Appendix

A Pre-Registered Analyses

A.1 Validation Tests (Experiment 1)

We assess whether we are able to detect disconfirmation bias by regressing the share of denigrating responses to the GPT-3-produced arguments and total response time for the thought-listing task on information condition indicators, priming condition indicators, and their interaction. We measure the share of denigrating responses by asking crowdsourcing workers on Amazon Mechanical Turk to count the number of denigrating responses for four sets of argument-response pairs. For each respondent, we divide the number of denigrating responses by the total number of responses provided ($\bar{x} = .17$). We detect clear evidence of disconfirmation bias. Roughly 9% of thought-listing responses in the "Pro" condition rejected or dismissed the arguments that were presented, whereas the share of denigrating comments was 8pp higher (SE = 1pp) in the "Mixed" condition and 16pp higher (SE = 2pp) in the "Con" condition. A larger (smaller) share of denigrating comments were observed when "Con" ("Mixed") information was paired with a directional prime. However, the difference-in-difference estimates for both are not statistically significant. Moving on to timer data, we find evidence of slower response times for those in "Mixed" and "Con" conditions. These correspond to approximately 5% (SE = .008; $p < .001$) and 6% decreases (SE = .008; $p < .001$) in response times, respectively. Turning to the interaction model, those in the "Pro" condition who receive a directional prime spend approximately 3% less time on the thought-listing task than those who receive an accuracy prime (SE = .012; $p < .001$). However, we fail to find evidence that the directional prime conditions the effects of the "Con" and "Mixed" conditions on response time. Taken together, the evidence is broadly consistent with research on disconfirmation bias finding that people are more likely to denigrate and expend cognitive resources on counter-attitudinal versus pro-attitudinal information.

TABLE A1. The Effect of Information Conditions on Share of Denigrating Comments and Response Time Data

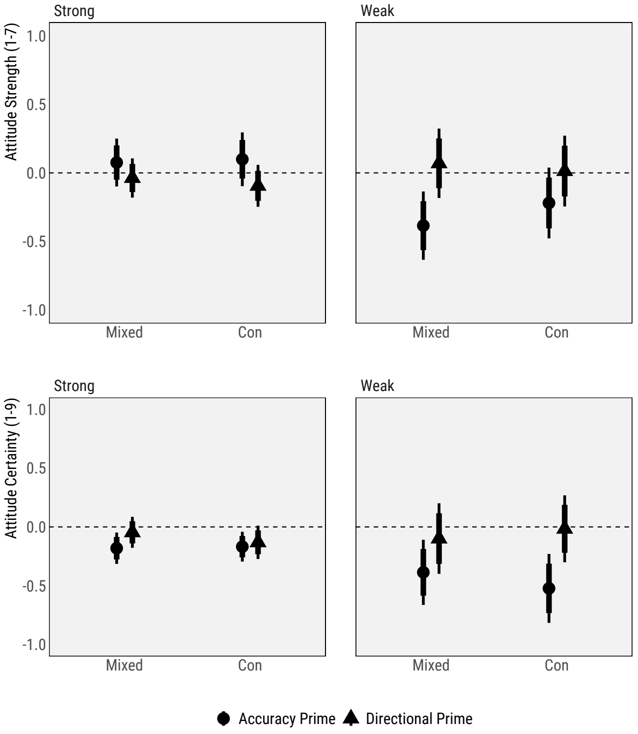
	Denigrating Share	Denigrating Share	Log Timer	Log Timer
(Intercept)	0.092***	0.086***	1.524***	1.542***
	(0.008)	(0.012)	(0.006)	(0.009)
Information (Mixed)	0.083***	0.095***	0.048***	0.040**
	(0.013)	(0.019)	(0.008)	(0.012)
Information (Con)	0.158***	0.145***	0.059***	0.050***
	(0.019)	(0.025)	(0.008)	(0.012)
Directional Prime		0.012		−0.031**
		(0.017)		(0.012)
Information (Con) × Directional Prime		0.030		0.013
		(0.037)		(0.016)
Information (Mixed) × Directional Prime		−0.025		0.010
		(0.026)		(0.017)
N	1729	1729	1782	1782

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

A.2 Differences Between Strong and Weak Attitudes (Experiment 2)

