American Political Science Review (2024) 1–18
doi:10.1017/S0003055424000819

Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments

YAMIL RICARDO VELEZ *Columbia University, United States*
PATRICK LIU *Columbia University, United States*

A long-standing debate in political psychology considers whether individuals update their beliefs and attitudes in the direction of evidence or grow more confident in their convictions when confronted with counter-attitudinal arguments. Though recent studies have shown that instances of the latter tendency, which scholars have termed attitude polarization and "belief backfire," are rarely observed in settings involving hot-button issues or viral misinformation, we know surprisingly little about how participants respond to information targeting deeply held attitudes, a key condition for triggering attitude polarization. We develop a tailored experimental design that measures participants' core issue positions and exposes them to personalized counter-attitudinal information using the large language model GPT-3. We find credible evidence of attitude polarization, but only when arguments are contentious and vitriolic. For lower valence counter-attitudinal arguments, attitude polarization is not detected. We conclude by discussing implications for the study of political cognition and the measurement of attitudes.

Whether people process political information in an even-handed fashion or actively resist ideologically inconvenient claims is crucial for understanding citizen competence (Kuklinski et al. 2000). Democratic accountability, it is said, hinges upon citizens' capacity to act on political preferences consistent with unbiased evaluations of economic and political circumstances. Accountability may be threatened if citizens process information chiefly in service of partisan and ideological goals (Achen and Bartels 2017, chap. 10; Shapiro and Bloch-Elkon 2008).

In their seminal article, Taber and Lodge (2006) observed troubling evidence that individuals exposed to a balanced set of pro and con arguments did not moderate their issue positions, but instead developed stronger attitudes. Nyhan and Reifler (2010) found misinformation corrections could "backfire," resulting in higher levels of factual inaccuracy. The finding that individuals assimilate congenial information and reject counter-attitudinal information seems particularly alarming amid a contemporary media age characterized by widespread misinformation. If people double down on false notions when challenged, then extreme caution is warranted when designing corrections or encouraging political deliberation—so the argument goes.

The ubiquity of these processes has been called into question by recent studies showing that "backfire" and "attitude polarization" are rare. Wood and Porter (2019) examine over 50 political claims and find robust evidence that corrections improve belief accuracy across statements varying in partisan relevance and valence. Meta-analyses of experimental studies testing the impact of fact-checking have also revealed that beliefs generally respond to factual corrections by moving in the direction of evidence (Walter et al. 2020). Assessing both attitudes and beliefs, Guess and Coppock (2020) find no evidence of attitude polarization for politically charged topics such as gun control, minimum wage, and capital punishment.

A central priority in this recent strain of research has been to employ salient issues or viral pieces of misinformation when evaluating attitude polarization and "belief backfire." But although these design choices increase personal relevance, the failure to detect more extreme positions may nonetheless reflect the possibility that issue areas or claims used in extant research are not sufficiently important or accessible. Defending attitudes in the way envisioned by scholars who have found evidence of "attitude polarization" might occur only when attitudes are deeply held, stable, and personally relevant—a defense frequently invoked by past proponents (e.g., Nisbett and Ross 1980, 180). Though positive belief updating and attitude change occur for more peripheral issues, attitude polarization might depend on whether people feel invested in an issue or cause.

Additional design choices may have further muted the appearance of this phenomenon in recent scholarship, magnifying uncertainty about the state of the literature. Studies priming directional motives by imploring participants to take partisan goals into account (Bayes et al. 2020) or encouraging the activation of "online" as opposed to "memory-based" processing (Redlawsk 2002) have found modest evidence of

Corresponding author: Yamil Ricardo Velez, Assistant Professor, Department of Political Science, Columbia University, United States, yrv2004@columbia.edu.
Patrick Liu, Ph.D. Student, Department of Political Science, Columbia University, United States, ppl2115@columbia.edu.

Received: March 22, 2023; revised: December 04, 2023; accepted: May 24, 2024.

polarization. Ceiling effects present persistent challenges for detecting attitude polarization, for as Taber and Lodge (2006, 757) note, "while the theory holds that those with the most extreme attitudes are the most prone to become even more extreme, detecting any such change is thwarted by the upper and lower bounds of the scale."

We conduct a critical test of the attitude polarization hypothesis that addresses these many critiques. Most notably, our design targets personal issue importance by (i) measuring participants' most deeply held issue attitudes and (ii) exposing them to personally relevant counter-attitudinal information. We accomplish this by leveraging a powerful large language model (LLM), GPT-3, that is capable of constructing personally tailored attitude measures and political arguments on the fly.[1] In addition, we (iii) encourage participants to engage in directional motivated reasoning and (iv) are careful to construct outcome measures that reduce ceiling effects. Indeed, since LLMs are capable of recovering attitudes that participants rank at the maximum of strength and certainty scales, we develop two new validated measures of attitudes—attitude defense and extremity—that are less susceptible to scale constraints.

We present a summary of our findings in Table 1. First, we carry out two studies modeled after Taber and Lodge (2006) that encourage participants to process a randomized mixture of pros and cons. Here, we fail to detect attitude polarization and find modest evidence of moderation for attitude certainty. Because these arguments are relatively brief and neutral in nature, we increase the intensity of the treatment in Study 3 by generating paragraph-long arguments written in the first person that disagree with the participants' issue positions, and find evidence of *moderation* in attitude strength. In Studies 4 and 5, we increase the intensity of treatments once more by generating arguments that harshly attack the issue positions of participants, and also develop new attitudinal measures that are less susceptible to ceiling effects. Across these two studies, we find that participants exposed to highly intense, emotionally charged counterarguments exhibit evidence of attitude polarization.

We find that mere exposure to counterarguments is insufficient to trigger attitude polarization even when arguments target deeply held issue positions. Instead, it is only when arguments cross into incivility and vitriol that polarization is observed. Our findings serve to reconcile more recent findings uncovering persuasive effects using tamer pieces of information, and broader concerns about attitude polarization that have proven elusive to detect. Though we contend that "attitudinal backlash" is a rare occurrence when individuals engage with mundane political content, toxic social media interactions and hostile attacks on issue positions may have the opposite effect. Indeed, we see these patterns as a mirror image of recent research reliably showing the power of a "nonjudgmental exchange of narratives" (e.g., Kalla and Broockman 2020). We conclude by discussing how the use of tailored interventions and outcome measures can improve our understanding of political behavior and persuasion while also highlighting ethical considerations that must be taken into account when using these technologies.

TABLE 1. Summarizing Our Findings

Experiment	1	2	3	5a	4	5b
Attitude Strength	∅	∅	–	∅	+	+
Attitude Certainty	∅	–	∅	–	+	∅
		Lower Valence			High Valence	

Note: +(–) indicates polarization (moderation), while ∅ indicates no measurable difference. Attitude strength is measured using a 7-point Likert item in Experiments 1–3 and a multi-item extremity scale in Experiments 4 and 5. Attitude certainty is measured using a 101-point scale in Experiments 1 and 2, a multi-item scale in Experiment 3, and a multi-item attitude defense scale in Experiments 4 and 5. Experiment 5 presented participants with either a lower valence or a higher valence counterargument. The values underneath 5a (5b) reflect the findings comparing the lower (higher) valence arm to a placebo condition.

DETECTING ATTITUDE POLARIZATION

Evidence of attitude polarization as widely pervasive has been elusive under the empirical literature, with recent studies suggest that polarization is "the exception, not the rule" in political learning (Guess and Coppock 2020, 1500). Below, we recount how the litmus test for detecting polarization has grown increasingly difficult to satisfy as detailed studies have failed to replicate the phenomenon. We pinpoint key assumptions in the theoretical foundations of the attitude polarization hypothesis deserving deeper inquiry.

The idea of attitude polarization received widespread attention following Lord, Ross, and Lepper's (1979) canonical demonstration, in which students reported strengthening their convictions after being exposed to mixed evidence for the deterrent effects of capital punishment. Lord and colleagues attributed this finding to *biased assimilation*, wherein individuals accept evidence that supports their initial view but discount evidence that contradicts it. At scale, they argued, biased assimilation would entail opinion divergence between people who hold opposing views when presented the same mixture of pro- and counter-attitudinal information.

Though many scholars criticized the use of self-described opinion change by Lord, Ross, and Lepper (1979) and failed to replicate polarization when they measured actual attitudes (e.g., Kuhn and Lao 1996; Miller et al. 1993), Taber and Lodge (2006, 756) argued

[1] Over the course of the study, OpenAI released an instruction fine-tuned version of GPT-3, davinci-003, which was described as GPT-3.5, as well as an entirely new model, GPT-4. Throughout the article, we use the GPT-3 as short-hand, but Experiments 3–5 use the more advanced class of GPT-3 models that is commonly described as GPT-3.5.

that these studies may simply have "failed to arouse sufficient partisan motivation to induce much biased processing." After displaying impassioned pro and con arguments by real interest groups to Stony Brook students with strong prior attitudes about affirmative action and gun control, Taber and Lodge found that many students reported higher levels of posttreatment attitude strength. They characterized these results as most consistent with the affect-based explanations offered by the theory of motivated reasoning.

The Motivated Reasoning Account

Motivated reasoning is primarily associated with the formulation of Kunda (1990), under which individuals confronted with new information are motivated by one of two goals: to reach an accurate conclusion or to minimize friction with prior beliefs. "Accuracy" and "directional" goals determine how individuals perceive, process, and respond to information inconvenient to their worldview.

Taber and Lodge (2006) brought motivated reasoning further into the mainstream and drew special attention to the role of affect. They pointed to "hot cognition," the hypothesis that previously evaluated sociopolitical concepts remain affectively charged in memory (Redlawsk 2002). Reexposure to a concept automatically and instantaneously activates the accompanying affect, which in turn generates directional motivations. However, for experimental subjects to experience the intensity of affect required to undergo attitude polarization, Taber and Lodge stressed the necessity of both *sufficiently strong priors* and *sufficiently heated stimuli.*

In the aftermath of Taber and Lodge (2006), motivated reasoning became the dominant rationale for attitude polarization and similar phenomena whenever they were observed. When Nyhan and Reifler (2010) found that corrections to misinformation strengthened the convictions of certain strongly committed subjects —a result they coined the "backfire effect"—they likewise relied on theories of affective and motivated reasoning (323), speculating that vigorously counterarguing belief-incongruent information can bring to the front of one's mind more congenial considerations that bolster prior beliefs. A wide range of studies since have observed polarization and backfire on varied issue topics, usually among narrow subgroups containing the most impassioned respondents (e.g., Ecker and Ang 2019; Hart and Nisbet 2012; Nyhan and Reifler 2015; Zhou 2016).

