.1111/ajps.12152

Jost, J. T., Nosek, B. A., & Gosling, S. D. (2008). Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science*, *3*, 126–136. https://doi.org/10.1111/j.1745-6916.2008.00070.x

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*, 109–128. https://doi.org/10.1037/0033-295x.100.1.109

Jurafsky, D., & Martin, J. (2017). *Speech and natural language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Jussim, L., & Zanna, M. P. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology*, *37*, 1–93. https://doi.org/10.1016/S0065-2601(05)37001-8

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*, 1–16.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*, 732–735. https://doi.org/10.1038/nclimate1547

Keltner, D., & Robinson, R. J. (1993). Imagined ideological differences in conflict escalation and resolution. *International Journal of Conflict Management*, *4*, 249–262. https://doi.org/10.1108/eb022728

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*, 833–848. https://doi.org/10.1177/0146167208315158

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*, 1094–1096. https://doi.org/10.1126/science.aao2998

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243. https://doi.org/10.1037//0022-3514.47.6.1231

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

McCabe, K. T. (2016). Attitude responsiveness and partisan bias: Direct experience with the Affordable Care Act. *Political Behavior*, *38*, 861–882. https://doi.org/10.1007/s11109-016-9337-9

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0

Messick, D. M., & Mackie, D. M. (1989). Intergroup relations. *Annual Review of Psychology*, *40*, 45–81. https://doi.org/10.1146/annurev.ps.40.020189.000401

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in pre-training distributed word representations*. arXiv. https://arxiv.org/abs/1712.09405

Mill, J. S. (1869). *On liberty*. Longmans, Green, Reader, and Dyer.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*, 665–675. https://doi.org/10.1177/0146167203029005010

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. https://doi.org/10.1037/0033-295X.84.3.231

O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. In: *Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, pp. 1–8.

Palfrey, T. R., & Poole, K. T. (1987). The relationship between information, ideology, and voting behavior. *American Journal of Political Science*, *31*, 511–530. https://doi.org/10.2307/2111281

Paluck, E. L., Green, S. A., & Green, D. P. (2018). The contact hypothesis re-evaluated. *Behavioural Public Policy*, *3*, 129–158. https://doi.org/10.1017/bpp.2018.25

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC* [Computer software]. liwc.net.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making*, *15*, 476–498.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*, 65–85. https://doi.org/10.1146/annurev.psych.49.1.65

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Pew Research Center. (2014). *Political polarization in American politics*. http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/

Porter, M. F. (1980). An algorithm for suffix stripping. *Program Electronic Library and Information Systems*, *14*, 130–137. https://doi.org/10.1108/00330330610681286

Prior, M., Sood, G., & Khanna, K. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science*, *10*, 489–518. https://doi.org/10.1561/100.00014127

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381. https://doi.org/10.1177/0146167202286008

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*, 404–417. https://doi.org/10.1037/0022-3514.68.3.404

Rutjens, B. T., Sutton, R. M., & van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Personality and Social Psychology Bulletin*, *44*, 384–405. https://doi.org/10.1177/0146167217741314

Safire, W. (2008). *Safire's political dictionary*. Oxford University Press.

Safire, W. (2009) "Straw-Man Issue". New York Times, June 2, 2009

Saguy, T., & Kteily, N. (2011). Inside the opponent's head: Perceived losses in group position predict accuracy in metaperceptions between groups. *Psychological Science*, *22*, 951–958. https://doi.org/10.1177/0956797611412388

Scherer, A. M., Windschitl, P. D., & Graham, J. (2015). An ideological house of mirrors: Political stereotypes as exaggerations of motivated social cognition differences. *Social Psychological and Personality Science*, *6*, 201–209. https://doi.org/10.1177/1948550614549385

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences of the USA*, *116*, 7662–7669. https://doi.org/10.1073/pnas.1805871115

Sherman, D. K., Nelson, L. D., & Ross, L. D. (2003). Naïve realism and affirmative action: Adversaries are more similar than they think. *Basic and Applied Social Psychology*, *25*, 275–289. https://doi.org/10.1207/S15324834BASP2504_2

Sidanius, J. (1984). Political interest, political information search, and ideological homogeneity as a function of sociopolitical ideology: A tale of three theories. *Human Relations*, *37*, 811–828. https://doi.org/10.1177/001872678403701003

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among US presidential and vice presidential candidates. *Journal of Research in Personality*, *41*, 63–75. https://doi.org/10.1016/j.jrp.2006.01.006

Stern, C., & Kleiman, T. (2015). Know thy outgroup: Promoting accurate judgments of political attitude differences through a conflict mindset. *Social Psychological and Personality Science*, *6*, 950–958. https://doi.org/10.1177/1948550615596209

Stoknes, P. E. (2015). *What we think about when we try not to think about global warming: Toward a new psychology of climate action*. Chelsea Green Publishing.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*, 111–147. https://www.jstor.org/stable/2984809

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*, 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Talisse, R., & Aikin, S. F. (2006). Two forms of the straw man. *Argumentation*, *20*, 345–352. https://doi.org/10.1007/s10503-006-9017-8

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, *45*, 74–83. https://doi.org/10.1037/0022-3514.45.1.74

Thompson, L., & Hastie, R. (1990). Social perception in negotiation. *Organizational Behavior and Human Decision Processes*, *47*, 98–123. https://doi.org/10.1016/0749-5978(90)90048-E

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288. https://www.jstor.org/stable/2346178

Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, *24*, 2454–2462. https://doi.org/10.1177/0956797613494848

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, *83*, 1298–1313. https://doi.org/10.1037//0022-3514.83.6.1298

Van Boven, L., Judd, C. M., & Sherman, D. K. (2012). Political polarization projection: Social projection of partisan attitude extremity and attitudinal processes. *Journal of Personality and Social Psychology*, *103*, 84–100. https://doi.org/10.1037/a0028145

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91–99. https://doi.org/10.1186/1471-2105-7-91

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151. https://doi.org/10.1126/science.aap9559

Waytz, A., Young, L. L., & Ginges, J. (2014). Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proceedings of the National Academy of Sciences of the USA*, *111*, 15687–15692. https://doi.org/10.1073/pnas.1414146111

Westfall, J., van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*, 145–158. https://doi.org/10.1177/1745691615569849

Yeomans, M., Minson, J., Collins, H., Chen, F., & Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, *160*, 131–148. https://doi.org/10.1016/j.obhdp.2020.03.011

Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry*, *22*, 159–182. https://doi.org/10.1080/1047840X.2011.551743