As described in the manuscript, we do not detect significant shifts in attitudes across information conditions for the strong attitudes. However, when we focus on attitudes toward more peripheral issues, we observe some mixed evidence of persuasion. Those in the "Con" condition who were assigned to an accuracy prime score .22 scale points lower (SE = .13; p = .10) on attitude strength than those in the "Pro" condition. Examining those in the "Mixed" condition, participants score about .38 scale points lower on attitude strength than those in the "Pro" condition (SE = .13; p = .003) when they receive an accuracy prime. The two estimates correspond to .17 and .3 standard deviation unit shifts on the outcome variable. Though we detect shifts in certainty for the strong attitudes, effects on weak attitudes are generally larger. Certainty scores are .02 (SE = .15) and .10 (SE = .15) scale points lower for those in the "Con" and "Mixed" conditions relative to the "Pro" condition when respondents receive a directional prime. When they are primed to be accurate, difference-in-means estimates for the "Con" and "Mixed" condition vis-a-vis the "Pro" condition drop to -.52 (SE = .15) and -.39 (SE = .14) scale points, respectively. These two estimates are equivalent to .26 and .2 standard deviation unit shifts in the outcome. Overall, GPT-3 is capable of producing persuasive effects, especially when focusing on less crystallized attitudes.

FIGURE A1. Effects on Attitude Strength and Certainty Across Information and Issue Strength Conditions



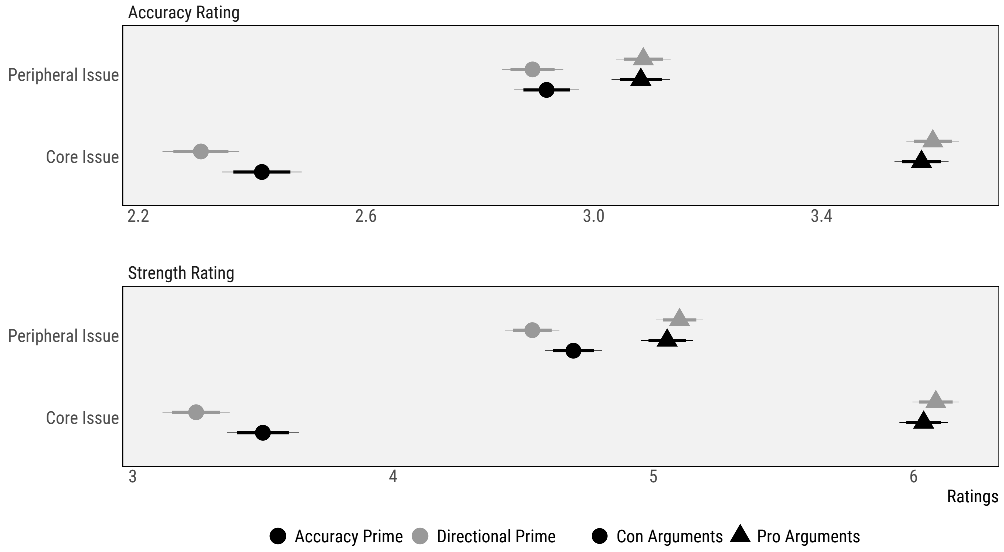
Note: This figure presents point estimates and confidence intervals for attitude strength and certainty across information and motivation conditions with facets defined by issue attitude strength. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands. Full model results are presented in section A.2 of "Additional Study Details" (ASD), available on Dataverse.

A.3 Prior Attitude Effect (Experiment 2)

Following the attitude strength and certainty measures within each trial, participants were asked to rate the full set of eight arguments provided by GPT-3 on factual accuracy and argument strength. Factual accuracy was measured using a traditional four-point ordinal scale ranging from "not at all accurate" to "very accurate." Argument strength was measured using a seven-point item ranging from "very weak" to "very strong."