Yet the most recent evidence to date raises real doubt about whether polarization is but a rare occurrence. Haglin's (2017) direct replication of Nyhan and Reifler (2015) failed to replicate the backfire effect. Wood and Porter (2019) failed to replicate backfire despite testing 52 issues across five experiments with more than 10,000 subjects, as did Guess and Coppock (2020) across three large survey experiments. Aggregating conclusions from these and numerous other studies that have failed to replicate the backfire phenomenon (e.g., Garrett, Nisbet, and Lynch 2013; Nyhan et al. 2020; Weeks 2015), Swire-Thompson, DeGutis, and Lazer's (2020, 288) review offers three potential explanations for its elusory nature: "either (a) the backfire effect is difficult to elicit on the larger group level, (b) it is extremely item-, situation-, or individual-specific, or (c) the phenomenon does not exist at all." Coppock (2023) found that across a wide range of survey experiments, respondents consistently updated their beliefs and attitudes in the direction of evidence and by approximately the same amount, regardless where their priors stood. Though Coppock's experiments excluded group cues, his findings are buttressed by Tappin, Berinsky, and Rand (2023), who found that persuasive messages' effects were not substantially diminished when respondents learned that their in-party leader (Trump or Biden) opposed the message's position.

In this article, we seek to advance the debate on attitude polarization by drawing renewed attention to the role of attitude strength in nurturing the conditions that purportedly give rise to motivated reasoning. That subjects experience a deep-seated desire to protect attitudes that are affect-laden or to which they assign great personal importance has long been presumed essential to the phenomenon, if not necessary for its occurrence. Taber and Lodge (2006, 757) termed this the *attitude strength effect*, whereby "those citizens voicing the strongest policy attitudes will be most prone to motivated skepticism." Existing studies that refute the attitude polarization hypothesis have not yet tackled this defense head on.

Revisiting Attitude Strength in Motivated Reasoning

At a high level, psychologists have defined four features of strong attitudes: resistance to change, stability over time, impact on judgment and cognition, and impact on behavior (Krosnick and Petty 1995). Different accounts have varied substantially how they frame the relationship between strength and other features of attitudes. As Boninger, Krosnick, and Berent (1995) note, psychologists long considered strength more a metaphor than a formal construct and measured attitude extremity, accessibility, certainty, and commitment each as indicators for strength. In a more recent review of the subject, Howe and Krosnick (2017) describe these many terms as stand-alone concepts that co-occur with strength in complex ways, while nonetheless subsuming them under the phrase "strength-related features."

For our purposes, we will speak broadly about attitude strength as an umbrella concept, relying where appropriate on such related factors as importance and certainty. Our reasoning is twofold. First, a wide-ranging literature largely predating Taber and Lodge (2006) explored potential relationships between various strength-related features of attitudes and the constituent processes and outcomes of motivated reasoning. According to these studies, attitude polarization may be exacerbated or more likely to occur among those who assign their relevant attitude greater levels of importance (Tesser and Leone 1977), commitment (Pomerantz, Chaiken, and Tordesillas 1995), and

extremity (McHoskey 1995). While nuanced differences abound, the literature suggests substantial overlap between these strength-related features, particularly in the ways they bear on motivated reasoning. A second rationale relates to the empirical strategy in this article. The theory of motivated reasoning as elaborated by Taber and Lodge (2006) relies on attitude strength without disambiguating between these varied features. We imitate their approach and consider a range of strength-related concepts and measures in the interest of constructing a generous test of the attitude polarization hypothesis.

Individuals may assign personal importance to issues and subsequently develop strong attitudes about them for heterogeneous reasons. The degree to which an individual perceives an issue as bearing upon their self-interest, social identification with important reference groups (i.e., partisan identification), and cherished social and personal values determines the level of personal importance allotted to that issue (Boninger, Krosnick, and Berent 1995). Highly important attitudes, in turn, might generate the motivations that induce selective information processing (Lavine, Borgida, and Sullivan 2000). There is suggestive evidence that more important attitudes are more resistant to change (Gopinath and Nyer 2009; Zuwerink Jacks and Devine 1996), and corrections to misinformed beliefs may be less effective when those beliefs are perceived as personally important (Vidigal and Jerit 2022).

The key question taken up in this article is whether strong and weak attitudes operate so differently on reasoning that motivated reasoning might only be observed when the former are at play. That is, would an experimental design that elicits subjects' most deeply held attitudes unearth evidence of attitude polarization? Taber and Lodge (2006, 754) questioned whether earlier studies had dismissed attitude polarization primarily because their selection of arguments and evidence were insufficiently affect-laden and thus unable to arouse the requisite motivations. When Nyhan and Reifler (2010, Study 2) corrected misperceptions about the discovery of weapons of mass destruction during the US's 2003 invasion of Iraq, their post hoc analysis found backfire effects present only among a subset of conservative respondents who rated Iraq as the most important issue for the US, leading the authors to similarly conclude that backfire effects are contingent on issue importance.

Several recent articles have endeavored to resolve this lingering uncertainty, whether by testing a great assortment of issues of "keen political interest" and with partisan valence (Wood and Porter 2019, 142) or by selecting issues made salient by the latest national news. For instance, Guess and Coppock (2020, Experiment 1) administered one study of backfire effects using information about gun safety just days after the 2016 Orlando mass shooting. Those studies found little evidence of polarization in spite of research designs generous toward the motivated reasoning hypothesis.

It is not clear, however, that personal salience and national salience can be merged as concepts. As Ryan and Ehlinger (2023, 5) argue, there is no consensus to date about what number of political topics the typical person cares about, nor the number of distinct topics that matter to the electorate at large. Not every issue of national import will induce in every individual the urge to muster up defenses against discordant evidence. A critical assessment of attitude polarization instead necessitates a research design that can, first, invite open-ended input from participants about the issues they hold deeply important and, second, assess the efficacy of persuasion on these core attitudes by confronting participants with tailored pro- and counter-attitudinal responses.

Tailoring Information with Large Language Models

We perform such a critical assessment by taking advantage of advances in LLMs—dense neural networks with high levels of performance in replicating human speech. Though training language models is time- and resource-intensive, *pre-trained*, *task-agnostic* models with general applicability and that can be adapted to specific natural language processing tasks with far fewer resources are growing increasingly accessible, presenting a window of opportunity for experimental social science (Linegar, Kocielnik, and Alvarez 2023). Porter and Velez (2022), for instance, demonstrate how using LLMs to automate a "placebo sampling" process can improve survey experiments.

We illustrate two novel applications of pre-trained language models using GPT-3, an autoregressive language model developed by OpenAI. First, using participants' open-ended responses detailing issues of personal importance, GPT-3 was able to construct tailored 7-point Likert items measuring participants' attitude strength and certainty about those issues. Second, we demonstrate GPT-3's capacity to generate a suite of persuasive arguments when provided only (a) a topic of political discussion by the research participant and (b) brief, issue-agnostic instructions from the researcher about the position to be taken (pro/con). Recent scholarship affirms the capacity of GPT-3 to generate policy arguments that are comparable to messages written by lay humans in terms of persuasive impact (Bai et al. 2023). The ability to develop personalized measurement scales and tailor persuasive messages in the context of large online experiments advances the persuasion literature beyond extant strategies for confronting personal issue importance.[2]

How can we know whether our design has, in fact, tapped an issue about which the respondent holds deeply held convictions, and thus for which motivated skepticism is likely to be activated? Taber and Lodge (2006) crystallized motivated reasoning's constitutive processes into hypotheses they termed *disconfirmation*

[2] In Studies 3–5, we use a more advanced model, GPT-3.5, to construct arguments and create tailored outcome measures. While writing this manuscript, OpenAI released GPT-4, an even more advanced model. We retain the use of GPT-3.5 to maintain relative consistency in model performance across the studies, and because the task of argument creation and summarization is not such an advanced task that it warrants the use of GPT-4.

bias and the *prior attitude effect.* Disconfirmation bias predicts an inclination to counterargue and denigrate attitudinally incongruent arguments more than congruent arguments. The prior attitude effect, whereby those who hold core issue preferences evaluate ideologically congruent arguments as stronger than incongruent ones, finds roots in Lord, Ross, and Lepper's (1979) classic study.[3] There, students read about two studies on the deterrent effects of the death penalty: one touting its efficacy and the other undermining it. Participants regarded the study that affirmed their prior view as significantly better designed and more convincing than the study opposing their view. We test for both of these mechanisms in our experiments.

Constructing a Critical Test of the Attitude Polarization Hypothesis

In addition to eliciting important issues and tailoring persuasive treatments, our experiments involve several other design choices aimed at raising the chances of detecting polarization through an "easy" test. Failing to detect attitude polarization despite all these conditions provides strong evidence that the phenomenon may occur only rarely. We briefly summarize these design choices.