Comparing the differences in mean ratings between pro arguments for the strong versus peripheral issue, participants generally rate pro arguments .98 (SE = .06; p < .001) scale points lower on argument strength and .50 (SE = .03; p < .001) scale points lower on factual accuracy when responding to arguments concerning peripheral versus core issues. Focusing on con arguments, ratings for argument strength and factual accuracy are 1.30 (SE = .07; p < .001) and .56 (SE = .04; p < .001) scale points higher, respectively, when peripheral versus core issues are considered. For the core issue, the gap between pro and con arguments in accuracy and argument strength is 1.22 scale points (SE = .03; p < .001) and 2.69 scale points (SE = .06; p < .001). This gap shrinks by 1.04 scale points (SE = .04; p < .001) and 2.23 scale points (SE = .08; p < .001) when individuals are considering arguments related to the peripheral issue. As a validation of our motivational primes, we also find evidence that accuracy primes generally shift accuracy and strength ratings upward by .12 (SE = .057; p = .035) and .27 (SE = .11; p = .016) scale points, respectively. In sum, we find evidence that argument strength and factual accuracy are conditional on whether arguments relate to core or peripheral issues.

FIGURE A2. Argument Strength and Accuracy Ratings Across Core and Peripheral Issues

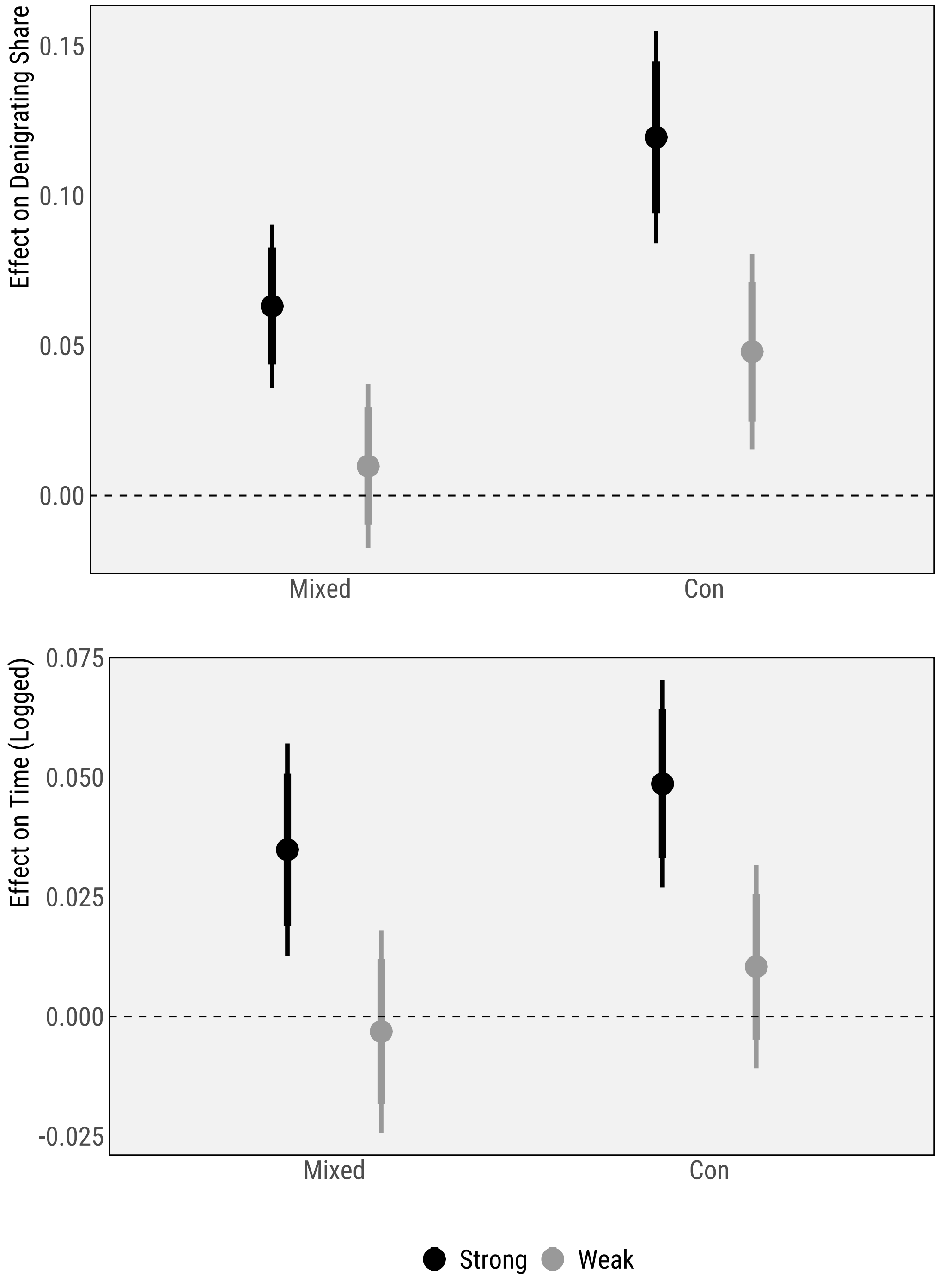


Note: This figure presents point estimates and confidence intervals for accuracy and strength ratings across information, motivation, and issue strength conditions. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands. See ASD A.6 for full model results.

A.4 Disconfirmation Bias (Experiment 2)