1. **Strong attitudes:** Respondents provide open-ended input about their deeply held attitudes. We validate the strength of these attitudes using multiple measures.
2. **Tailored arguments:** We employ GPT-3 to generate arguments tailored to respondents' strong attitudes. We test single-sentence arguments (Experiments 1 and 2) and full paragraphs (Experiments 3–5).
3. **Treatment intensity**: We increase the negative valence of the arguments throughout our studies. Experiments 1 and 2 present fairly neutral counterarguments in the form of a thought-listing task; Experiment 3 presents a negative argument written in the first person; Experiments 4 and 5 use a style of argumentation that relies on vitriol.
4. **Motivational primes:** We randomly assign participants to be primed for either directional or accuracy motivations prior to argument exposure in Experiments 1 and 2.
5. **Multiple information conditions:** In Experiments 1 and 2, we randomly assign participants to view and respond to four pro-attitudinal arguments, four counter-attitudinal arguments, or two of each. Past studies typically included only a con or a mixed condition. The mixed argument condition offers a potential test of the biased assimilation hypothesis à la Lord, Ross, and Lepper (1979), while evidence of polarization under mixed or con conditions could align with the motivated reasoning view.
6. **Multiple outcome measures:** Studies of attitude polarization often face limitations due to ceiling effects. Similar to Taber and Lodge (2006), we minimize this issue by employing multiple outcome measures as well as a multi-item scale of attitude certainty, improving our chances of detecting the strengthening of already strong attitudes. In Studies 4 and 5, we develop two new attitudinal measures that are less susceptible to ceiling effects.
7. **Measuring intervening processes:** We validate whether our experimental design successfully set conditions for triggering motivated reasoning using measures of two theorized mechanisms of motivated reasoning—the "prior attitude effect" and disconfirmation bias.

[3] Throughout the article, we use "core issue" as shorthand for an issue that the participant prioritizes in the context of an open-ended question. We show that issue attitudes elicited this way tend to score high on strength, certainty, and stability.

DATA, METHODS, AND RESULTS

We develop a tailored experimental design that constructs personalized interventions and outcome measures using open-ended responses. The basic structure of the experimental design is the following (Table 2): (1) participants report their position on a core political issue using an open-ended question, (2) text from this question is passed to GPT-3, returning a one-sentence summary that is used to construct Likert-style items and arguments related to the issue position, and (3) participants report attitude strength and certainty with respect to this deeply held issue position. Experiment 1 implemented this approach in one wave using a pre-post design (Clifford, Sheagley, and Piston 2021). Experiment 2, carried out in two waves, assessed attitudes toward core and peripheral issues to evaluate variation in effects across levels of attitude strength. A modified design was employed in Experiments 3–5 to explore how attitudes respond to more heavy-handed persuasive messages, compared against a placebo control wherein subjects read a random vignette drawn from a validated corpus of placebo texts.[4]

In Appendix B.1 of the Supplementary Material, we show that participants' self-reported core attitudes map onto 66 unique issue topics. Although abortion and other salient issues are mentioned frequently, no single issue accounts for more than a quarter of responses. Eliciting open-ended responses avoids assuming any particular topic of national salience will be *personally* salient to participants.

Experiment 1

We recruited 2,141 participants using the online survey platform CloudResearch Connect (CR).[5] The survey was in the field from September 28, 2022 to October

[4] All experiments were reviewed and approved by the institutional review board at the authors' institution (Protocol AAAU3638).

[5] To ensure high-quality open-ended responses and compliance with the thought-listing, we sought a survey vendor with especially attentive subjects. Online convenience samples often replicate average treatment effects (ATEs) and conditional average treatment effects (CATEs) observed using nationally representative samples (Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2015).

TABLE 2. Outline of Five Experiments

Feature	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Examining attitude strength					
Strong attitudes	✓	✓	✓	✓	✓
Measuring intervening processes					
Test of biased assimilation (mixed info condition)	✓	✓			
Directional/accuracy goals (motivational primes)	✓	✓			
Disconfirmation bias (thought listings)	✓	✓			
Prior attitude effect (argument strength ratings)		✓			
Attitude strength moderates mechanisms (within-subjects core vs. peripheral issue)		✓			
Addressing ceiling effects (outcome measures)					
7-point attitude strength	✓	✓	✓		
101-point attitude certainty	✓	✓	✓		
Multi-item attitude certainty		✓	✓		
Multi-item attitude extremity				✓	✓
Multi-item attitude defense				✓	✓
Varying treatment intensity					
Tailored arguments	✓	✓	✓	✓	✓
Set of four arguments (pro/mixed/con)	✓	✓			
Argumentative paragraph			✓		✓
Argumentative paragraph with vitriolic language				✓	✓

1, 2022.[6] Upon reading the consent form and agreeing to participate, participants were taken to the following open-ended question: "Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? For example, if you care about farm subsidies, you can write 'I believe farm subsidies should be increased to help farmers.' Please write a brief sentence about an issue that you care about and where you stand on the issue." Their response was passed to OpenAI's GPT-3 text completion API, a one-sentence summary was produced, and this summary was presented as a 7-point Likert item, ranging from "strongly disagree" to "strongly agree." In 17% of cases, GPT-3 was unable to provide output. All prompts were passed through OpenAI's content filter to minimize the possibility that GPT-3 would generate "toxic content." If this condition could not be met, we flagged the observation in Qualtrics and provided a generic set of arguments. As noted in our pre-analysis plan, we exclude these cases because participants did not receive tailored information. This leaves us with an effective sample size of 1,782.[7]

The text completion API was instructed to summarize each statement in one sentence. For example, one participant wrote "Congress should address the issue on healthcare cost" [sic] (see Table 3). The GPT-3 completion API produced the following Likert item for this respondent: "To what extent to do you agree or disagree with the following statement? I believe that Congress should address the issue of healthcare costs." After responding to a personalized Likert item, participants were asked about their level of certainty regarding this issue position on a 0–100 scale.

Participants were randomized to one of two motivational conditions. These conditions described the thought-listing task, but emphasized engaging with the argument from a more fair-minded perspective ("accuracy" motivation) or from the perspective of maintaining consistency ("directional" motivation). Those in the former condition were explicitly instructed to "ignore any personal feelings or emotions" and focus on the "truth of each statement," whereas those in the latter condition were instructed to "not worry" about the accuracy of each statement but instead focus on understanding what the statement means to them, given their existing beliefs. Participants were then randomized to one of three information conditions for the thought-listing task, exposing them to four pro-attitudinal arguments ("Pro" condition), four counter-attitudinal arguments ("Con"), or two pro- and two counter-attitudinal arguments ("Mixed"). Arguments were produced by passing a summary of the participant's issue position to GPT-3's text completion API. This output was returned as an eight-item JSON file that could be displayed in Qualtrics. In total, four pros and four cons were generated for each open-ended response. In the "Pro" ("Con") condition, respondents saw all of the pro (con) arguments. In the "Mixed" condition, respondents saw two randomly selected pro and con arguments. A participant who strongly supported universal healthcare, for example, could see pro-attitudinal arguments such as "it's unfair that

[6] The pre-analysis plan for Experiment 1 is available here: https://aspredicted.org/2xf65.pdf. All studies presented in this manuscript were approved by Columbia University's Human Research Protection Office (Protocol #AAAU3638).

[7] See "Additional Study Details" (hereafter ASD) documentation on Dataverse for an analysis of whether certain issues are less likely to receive valid GPT-3 output. In ASD B.3, we note that the API calls occur before treatment, thus avoiding confounding. Though there is modest variation across issues with respect to error-free completion rates, we fail to reject the null of equal completion rates ($F(14, 1{,}310) = 0.79$, $p = 0.68$).

TABLE 3. Representative GPT-3 Output

Open-ended response	Congress should address the issue on healthcare cost.
Likert	I believe that Congress should address the issue of healthcare costs.
Pro1	Healthcare costs are a burden on many Americans.
Pro2	Reducing healthcare costs could free up money for other important programs.
Pro3	Addressing the issue of healthcare cost could improve the quality of life for many Americans.
Pro4	Congress has a responsibility to represent the people and their interests.
Con1	There is no one-size-fits-all solution to reducing healthcare costs.
Con2	Some solutions to reducing healthcare costs could be unpopular with voters.
Con3	Addressing the issue of healthcare cost could be costly in itself.
Con4	It is not clear that Congress has the power to directly address healthcare cost.

people have to choose between basic needs and medical care" or counter-attitudinal arguments such as "there would likely be long wait times for non-urgent medical procedures if everyone had equal access to affordable healthcare," depending on the information condition. Each argument was presented with a corresponding text box below it so that participants could write their thoughts about each argument.[8] Attitude strength and certainty were then measured using the same procedure as in the pretreatment phase. Participants filled out a brief demographic battery and read a debriefing statement revealing that the information they saw was produced by GPT-3.

Models

We regress posttreatment measures of attitudes (i.e., attitude strength and attitude certainty) on *pretreatment strength and certainty measures*, information condition indicators, motivation condition indicators, and their interaction using OLS regression. Due to the inclusion of pretreatment measures, these models capture within-study variation in attitudes:

$$Y_i = \alpha + \beta_1 \text{Mixed}_i + \beta_2 \text{Con}_i + \beta_3 \text{Pretreatment Strength}_i + \beta_4 \text{Pretreatment Certainty}_i + \beta_5 \text{Mixed}_i \times \text{Directional Prime}_i + \beta_6 \text{Con}_i \times \text{Directional Prime}_i + \epsilon. \quad (1)$$

For ease of interpretation, we plot treatment effects across the various information and motivation conditions. Based on the previous literature, our key expectation is that exposure to counter-attitudinal information in the "Con" and "Mixed" conditions should increase attitude strength and polarization relative to the "Pro" condition. Although previous research has stressed the importance of balanced information in activating motivated reasoning, we are agnostic regarding the ordering of effect sizes for the "Con" and "Mixed" conditions and consider *any* positive estimates associated with the "Con" condition (β_2, β_6) or "Mixed" condition (β_1, β_5) as evidence of attitude polarization.

Before moving on to our key analyses, we conduct validation tests of whether intervening processes such as disconfirmation bias can be detected in our experiment (see Appendix A.1 of the Supplementary Material). We find that those exposed to pro-attitudinal arguments generally spend approximately 5% less time on the thought-listing task than those exposed to counter-attitudinal arguments (SE = 0.007; $p < 0.001$). The share of denigrating responses to arguments also increases from about 9% in the "Pro" condition to 25% in the "Con" condition (SE = 0.019; $p < 0.001$). These results are consistent with an extensive literature suggesting that people expend more cognitive effort when considering arguments inconsistent with their prior beliefs and attitudes.