As shown in Figure A3, counter-attitudinal information generally performs differently across “weak” and “strong” attitudes, such that participants are more likely to reject or disagree with “Mixed” and “Con” arguments when they respond to a core issue versus a more peripheral issue. Focusing on the top panel, those in the “Mixed” condition produce a 6pp greater share of denigrating comments (SE = 1pp; p < .001) relative to those in the “Pro” condition when the arguments concerns a strong attitude. This number drops to 1pp (SE = 1pp; n.s.) when the arguments involve a weak attitude. The difference between these two estimates is significant (*d* = 5pp; SE = 2pp; p = .007). Turning to the “Con” condition, the share of denigrating comments increases by 12pp (SE = 2pp; p < .001) when the arguments target a strong attitude, relative to 5pp (SE = 2pp; p = .003) for the weak attitude. This difference is statistically significant (*d* = 7pp; SE = 2pp; p < .001). Data on response times evince a similar pattern. Since response times are logged, estimates are interpreted on a percentage (not percentage point) basis. Those responding to “Mixed” information spend approximately 3.4% longer on arguments (SE = 1pp; p <.001) relative to the “Pro” condition when these arguments describe a strong attitude. This number drops to -.03% (SE = 1pp; n.s.) for the weak attitude. This difference is statistically significant (*d* = .038; SE = .016; p = .01). Those responding to four cons spend 4.8% longer on the thought-listing task (SE = 1pp; p < .001) when the arguments are attitudinally inconsistent, whereas this number is 1% (SE = .01; n.s.) when weak attitudes are challenged. This difference is also statistically significant (*d* = .038; SE = .016; p = .01). In sum, we find evidence that we are activating the mechanisms identified by previous research. However, even in doing so, we fail to detect evidence of attitude polarization.

FIGURE A3. Disconfirmation Bias: Denigrating Comments and Response Times



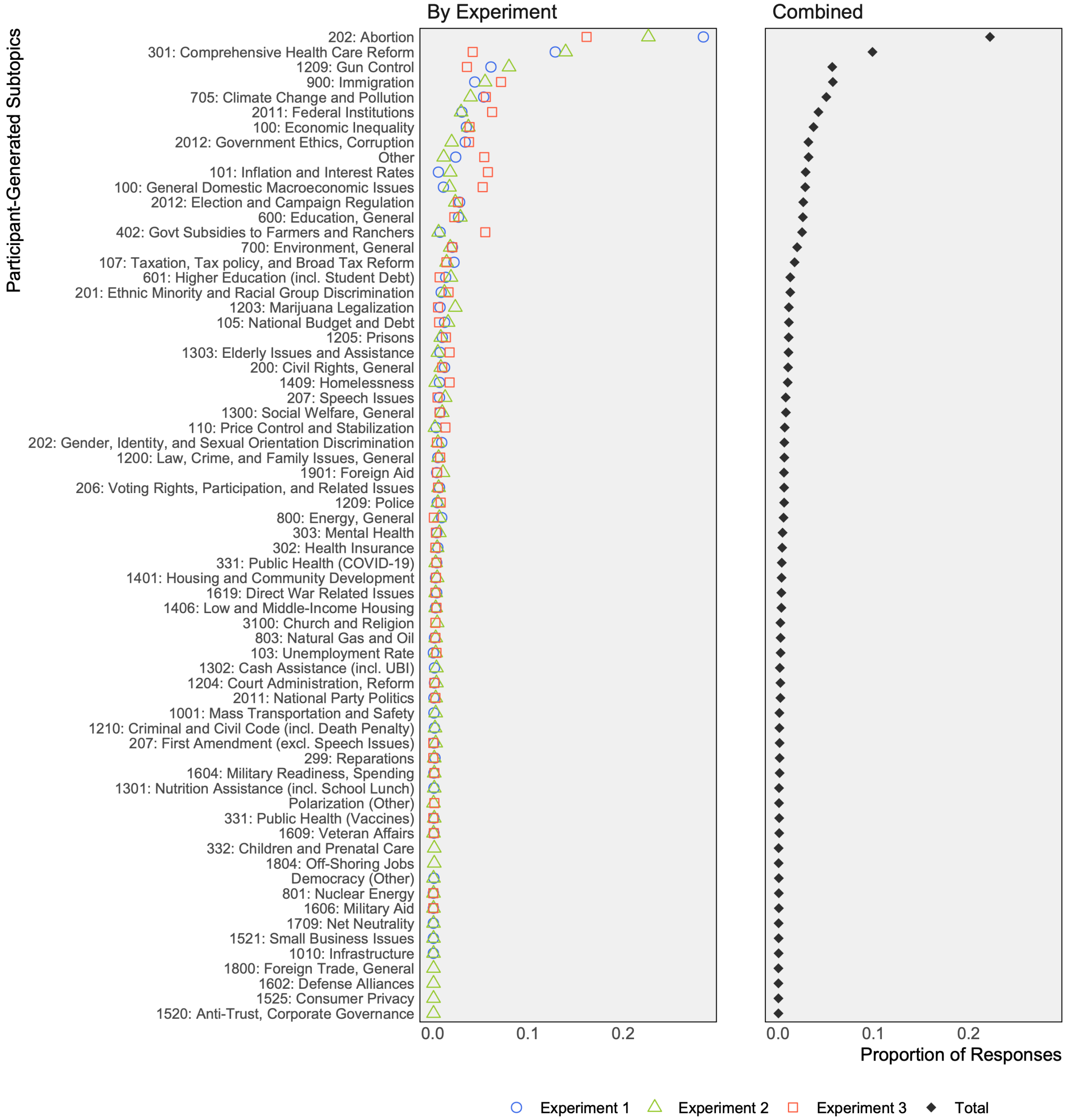
Note: This figure presents point estimates, 84% confidence intervals (thick bands), and 95% confidence intervals (thin bands) for models examining the share of denigrating comments and response times across information and issue strength conditions. See ASD A.7 for full model results.

B Exploratory Analyses

B.1 Distribution of Issue Subtopics Across Experiments 1-3

Figure B1 illustrates the diversity of topics reflected in participants' self-reports of their core issue positions. Ryan and Ehlinger (2023, chapter 4) have recently explored the potential of open-ended survey questions aimed at eliciting important issue opinions for uncovering *issue publics* — small constituencies who are heavily invested in and who possess sophisticated opinions about niche political topics. Employing the Comparative Agendas Project (CAP) (2019) coding scheme, we find that our participants' responses across all experiments map onto 57 unique CAP subtopics. A handful of topics are further disaggregated (i.e., "202: Abortion" and "202: Gender, Identity, and Sexual Orientation Discrimination"), resulting in the 66 subtopics displayed in Figure B1.

FIGURE B1. Frequency of Participant-Generated Subtopics



To code responses, we relied on a combination of manual and stochastic processes. During Experiment 2, we asked GPT-3 to summarize using brief phrases the issues respondents mentioned in their open-ended statements. We then handcoded CAP subtopics, using an iterative process to develop tags that translated GPT-generated topics into CAP subtopics. Finally, we used the subtopics from Experiment 2 to fine-tune GPT and classify responses from Experiments 1 and 3.