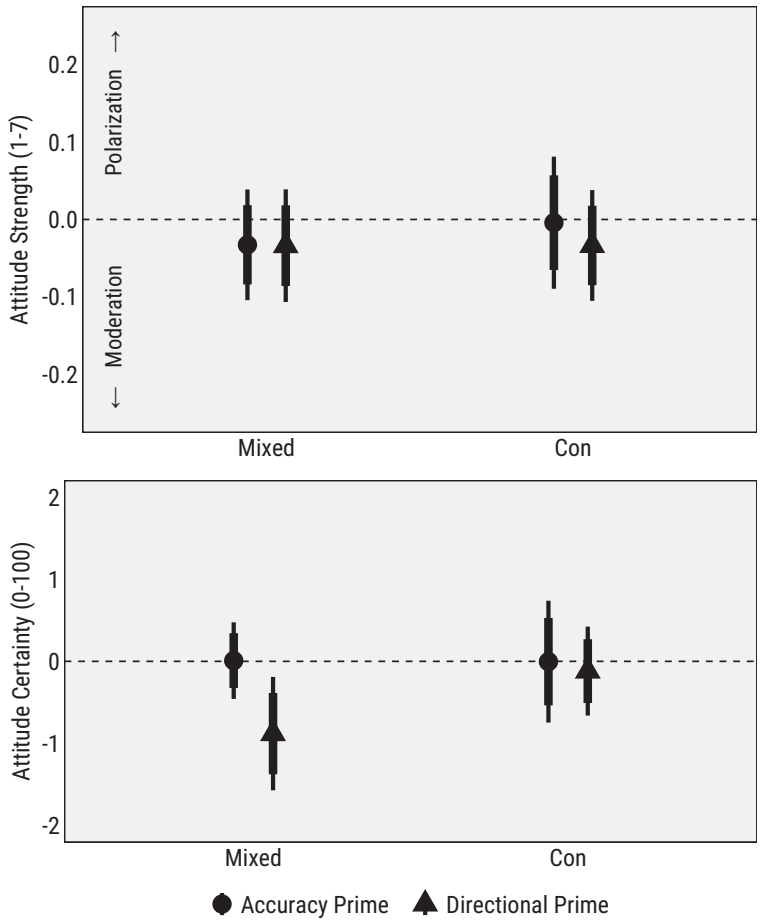
We now consider the effects of the different experimental conditions on attitude strength (see ASD A.1 for full model results). Recall that participants provided us with open-ended responses to a question asking about their most deeply-held issue position and we used GPT-3 to generate a personalized 7-point Likert item (see Appendix B.3 of the Supplementary Material for pretreatment distributions). If attitudes are polarizing due to exposure to counter-attitudinal information, we ought to expect a larger point estimate in the "Mixed" and "Con" conditions relative to the "Pro" condition. If attitudes are moderating, estimates should move in a negative direction.

As shown in Figure 1, differences across information and motivation conditions are small, negative, and statistically indistinguishable from zero. Among participants primed to consider accuracy, those in the "Mixed" and "Con" conditions score 0.015 (SE = 0.038) and 0.022 (SE = 0.035) scale points lower on attitude strength than those in the "Pro" condition, respectively. For those assigned to the directional prime, differences between the "Pro" condition and the "Mixed" and "Con" conditions are –0.022 (SE = 0.034) and –0.026 (SE = 0.034) scale points, respectively. In sum, we do not find evidence that the different information and motivation conditions are shifting attitudes.

We next turn to our analyses of the 101-point attitude certainty scale. As the pretreatment distributions illustrate (Appendix B.3 of the Supplementary Material), the mean attitude certainty score is 97. 71% of the sample selected the maximum certainty score, while 90% provided a score above 90. Figure 1 shows minimal differences between information and motivation

[8] The full list of arguments can be found here: https://bit.ly/3ZlZu26.

FIGURE 1. Effect Estimates on Attitude Strength and Certainty



Note: This figure presents point estimates and confidence intervals for attitude strength and certainty. Thick bands are 84% confidence intervals, used to facilitate visual comparisons of coefficients. Thin bands are 95% confidence intervals. ASD A.1 presents full model results.

conditions. For those in the accuracy condition, the difference between the "Con" and "Pro" conditions is –0.025 (SE = 0.359) scale points, and the difference between the "Mixed" and "Pro" conditions is 0.004 (SE = 0.23) scale points. Those in the directional condition evince a slightly different pattern. Those in the "Mixed" condition score approximately 0.88 scale points lower on attitude certainty relative to those in the "Pro" condition (SE = 0.35; $p = 0.01$). This corresponds to a shift of 0.12 standard deviations and runs in the opposite direction of what the motivated reasoning paradigm leads us to expect: those in the "Mixed" condition should express *more* certainty in their attitudes, relative to those who receive pro-attitudinal information.

In our validation tests, we detect evidence of disconfirmation bias, with those in the "Con" condition responding more critically to their assigned arguments than those in the "Pro" and "Mixed" conditions. However, despite providing arguments that target core issues and priming participants to consider information from a directional perspective, we fail to uncover evidence of attitude polarization across two measures. In one case, we observe reductions in certainty, which runs counter to expectations about attitude polarization.

That being said, certain design choices could explain the non-findings. First, we conducted the study in a single wave, with pre- and post-intervention measures of outcomes spaced several questions apart. Although this can improve precision without significantly altering treatment effects (Clifford, Sheagley, and Piston 2021), participants may feel compelled to report a consistent outcome value within the same survey, which would hinder our ability to detect effects. Second, even in the absence of a "consistency bias," our design might foreclose the possibility of detecting polarization due to ceiling effects. We included a continuous measure of attitude certainty for this reason. However, the baseline level of certainty in our sample was 97 on a 0–100 scale. Thus, even with a finer-grained measure, we may still have trouble detecting positive treatment effects.

Experiment 2

We conducted a multi-wave experiment on CR to address the aforementioned limitations. Wave 1 was in the field from October 10, 2022 to October 13, 2022 (N = 1, 591).[9] In Wave 1, we obtained pretreatment measures of attitude strength, certainty, duration, and discussion frequency for core and peripheral issues. We measured core issue attitudes using the open-ended approach described in Experiment 1. To assess if effects differed depending on attitude strength, we also measured attitudes toward more peripheral issues. Respondents selected from a prepared list of 10 issues and responded to attitudinal questions (e.g., strength and certainty). To ensure that we were measuring weaker attitudes, we explicitly asked participants to select issues that they follow, but otherwise do not have a strong opinion on. The 10 issues included foreign aid, school funding, inflation reduction, public transportation, universal healthcare, free speech, gun control, minimum wage, student loans, and marijuana legalization. Participants could choose any side of the issue.

For both core and peripheral issue positions, we measured attitude strength (7-point Likert item), attitude certainty (101-point certainty scale), and "attitude clarity and correctness" (seven items, hereafter "multi-item certainty"). Following Petrocelli, Tormala, and Rucker (2007), each item in the multi-item certainty scale is a 9-point Likert item ranging from 1 (not at all certain) to 9 (very certain). This scale scores high on reliability (α = 0.92 for peripheral issues; α = 0.96 for core issues). After a 1-week washout period, we carried out randomization and posttreatment outcome measurement in Wave 2 ($N = 1,313$).

We conducted two within-subjects trials by issue (peripheral vs. core) that randomly assigned participants to one of two primes (directional vs. accuracy) and one of three information conditions (100% pro-attitudinal arguments, 100% counter-attitudinal arguments, or 50% pro- and counter-attitudinal arguments). Conditions were independently randomized such that a respondent could have been assigned first to a peripheral issue trial with a directional prime and counter-attitudinal arguments, then to a core issue trial with an accuracy prime and a mixture of pro- and counter-attitudinal arguments. We randomized the order of the trials. Within each trial, we (i) presented arguments and the thought-listing task, (ii) measured strength and certainty outcomes, then (iii) asked participants to rate the strength and accuracy of all four pro and all four con arguments generated by the model, including arguments not shown during the thought-listing task. After both trials were complete, we measured demographics, thanked participants for their participation, and debriefed them. Of the 1,313 participants who completed Wave 2, 1,137 received a valid GPT-3 response for their "core issue" and 1,225 received a valid GPT-3 response for their "peripheral issue." Below, we focus on the findings with respect to the "core issue" before moving on to our discussion of the "peripheral issue" findings.[10]

The top panel of Figure 2 presents effect estimates on attitude strength for core issues (see ASD A.2 for full model results). We observe little attitude change across the information and motivation conditions. In all four tests, effect sizes are small and statistically indistinguishable from zero. We now consider the multi-item certainty measure. As was the case with the other attitudinal measures in Experiment 1, scores are generally at the upper end of the distribution, with a mean score of 6.75 and a standard deviation of 0.80. Focusing on this outcome, we find evidence of moderation when comparing the "Con" and "Mixed" conditions to the "Pro" condition. We detect shifts of −0.13 (SE = 0.07) and −0.05 (SE = 0.07) scale points in the "Con" and "Mixed" conditions relative to the "Pro" condition when respondents are primed to operate in a directional mode. These estimates are magnified in the accuracy condition such that those in the "Con" and "Mixed" conditions score 0.17 (SE = 0.07) and 0.18 (SE = 0.07) scale points lower on certainty than those in the "Pro" condition. This is equivalent to a movement of approximately 0.09 standard deviations on the outcome.

Importantly, in ASD B.1, we report an exploratory change score analysis demonstrating that these shifts reflect those in the "Con" and "Mixed" condition moderating, rather than those in the "Pro" condition growing more confident. By and large, participants in the "Pro" condition either moderated or showed little change between the pre- and posttreatment measures of attitude strength and certainty on their core issues. In ASD B.8, we additionally report an exploratory analysis of conditional ATEs by subtopic. We find no evidence that attitude polarization is observed for certain subtopics but not others. In sum, we observe attitude change inconsistently, but when we do, estimates uniformly move toward moderation, instead of polarization.

In Experiment 2, we fail to detect evidence of attitude polarization. Although we observe moderate differences in our outcomes due to different information conditions, we do not detect evidence that those in the "Mixed" and "Con" conditions become more extreme or certain relative to those in the "Pro" condition. Instead, we find some evidence of moderation in the presence of counter-attitudinal arguments, especially when participants are reminded to be accurate.

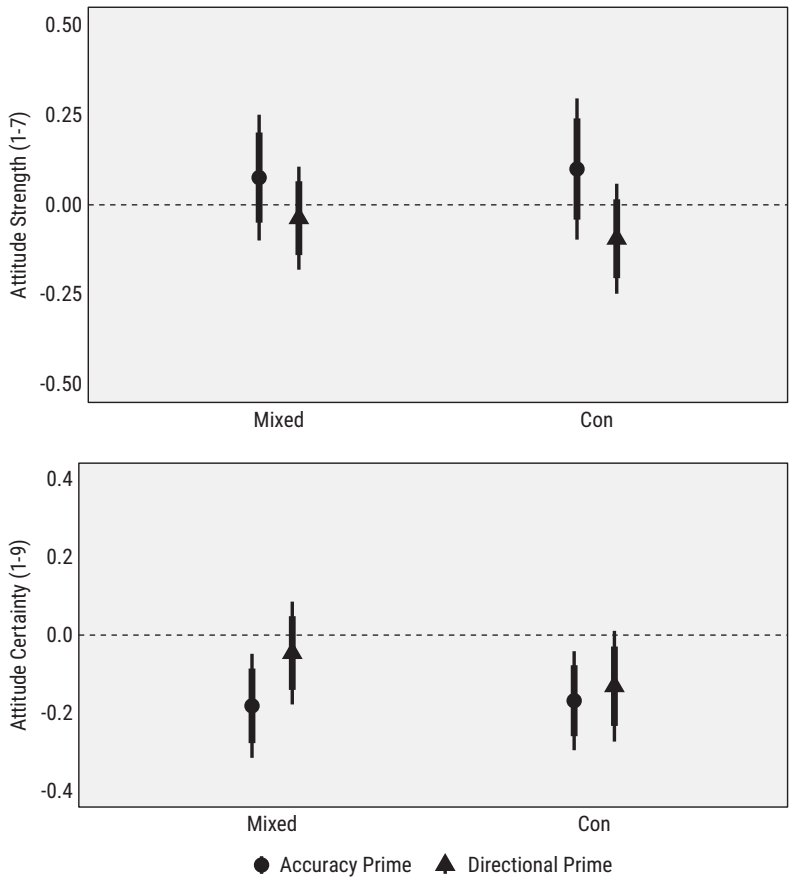
Potential Concerns

Across two studies that primed directional motives, varied exposure to different kinds of information, and targeted deeply held issue positions, we failed to detect attitude polarization. Despite this, one might object that we have not successfully measured strong attitudes

[9] The pre-analysis plan for Experiment 2 can be found here: https://aspredicted.org/9xf6p.pdf.

[10] The full list of arguments can be found here: https://bit.ly/3GR0Z15.

FIGURE 2. Effect Estimates on Attitude Strength and Certainty



Note: This figure presents point estimates, 84% confidence intervals (thick bands), and 95% confidence intervals (thin bands) for attitude strength across information and motivation conditions. Facets are defined by issue attitude strength. ASD A.2 presents full model results.

or that our information interventions are too weak. We address these potential concerns below.

Insufficiently Strong Attitudes

Might our failure to observe attitude polarization indicate we are still assessing weak attitudes? As a benchmarking exercise, we compare core and peripheral issues across several dimensions to assess if we can recover a distinct set of issue attitudes using our open-ended method. Mean responses on a 7-point Likert item are 6.76 (s = 0.78) for the core issue and 5.03 (s = 1.29) for the peripheral issue. Using the 0–100 certainty scale, the mean response was 96 (s = 9.15) for the core issue and 65 (s = 23) for the peripheral issue. Focusing on multi-item certainty, mean responses are 8.41 (s = 0.88) and 5.8 (s = 1.96). Across all scales, attitudes elicited using our open-ended method possess averages close to the maximum.

76% of respondents report that they have held their core issue attitude for more than 4 years, while this number drops to 41% for the peripheral issue attitude (16% reported forming the peripheral attitude in the middle of the survey vs. 2% for the core issue). 76% of respondents report "never" speaking about the peripheral issue in the past 6 months, compared to 37% for the core issue. Examining attitude stability, approximately 44% of respondents retain the same level of attitude strength across waves for the peripheral issue, while this number is 81% for the core issue. When subsetting on those who report the maximum level of attitude strength in Wave 1, 62% of respondents remain at the maximum for the peripheral issue, compared to 87% for the core issue. In sum, our evidence suggests attitudes pertaining to the core issue are stable, durable, and personally relevant.

We detect strong evidence of a prior attitude "effect." As we report in Appendix A.3 of the Supplementary Material, we find that people generally rate pro-attitudinal (counter-attitudinal) arguments as stronger (weaker) when they possess stronger attitudes toward a given topic. We also find that the gap between pro and con ratings is larger for core versus peripheral issues. In Appendix A.4 of the Supplementary Material, we also successfully replicate disconfirmation bias, whereby individuals expend cognitive resources on combatting counter-attitudinal information. Participants generally spend more time on the thought listing

and are more likely to denigrate arguments when they challenge (vs. support) one's pre-existing attitudes. In sum, our results indicate that we have measured attitudes that are strong enough to provoke an attitude defense.

Insufficiently Strong Interventions

Because we detect evidence of disconfirmation bias and prior attitude effect—essential mechanisms for triggering polarization according to theories of motivated reasoning—our design appears to satisfy most of the necessary conditions that cause participants to "deposit more supportive evidence and affect in memory" (Taber and Lodge 2006, 757). One could contend that the counter-attitudinal arguments provided by GPT-3 are insufficiently strong to trigger the self-defense mechanisms described in previous work. To address this claim, we provide an empirical test of GPT-3's capacity to generate strong arguments in Appendix B.2 of the Supplementary Material. We find that while pro arguments generated by GPT-3 are rated as weaker than human-generated arguments, con arguments—the key driver of attitude polarization—generated by GPT-3 are rated to be just as persuasive as human-generated con arguments. Still, even if we grant that GPT-3 cannot generate strong arguments, previous studies have found that *weak* con arguments generate more "refutative thoughts" than strong con arguments (Benoit 1987; Petty and Cacioppo 1986). Thus, if GPT-3 is underperforming humans in the creation of counter-attitudinal arguments, low-quality con arguments ought to produce *more* attitude polarization than stronger arguments (to the extent that an accumulation of "refutative thoughts" renders it easier to bolster one's attitudes).

That being said, perhaps GPT-3 arguments are not sufficiently confrontational. The deliberative nature of the thought-listing task could also mute reflexive responses to politically incongenial content that might otherwise be detected using more conventional designs. Moreover, there could be heterogeneity in responses to counter-attitudinal information that our design is ill-equipped to detect, given the large number of experimental conditions. Though the average participant does not appear to polarize, there could be politically relevant subgroups who do. We conduct additional experiments that strengthen exposure to counter-attitudinal information by confronting participants with long-form arguments. We also simplify the experimental design by devising a traditional survey experiment where participants are randomly assigned to different pieces of information. Experiment 3 presents participants with a strongly worded, but still civil, counter-argument, whereas Experiments 4 and 5 explore whether highly uncivil and emotionally charged arguments trigger attitude polarization when targeting deeply held beliefs.

Experiment 3

From December 12 to 13, 2022, we recruited two thousand participants using the online sample provider Lucid.[11] Given concerns about data quality on this platform (Aronow et al. 2020), we created an additional GPT-3 script that assessed the quality of the open-ended responses as they were submitted and filtered out participants who provided unintelligible responses before treatment assignment.[12] Since our design hinges on legible open-ended responses to produce counter-attitudinal arguments, these quality checks were necessary. As described in our pre-analysis plan, our analysis omits participants flagged by GPT-3 as "low quality," those who repeated the example issue position in our instructions, and those who did not receive any output from GPT-3. These variables are measured before treatment assignment, and thus do not bias our estimates.

Upon completing the open-ended question, participants responded to a single-question attention check that measured retention of information provided in a news vignette (Kane, Velez, and Barabas 2023) and a set of demographic questions. Participants were then randomly shown either (1) one of four thousand placebo texts drawn from the Porter and Velez (2022) corpus or (2) a tailored counter-argument produced by GPT-3. The placebo texts are news-like vignettes that are approximately a paragraph long, matching the average length of the tailored arguments. Instead of retrieving four pros and cons, we instructed GPT-3 to write a "paragraph-long passionate rebuttal" where the author "strongly disagrees" with the statement provided by the participant (see Appendix C.2 of the Supplementary Material for example arguments).[13]

Following our preregistration plan, we estimate a covariate-adjusted ATE for attitude strength and certainty, adjusting for age, education, income, partisanship, ideology, and race (white = 1; non-white = 0). We estimate these models using OLS regression with HC2 standard errors. Focusing on the first panel of Figure 3 (see ASD A.3 for full model results), the covariate-adjusted ATE of counter-attitudinal information on attitude strength is −0.510 scale points on a 7-point scale (SE = 0.09; $p < 0.001$).[14] This corresponds to a shift of 0.26 control-group standard deviation units and is comparable to other effect sizes detected in the persuasion literature (e.g., Broockman and Kalla 2016 detect a shift of 0.40 scale points in response to their canvassing intervention). An exploratory analysis of treatment effect heterogeneity reveals that effects on attitude strength are consistent across groups varying in open-ended response quality, attentiveness, ideology, and political knowledge. In general, we find that CATEs are always negative and statistically significant

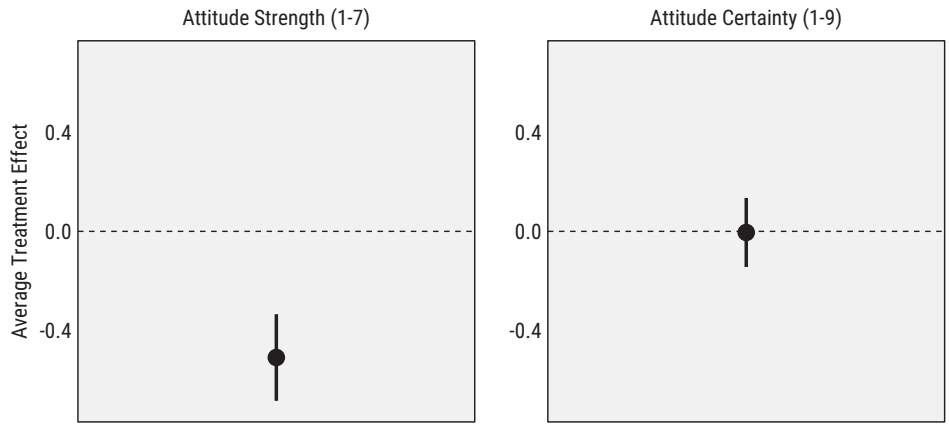
[11] The pre-analysis plan for Experiment 3 can be found here: https://aspredicted.org/km7ya.pdf.