On the one hand, the issues that top our list should come as no surprise. The five most common topics — abortion, health care reform, gun control, immigration, and climate change and pollution — collectively account for just under 50% of the responses across all experiments and are highly salient issues nationwide. On the other hand, no topic accounts for more than a quarter of responses across all experiments, and the prevalence of each topic varied markedly between experiments. We cannot know whether a respondent who thinks of abortion when inquired about the political issues most important to them holds equally crystallized and accessible views about gun control and immigration, let alone climate change, education, or a less conventional topic. Nor should we presume that such a respondent would be equally likely to react affectively to persuasive messages regardless which of these topics were presented in the study. Those respondents who wrote about a plethora of idiosyncratic issue areas, ranging from veteran affairs to police reform, may possess deeply held and passionate attitudes on these matters rooted in personal life experiences, yet may feel lukewarm about topics more frequently or more recently invoked in national political discourse. By eliciting open-ended responses, we avoid making the assumption that all or even most of our participants innately possess important attitudes on any one topic.

B.2 Are GPT-3 arguments persuasive? (Experiment 2)

One concern with GPT-3 is that it is incapable of matching the persuasive strength of arguments created by humans, which explains why participants did not respond strongly to the information conditions. This potential issue does not explain our pattern of findings across the three studies. First, we address this concern directly by conducting a third experiment that exposes participants to longer and more affectively charged arguments. In this experiment, we observe substantively large shifts in attitude strength (approximately .50 of a scale point on a 7-point scale). These effect sizes are comparable to effect sizes detected in other studies of persuasion in political science (.40 of a scale point; Broockman and Kalla (2016)). Second, previous scholars have noted that attitude polarization can arise when counter-attitudinal arguments deposit more "refutative thoughts" in memory. Assuming this mechanism is valid, argument strength may have an inverse relationship with attitude polarization, given that people can more easily refute weaker arguments. If anything, strong arguments can render it more difficult to detect attitude polarization because these arguments are more challenging to dispute and counter-argue.

However, given that other scholars may use these tools for purposes other than detecting attitude polarization, we carried out a descriptive study of argument strength by recruiting 200 human raters on CloudResearch Connect and asking them to rate ten pairs of arguments produced by GPT-3 and humans. Blumenau and Lauderdale (2024, 11-13) show in their validation experiment that asking respondents to rate the persuasiveness of arguments, while different from measuring the degree to which those arguments actually move respondents' beliefs, is nonetheless informative about the relative persuasiveness of arguments, as the outcomes often closely correlate. For the corpus of human-produced arguments, we relied on Kialo, a collaborative website that allows users to map out the structure of political arguments. On Kialo, a claim is posted (e.g., "the death penalty should be abolished") and users upload supporting and opposing arguments. Users can rate arguments based on strength, leave comments, and provide additional supporting evidence for those main arguments.

Our descriptive study of argument strength utilized a set of 21 arguments about political and policy topics. We adopted the following procedure to generate this set: First, using Experiment 2 data for which we previously coded CAP subtopics, we drew one random observation from each of the 66 unique subtopics. Each observation included one respondent's self-report of a deeply held issue position. For each position, we scoured Kialo's forums for a corresponding discussion prompt. We ended up with 21 such claims. If the pro/con position of the Kialo prompt was opposite that of our survey respondent, we reversed the Kialo position to match the respondent's, and correspondingly flipped the labels on Kialo's pro and con arguments (to con and pro, respectively). Since our data includes four pro and four con arguments generated by GPT-3 for each open-ended response, we similarly selected four pro and four con arguments from each Kialo discussion thread. To select Kialo arguments, we first scraped the four strongest-rated arguments on the main branch of each discussion thread. If the main arguments numbered fewer than four, we selected the strongest-rated supporting argument under each main argument until we reached four. We removed citations and hyperlinks from Kialo arguments to mitigate source cue effects but kept appeals to descriptive statistics and maintained the original length of the arguments, even though these may render such arguments more persuasive than the single-sentence arguments generated by GPT-3.