[12] In a pilot study, we compared hand-coded data and the predictions of the data quality script. Approximately 5% of open-ended responses tagged as "low quality" had legible input, compared to 81% of those tagged "medium quality" and 91% of those tagged "high quality."

[13] The full list of arguments can be found here: https://bit.ly/3XjLjZC.

[14] The ATE (without covariate adjustment) is −0.487 scale points (SE = 0.09; $p < 0.001$).

FIGURE 3. Effect of Counter-Attitudinal Information on Attitude Strength and Certainty



Note: This figure presents point estimates and 95% confidence intervals for the covariate-adjusted ATE of counter-attitudinal information on attitude strength and certainty. ASD A.3 presents full model results.

across the range of moderators. In the case of political knowledge, we find that CATEs are even *more negative* among the most politically sophisticated. The difference between the lower and upper tertiles is –0.433 scale points (SE = 0.22; $p = 0.048$), suggesting a larger "attitude moderation" effect among those who are more politically knowledgeable. This runs contrary to expectations from the motivated reasoning literature that political sophisticates are especially likely to reject counter-attitudinal information. Turning to certainty, we find that the effect of counter-attitudinal information on certainty is approximately zero ($\hat{ATE} = -0.005$; SE = 0.07; $p = 0.95$).[15] In sum, we find robust evidence of a decrease in attitude strength when participants are exposed to tailored counter-attitudinal information but are unable to detect evidence of shifts in attitude certainty.

Lingering Concerns: Ceiling Effects and Contentious Counterarguments

Despite our efforts to address ceiling effects with multiple scales, including a 101-point and seven-item certainty scale, we show in Appendix B.3 of the Supplementary Material that a majority of participants score at the maximum of the attitude certainty and Likert items. Even our seven-item certainty scale is susceptible to this issue, with 45% scoring at the maximum in Experiment 2 and 31% doing so in Experiment 3. Though this is further validation that we are eliciting deeply held issue positions, this poses a challenge for detecting attitude polarization.

Insufficient treatment intensity is another potential concern. Despite the stronger arguments presented in Experiment 3, one could argue that they are neutrally valenced and do not necessarily provoke a defense of one's attitudes. Previous studies finding evidence of "boomerang effects" (i.e., attitude polarization) have documented its emergence in highly contentious situations and in the presence of affect-laden arguments or insults (Kim, Levine, and Allen 2017; Taber and Lodge 2006). With increasing consumption of social media and growing political incivility, there exists a style of argumentation grounded in vitriol and personal attacks that may be a strong candidate for triggering attitude polarization. Assessing the impact of negatively valenced messages constitutes an important class of arguments that we have yet to explore and that the most recent literature on motivated reasoning has not directly examined. We address these concerns in Experiments 4 and 5 by introducing new measures of attitude strength and intensifying counterarguments.

Experiment 4

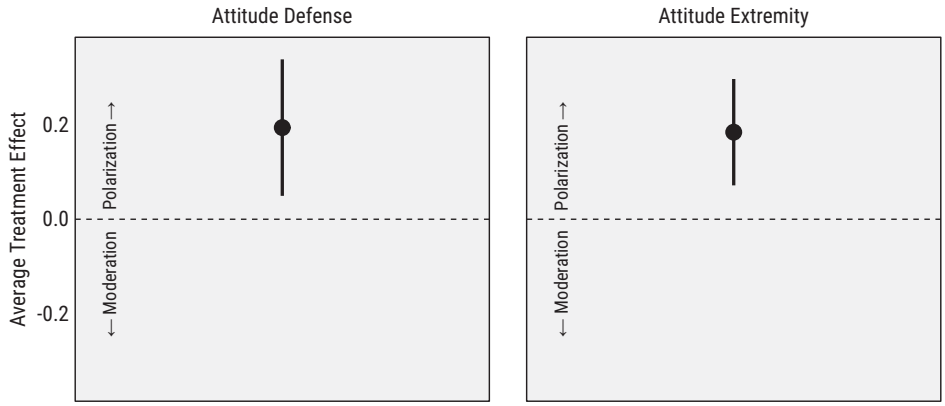
Following Experiment 3, we implemented a two-arm placebo-controlled persuasion experiment. To address the concern of insufficient treatment intensity, we explicitly instructed GPT-3 to "create a paragraph-long affectively charged counterargument" that is "emotionally charged" and "directly attacks what the person said." In contrast to the previous arguments used in our studies, which were mostly respectful and dispassionate, the new LLM-generated messages were designed to engage participants on an emotional level, potentially triggering a defensive response. For example, a participant writing about free public college received the following argument:[16]

> It is absolutely absurd to suggest that public universities should be tuition free. Are you suggesting that the

[15] To minimize multiple comparisons, we preregistered an analysis of only these two outcomes. However, an exploratory analysis of the single item certainty scale reveals a small negative ATE ($\hat{ATE} = -1.708$; SE = 0.87; $p = 0.05$).

[16] See https://bit.ly/47DTYvL for a full list of arguments.

FIGURE 4. Effect of Emotionally Charged Arguments on Attitude Defense and Extremity (Experiment 4)



Note: This figure presents point estimates and 95% confidence intervals for the covariate-adjusted ATE of counterarguments on attitude defense and extremity. ASD A.4 presents full model results.

> hardworking taxpayers should foot the bill for every student's education, regardless of their academic abilities or financial situation? This is not only unfair, but it goes against the very principles of a merit-based society. Education is a privilege, not a right, and it should be earned through hard work and dedication. By making public universities tuition free, you are essentially devaluing the hard work and sacrifices of those who have paid for their education. Furthermore, this proposal would only lead to a decrease in the quality of education, as universities would not have the necessary funds to provide top-notch resources and facilities. It is time to stop expecting handouts and start taking responsibility for our own education.

In Appendix B.5 of the Supplementary Material, we explicitly compare these negatively valenced arguments to the more anodyne style used in the previous studies. We find that these arguments are more likely to include moralizing language and emotionally charged words.

To address ceiling effects, we carried out a measurement study (N = 338) between Experiments 3 and 4 where we assessed the performance of four new measures capturing attitudinal intensity. Our goal was to identify measures that were less subject to ceiling effects, were correlated with other constructs (e.g., Likert ratings and certainty scores), and exhibited high levels of concurrent validity. As in our previous studies, participants were asked to report a core issue in an open-ended question. Once again, we found that participants rated their own issue highly on a 7-point Likert scale ($\bar{x} = 6.72$; s = 0.71) and 101-point certainty scale ($\bar{x}$ = 94; s = 12).

We tested four new measures that capture strong attitudes: (1) a personalized conjoint that asked participants to choose between hypothetical candidates taking positions on their core issue (who also varied on important dimensions such as age, career, race, and party)[17]; (2) an attitude defense measure that captured participants' willingness to defend their position across a variety of scenarios (e.g., a public interview and a campus speech); (3) an attitude extremity scale that captured willingness to incur costs to support one's issue position; and (4) an allocation task where participants were asked to allocate funds across four issue domains, including their own. Our measurement study revealed that attitude defense and extremity were less susceptible to ceiling effects, were modestly correlated with traditional measures, and exhibited fairly high levels of concurrent validity (see Appendix B.4 of the Supplementary Material).

For Experiment 4, 2,017 participants were recruited on CR over a period of 4 days, from October 9 to 12, 2023.[18] We estimate a covariate-adjusted ATE with age, political sophistication, education, income, ideology, political party, 101-point attitude certainty, 7-point attitude strength, and self-reported political behavior as covariates. Our key outcomes are attitude extremity and defense.[19]

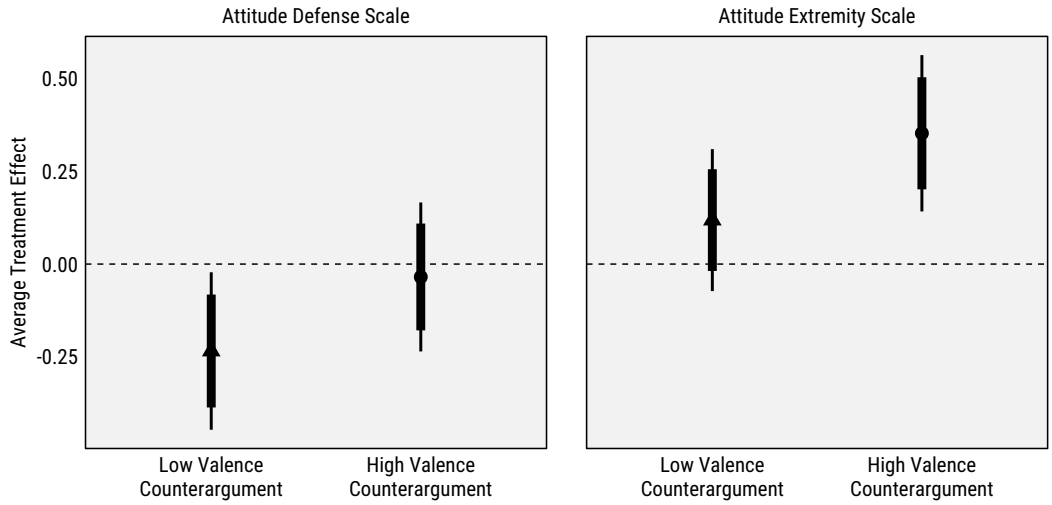
We present the covariate-adjusted ATEs for the two measures in Figure 4. Beginning with the attitude defense estimates, we find that exposure to a hostile counterargument *increases* attitude defense by 0.19 scale points (SE = 0.07; p = 0.01). This is equivalent to approximately 17% of the gap in scores when comparing the least and most politically knowledgeable individuals in the control group. Turning now to attitude extremity, we observe a positive effect on extremity of 0.18 scale points (SE = 0.06; $p < 0.001$). This is equivalent to 56% of the extremity gap between low and high political knowledge individuals in the control group. Thus, after addressing ceiling effects and increasing treatment intensity, we

[17] See Velez (2023) and Ryan and Ehlinger (2023) for examples of this methodology.

[18] The pre-analysis plan for Experiment 4 can be found here: https://aspredicted.org/3dt9n.pdf.

[19] Appendix B.3 of the Supplementary Material confirms that these measures were less susceptible to ceiling effects. Approximately 13% of participants scored at the maximum of attitude defense, whereas a meager 0.77% scored at the maximum of extremity.

FIGURE 5. Effect of Low and High Valence Counterarguments on Attitude Defense and Extremity (Experiment 5)



Note: This figure presents point estimates and 95% confidence intervals for the covariate-adjusted ATEs of two treatment arms on attitude defense and extremity. ASD A.5 presents full model results.

observe a trend where participants are more inclined to bear costs in defense of their issue position and to speak out in a range of high-stakes scenarios. Given that attitude strength is a latent variable, with increasing willingness to act and incur costs as potential behavioral indicators, these findings are consistent with a process of attitude polarization.

Experiment 5

The findings from Experiment 4 suggest that stronger interventions are capable of producing attitude polarization. However, given the elusiveness of this effect throughout the literature and our inability to detect it in the previous studies, we sought to replicate and extend Experiment 4 by including an additional arm that mirrored the tamer style of argumentation used in the earlier studies. Specifically, we added a treatment arm to Experiment 4's design that presented participants with a more neutrally valenced argument, modeled after the instructions used to generate the output in Experiment 3 (see Appendix B.5 of the Supplementary Material for a comparison of the two argument styles).[20] We collected data on 1,922 CR participants from October 19 to 23, 2023.[21] As in the previous experiment, we estimated covariate-adjusted ATEs with the same vector of covariates and outcome measures.

Figure 5 displays the covariate-adjusted ATEs for the two outcomes. We find that the lower valence counterargument produces a *decrease* on the attitude defense scale of 0.22 scale points (SE = 0.09; $p = 0.02$), whereas the effect for the higher valence counterargument is much smaller ($\hat{ATE} = -0.02$; SE = 0.09) and not significant ($p = 0.48$). In contrast, the contentious counterargument produces a positive increase of 0.20 scale points (SE = 0.07; $p = 0.01$) on the attitude extremity scale, replicating Experiment 4. The effect for the low valence counterargument is about half the size of the high valence counterpart ($\hat{ATE} = 0.09$; SE = 0.07) and not significant ($p = 0.17$). Overall, higher emotional valence arguments appear to have a more pronounced impact on attitude extremity, with lower valence counterarguments mirroring the pattern of attitude moderation we had observed previously.[22]

CONCLUSION

In this article, we reviewed the political science literature on attitude polarization and aimed for a critical assessment of the phenomenon. We exposed participants to tailored counter-attitudinal arguments, measured attitudes using various methods, and primed participants to think in more directional terms. Based on a careful reading of the literature on MR, we sought to create ideal conditions for detecting attitude polarization. In Experiments 1 and 2, we observed small but statistically significant effects in the opposite direction: attitudes became *less* certain after seeing counter-attitudinal material. The third study simplified the experimental design and exposed participants to longer, more affectively charged arguments. Here, we observed significant

[20] See https://bit.ly/47Kqtbx for a full list of arguments.
[21] The pre-analysis plan for Experiment 5 can be found here: https://aspredicted.org/34pa5.pdf.
[22] ASD B.9 addresses whether exposure to vitriolic content could have precipitated differential attrition across conditions. Though attrition is low in Experiments 4 and 5 and appears not to affect our findings, we discuss in ASD B.14 why scholars should be vigilant about attrition when exposing subjects to aggressive language.

decreases in attitude strength. Our findings from these three studies suggest that mere exposure to counter-attitudinal information is insufficient to trigger attitude polarization, even when attitudes are strongly held.

To assess potential scope conditions, we conducted additional experiments that varied the intensity of treatments by generating tailored counter-attitudinal arguments that were contentious and vitriolic. We also mitigated the persistent issue of ceiling effects by devising new attitudinal measures. Contrary to the findings from the first three studies, we found robust evidence of attitude polarization when a sharper, more confrontational style of argumentation was used. Experiment 5 extended this study by including tamer arguments modeled after our earlier attempts as an additional treatment arm. Here, we observed that vitriolic arguments were uniquely polarizing. Our ability to replicate attitude polarization across these two studies, but not in the initial set of experiments, suggests that the phenomenon is more likely to emerge when deeply held attitudes are confronted with antagonistic arguments.

We view the totality of our evidence as a kind of microcosm of the motivated reasoning literature. When scholars have used more neutral content encountered in op-eds or fact-checks, as in more recent work, moderation has been observed. However, more extreme arguments have been shown to trigger attitude polarization, a point highlighted in the canonical article by Taber and Lodge. Our results offer an important corrective, highlighting a potential limit at which negative persuasive communication may begin to strengthen convictions. Though "attitudinal backlash" is uncommon, it may be detected when the aforementioned conditions hold.

At first blush, our findings appear to vindicate popular views of motivated reasoning. However, such an interpretation would be premature. Although prominent studies invoking the theory argued for the importance of argument valence in triggering the phenomenon (Taber and Lodge 2006), subsequent descriptions held that it could be triggered even in the presence of neutral stimuli through a process of hot cognition, where political concepts and their "affective charge" are automatically retrieved from memory without much conscious processing (Lodge and Taber 2013, 19). Our early experiments—as well as other studies employing op-eds, fact-checks, and candidate statements—do not yield significant attitude polarization, calling into question the inflexibility and inevitability often attributed to this "strong" view of motivated reasoning. Furthermore, the narrowness of our conditions—high attitude strength and negatively valenced counterarguments—suggests that a theory of motivated reasoning predicting attitude polarization as the modal response to persuasion is unlikely to explain the lion's share of observed effects.

Though the political science literature on attitude change has been dominated by debates over Bayesian updating, with motivated reasoning playing the role of a useful foil, a well-developed literature in psychology suggests alternative mechanisms that can also influence attitude change. One such mechanism could be a tit-for-tat process whereby individuals reciprocate the emotional and communicative style they encounter in a persuasive exchange. When confronted with an overly aggressive or uncivil argument, individuals may respond in kind, leading to an escalation of conflict and further entrenchment of pre-existing attitudes. Conversely, individuals may be more amenable to attitude change when encountering respectful discourse, especially the kind rooted in personal experiences rather than in facts alone (Kubin et al. 2021). This latter point has been supported by a growing body of research suggesting that a "non-judgmental exchange of narratives" can promote attitude change even for politically controversial topics (Kalla and Broockman 2020).

This encourages us to move beyond the dichotomy of Bayesian updating and motivated reasoning when understanding attitude polarization. Future research could systematically examine how different combinations of source characteristics, conversational norms, and argument features influence attitude formation and change. This exploration could involve experimental designs that manipulate the tone and content of arguments, as well as the characteristics of the sources presenting them. Additionally, these studies could consider the role of audience characteristics, such as cognitive styles and tendencies toward incivility (e.g., Petersen, Osmundsen, and Arceneaux 2023), in shaping responses to persuasion.

Our study has important implications for the study of persuasion when deeply held attitudes are at stake. If attitude polarization depends on strong attitudes, conventional experimental designs might underestimate its prevalence. Attitudes measured using our tailored approach were generally at the upper end of dimensions such as stability, strength, and certainty. Standard measurement approaches involving closed-ended questions may have trouble capturing deeply personal issue positions that define individuals' political orientations. As we discuss in Appendix B.1 of the Supplementary Material, while salient issues such as abortion and healthcare are mentioned frequently by participants, no single issue accounts for more than a quarter of responses. Future research could investigate ways of improving tailored outcome measures, as well as how these measures differ from more traditional open-ended scales. Tailoring messages may yield important insights in domains such as misinformation when targeting rare but socially consequential conspiracy theories (Costello, Pennycook, and Rand 2024).

Despite the promise of LLMs, there are ethical and practical considerations researchers must account for. First, scholars should apply content filters to ensure that text generations do not inadvertently expose participants to harmful messages. In our studies, we applied these filters and observed zero instances of derogatory terms.[23] This does not negate the need to actively monitor output and ensure that LLMs are not inadvertently perpetuating harmful biases or stereotypes. Second, scholars should be

[23] OpenAI's models actively refrain from making negative statements about marginalized groups and have been accused of exhibiting a "liberal bias" (https://www.theverge.com/2023/2/17/23603906/openai-chatgpt-woke-criticism-culture-war-rules).

mindful of errors or false claims that may be produced by these models. All of our studies included a debriefing protocol describing our use of GPT-3 in constructing the arguments shown to participants. We also found that the model occasionally produced double-barreled questions or questions containing justifications when constructing tailored outcome measures, and thus would fail to meet the standards of survey researchers. We revised our pipeline to address this issue in Experiments 4 and 5, but this demonstrates the need to be vigilant when working with these models. At a minimum, using best-available content filters, actively monitoring output, and disclosing the use of these technologies to participants is a necessary and important step.[24] As these technologies grow more ubiquitous, countering some of their potentially harmful consequences will likely require active participation by academics and journalists (e.g., building dynamic interventions to counter false claims). Ultimately, the responsible integration of LLMs into survey research may help us gain a deeper understanding of political psychology, but also the opportunities and challenges associated with these emerging technologies.[25]

Motivated reasoning has had a profound impact on political psychology and our larger discipline. In this article, we consider one of the most troubling implications of motivated reasoning, as defined by Kunda (1990) and Taber and Lodge (2006): attitude polarization. Attitude polarization is troubling because it suggests that deliberation can harden opinions, factual corrections can further reinforce false beliefs, and persuading "true believers" can backfire. Our findings suggest that attitude polarization is not a robust or generalizable effect, but rather a contingent one that depends on the intensity and valence of the counterarguments. When people are exposed to arguments that they may encounter in op-eds, fact-checks, or in credible news media, attitudes are generally not bolstered, even when those arguments are relevant to deeply held issue positions or dispel congenial, but otherwise inconsequential, facts. However, violating the norms of civil discourse can have the deleterious effects feared by skeptics of open deliberation. Whether those fears turn out to be well-founded will depend on the future trajectory of political discourse which, at present, shows signs of increasing hostility and antagonism.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055424000819.