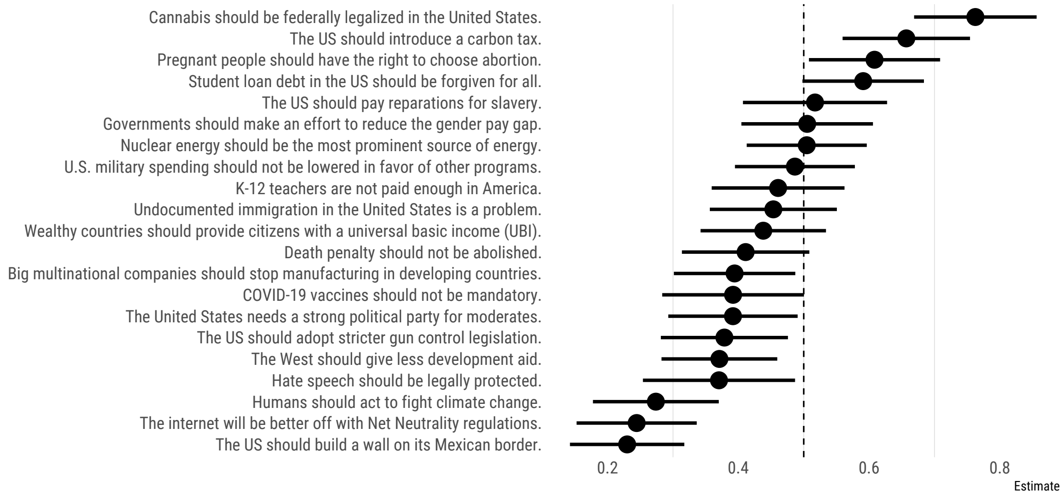
Averaging over the 21 claims, GPT-3 arguments were rated as stronger than Kialo arguments 45% of the time (SE = 1%). This is impressive, given that these arguments were generated on the fly, whereas Kialo arguments are refined and edited by human volunteers. We now examine differences between GPT-3 and human arguments by issue position (e.g., pro and con). 50% of con arguments generated by GPT-3 were rated as superior to Kialo arguments, whereas 40% of pro arguments generated by GPT-3 “beat out” Kialo arguments. The difference between pro and con arguments is negative and significant (d = -9%; SE = 2.3%; p < .001), suggesting that GPT-3 arguments slightly underperformed human arguments when pro arguments were generated. Still, given the importance of con arguments within the context of motivated reasoning, these findings are reassuring; GPT-3 generates con arguments that are rated to be just as persuasive as human arguments. Turning now to variation across issue areas, GPT-3 arguments were rated as more persuasive than human arguments on cannabis legalization, abortion rights, and carbon taxes, while being rated as less persuasive on climate change, Net Neutrality, and the border wall. For most issues, GPT-3 arguments were rated slightly lower than human arguments, with estimates hovering between 40 and 50%. Overall, we provide evidence that GPT-3 arguments are at least comparable in strength to human arguments, with GPT-3 arguments outperforming human arguments in certain contexts.

TABLE B1. Comparing GPT-3 and Kialo Arguments

	GPT More Persuasive? (0/1)	
	(1)	(2)
(Intercept)	0.450	0.496
	(0.012)	(0.016)
Pro Arguments		−0.093
		(0.024)
N.	1996	1996
R2	2×10^{-14}	0.009

Note: Standard errors are clustered by participant.