[24] All projects described here were approved through OpenAI's internal app review process on September 15, 2022. OpenAI has since discontinued this process, allowing API users to build applications as long as they adhere to the standards outlined here: https://platform.openai.com/docs/usage-policies/disallowed-usage.

[25] To support researchers implementing similar designs, our pipeline will be publicly available upon publication (Velez and Liu 2024).

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/SNG5BW.

ACKNOWLEDGEMENTS

We are grateful for feedback from the Editors, anonymous reviewers, and participants at the Columbia University Political Methodology Colloquium, Sciences Political Behavior Workshop, and Dutch Political Psychology Meeting. We also thank Donald P. Green, Ethan Porter, and Charles Taber for incisive comments on our manuscript. All errors are our own.

CONFLICT OF INTERESTS

The authors declare no ethical issues or conflict of interests in this research.

ETHICAL STANDARDS

The authors declare the human subjects research in this article was reviewed and approved by Columbia University's Human Research Protection Office and certificate numbers are provided in the text. The authors affirm that this article adheres to the principles concerning research with human participants laid out in APSA's Principles and Guidance on Human Subject Research (2020).

REFERENCES

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*, Revised edition. Princeton, NJ: Princeton University Press.

Aronow, Peter Michael, Joshua Kalla, Lilla Orr, and John Ternovski. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020." Working Paper. https://doi.org/10.31235/osf.io/8sbe4

Bai, Hui, Jan G. Voelkel, Johannes C. Eichstaedt, and Robb Willer. 2023. "Artificial Intelligence Can Persuade Humans on Political Issues." *OSF Preprints*. https://doi.org/10.31219/osf.io/stakv

Bayes, Robin, James N. Druckman, Avery Goods, and Daniel C. Molden. 2020. "When and How Different Motives Can Drive Motivated Political Reasoning." *Political Psychology* 41 (5): 1031–52.

Benoit, William L. 1987. "Argumentation and Credibility Appeals in Persuasion." *Southern Journal of Communication* 52 (2): 181–97.

Boninger, David S., Jon A. Krosnick, and Matthew K. Berent. 1995. "Origins of Attitude Importance: Self-Interest, Social Identification, and Value Relevance." *Journal of Personality and Social Psychology* 68 (1): 61–80.

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352 (6282): 220–4.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115 (3): 1048–65.

Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago, IL: University of Chicago Press.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences* 115 (49): 12441–6.
Costello, Thomas H., Gordon Pennycook, and David G. Rand. 2024. "Durably Reducing Conspiracy Beliefs Through Dialogues with AI." *PsyArXiv*. doi:10.31234/osf.io/xcwdn.
Ecker, Ullrich K. H., and Li Chang Ang. 2019. "Political Attitudes and the Processing of Misinformation Corrections." *Political Psychology* 40 (2): 241–60.
Garrett, R. Kelly, Erik C. Nisbet, and Emily K. Lynch. 2013. "Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory." *Journal of Communication* 63 (4): 617–37.
Gopinath, Mahesh, and Prashanth U. Nyer. 2009. "The Effect of Public Commitment on Resistance to Persuasion: The Influence of Attitude Certainty, Issue Importance, Susceptibility to Normative Influence, Preference for Consistency and Source Proximity." *International Journal of Research in Marketing* 26 (1): 60–8.
Guess, Andrew, and Alexander Coppock. 2020. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* 50 (4): 1497–515.
Haglin, Kathryn. 2017. "The Limitations of the Backfire Effect." *Research & Politics* 4 (3). https://doi.org/10.1177/2053168017716547
Hart, P. Sol, and Erik C. Nisbet. 2012. "Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization about Climate Mitigation Policies." *Communication Research* 39 (6): 701–23.
Howe, Lauren C., and Jon A. Krosnick. 2017. "Attitude Strength." *Annual Review of Psychology* 68 (1): 327–51.
Kalla, Joshua L., and David E Broockman. 2020. "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments." *American Political Science Review* 114 (2): 410–25.
Kane, John V., Yamil R. Velez, and Jason Barabas. 2023. "Analyze the Attentive and Bypass Bias: Mock Vignette Checks in Survey Experiments." *Political Science Research and Methods* 11 (2): 293–310.
Kim, Sang-Yeon, Timothy R. Levine, and Mike Allen. 2017. "The Intertwined Model of Reactance for Resistance and Persuasive Boomerang." *Communication Research* 44 (7): 931–51.
Krosnick, Jon A. and Richard E. Petty. 1995. "Attitude Strength: An Overview." In *Attitude Strength: Antecedents and Consequences*, eds. Richard E. Petty and Jon A. Krosnick, 1–24. Hillsdale, NJ: Lawrence Erlbaum Associates.
Kubin, Emily, Curtis Puryear, Chelsea Schein, and Kurt Gray. 2021. "Personal Experiences Bridge Moral and Political Divides Better Than Facts." *Proceedings of the National Academy of Sciences* 118 (6): e2008389118. https://doi.org/10.1073/pnas.2008389118
Kuhn, Deanna, and Joseph Lao. 1996. "Effects of Evidence on Attitudes: Is Polarization the Norm?" *Psychological Science* 7 (2): 115–20.
Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *Journal of Politics* 62 (3): 790–816.
Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98.
Lavine, Howard, Eugene Borgida, and John L. Sullivan. 2000. "On the Relationship between Attitude Involvement and Attitude Accessibility: Toward a Cognitive-Motivational Model of Political Information Processing." *Political Psychology* 21 (1): 81–106.
Linegar, Mitchell, Rafal Kocielnik, and R. Michael Alvarez. 2023. "Large Language Models and Political Science." *Frontiers in Political Science* 5: 1257092. https://doi.org/10.3389/fpos.2023.1257092
Lodge, Milton, and Charles S. Taber. 2013. *The Rationalizing Voter*. Cambridge: Cambridge University Press.
Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37 (11): 2098–109.
McHoskey, John W. 1995. "Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization." *Basic and Applied Social Psychology* 17 (3): 395–409.
Miller, Arthur G., John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd. 1993. "The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change." *Journal of Personality and Social Psychology* 64: 561–74.
Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2 (2): 109–38.
Nisbett, Richard E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.
Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42 (3): 939–60.
Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32 (2): 303–30.
Nyhan, Brendan, and Jason Reifler. 2015. "Does Correcting Myths about the Flu Vaccine Work? An Experimental Evaluation of the Effects of Corrective Information." *Vaccine* 33 (3): 459–64.
Petersen, Michael Bang, Mathias Osmundsen, and Kevin Arceneaux. 2023. "The 'Need for Chaos' and Motivations to Share Hostile Political Rumors." *American Political Science Review* 117 (4): 1486–505.
Petrocelli, John V., Zakary L. Tormala, and Derek D. Rucker. 2007. "Unpacking Attitude Certainty: Attitude Clarity and Attitude Correctness." *Journal of Personality and Social Psychology* 92 (1): 30–41.
Petty, Richard E., and John T. Cacioppo. 1986. *The Elaboration Likelihood Model of Persuasion*. New York: Springer.
Pomerantz, Eva M., Shelly Chaiken, and Rosalind S. Tordesillas. 1995. "Attitude Strength and Resistance Processes." *Journal of Personality and Social Psychology* 69 (3): 408–19.
Porter, Ethan, and Yamil R. Velez. 2022. "Placebo Selection in Survey Experiments: An Agnostic Approach." *Political Analysis* 30 (4): 481–94.
Redlawsk, David P. 2002. "Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making." *Journal of Politics* 64 (4): 1021–44.
Ryan, Timothy J., and J. Andrew Ehlinger. 2023. *Issue Publics: How Electoral Constituencies Hide in Plain Sight*. Cambridge: Cambridge University Press.
Shapiro, Robert Y., and Yaeli Bloch-Elkon. 2008. "Do the Facts Speak for Themselves? Partisan Disagreement as a Challenge to Democratic Competence." *Critical Review: A Journal of Politics and Society* 20 (1–2): 115–39.
Swire-Thompson, Briony, Joseph DeGutis, and David Lazer. 2020. "Searching for the Backfire Effect: Measurement and Design Considerations." *Journal of Applied Research in Memory and Cognition* 9 (3): 286–99.
Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–69.
Tappin, Ben M., Adam J. Berinsky, and David G. Rand. 2023. "Partisans' Receptivity to Persuasive Messaging Is Undiminished by Countervailing Party Leader Cues." *Nature Human Behaviour* 7 (4): 568–82.
Tesser, Abraham, and Christopher Leone. 1977. "Cognitive Schemas and Thought as Determinants of Attitude Change." *Journal of Experimental Social Psychology* 13 (4): 340–56.
Velez, Yamil Ricardo. 2023. "Trade-Offs in Latino Politics: Exploring the Role of Deeply-Held Issue Positions Using a Dynamic Tailored Conjoint Method." *Aletheia.* https://doi.org/10.17605/OSF.IO/G5RWK.
Velez, Yamil Ricardo, and Patrick Liu. 2024. "Replication Data for: Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments" Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/SNG5BW.
Vidigal, Robert, and Jennifer Jerit. 2022. "Issue Importance and the Correction of Misinformation." *Political Communication* 39 (6): 715–36.
Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37 (3): 350–75.

Weeks, Brian E. 2015. "Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation." *Journal of Communication* 65 (4): 699–719.
Wood, Thomas, and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41 (1): 135–63.
Zhou, Jack. 2016. "Boomerangs versus Javelins: How Polarization Constrains Communication on Climate Change." *Environmental Politics* 25 (5): 788–811.
Zuwerink Jacks, Julia R., and Patricia G. Devine. 1996. "Attitude Importance and Resistance to Persuasion: It's Not Just the Thought That Counts." *Journal of Personality and Social Psychology* 70: 931–44.