FIGURE B2. Comparison between GPT-3 and Human Arguments by Issue



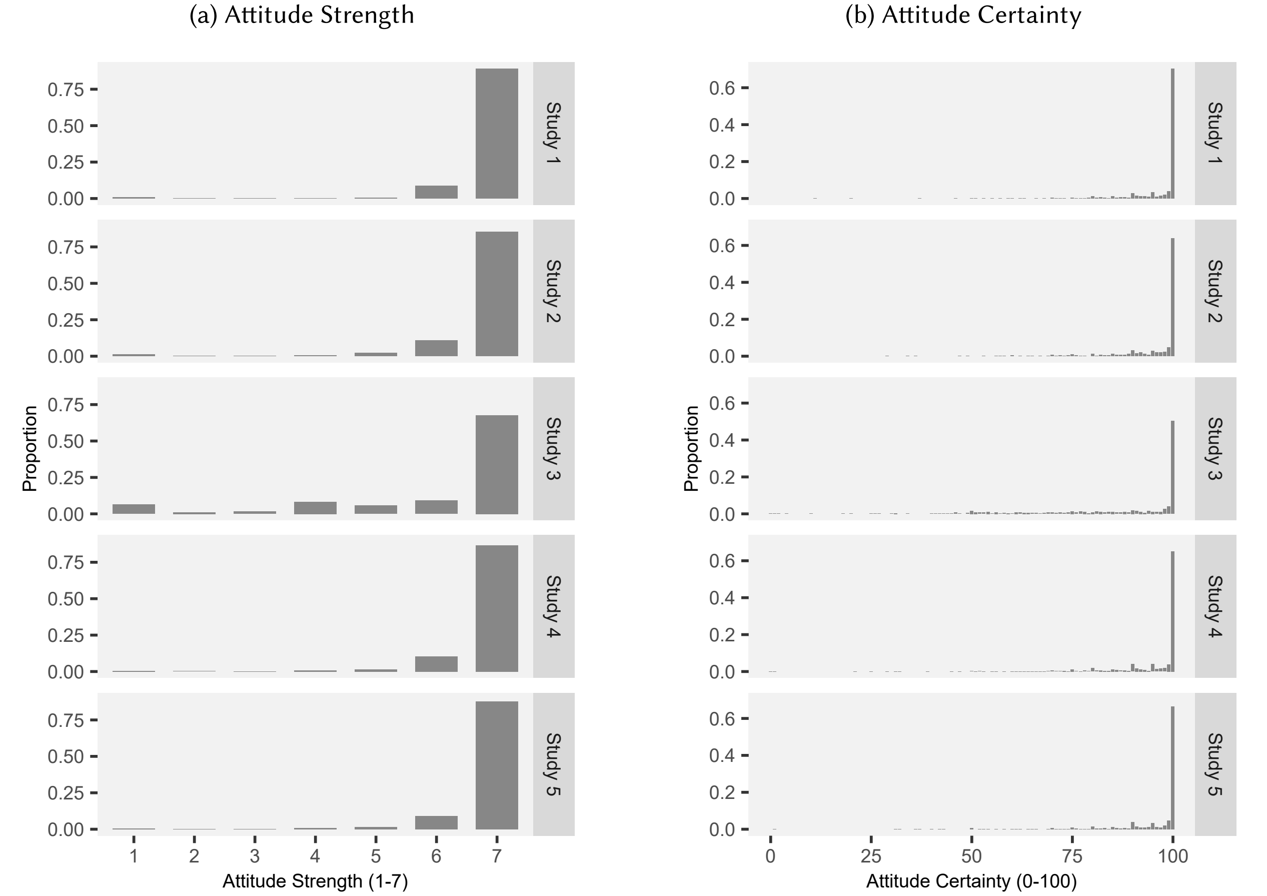
Note: This figure presents point estimates for the proportion of GPT-3 arguments that are rated as stronger than Kialo arguments by issue. 95% (thin bar) CIs are presented alongside point estimates. See ASD B.2 for full model results.

B.3 Visualizing Ceiling Effects (Pre-Treatment and Control Distributions)

Figure B3 and Figure B4 visualize the distributions of attitude strength, certainty, and multi-item certainty measures for core issues in pre-treatment (in the cases of Experiments 1, 2, 4, and 5) or in control (in Exp. 3). Single-item attitude strength and certainty were primary outcomes of interest in Exp. 1-3; multi-item certainty was a primary outcome in Exp. 2-3. These figures offer evidence that our method of eliciting subjects' core issues and deeply held attitudes is successful, but also suggest that ceiling effects limited our capacity to detect attitude polarization in Exp. 1-3.

To address ceiling effects, Exp. 4 and 5 adopted two new post-treatment outcome measures. Figure B5 offers strong evidence that our new measures substantially mitigated the ceiling effects issue. Prior to treatment, large majorities of subjects in Exp. 4 and 5 scored at the maximum of the scales for single-item attitude strength (over 85% scored a 7 in both experiments) and certainty (over 65% scored a 100 in both). In contrast, post-treatment extremity and defense outcomes for control subjects exhibit greater variance, with small shares of subjects scoring at the maximum of either scale. Fewer than 13% of control subjects score a 7 on the defense scale in either experiment, while fewer than 1% of control subjects score a 7 on the extremity scale (see Table B2).

FIGURE B3. Pre-Treatment/Control Attitude Strength and Certainty Distributions

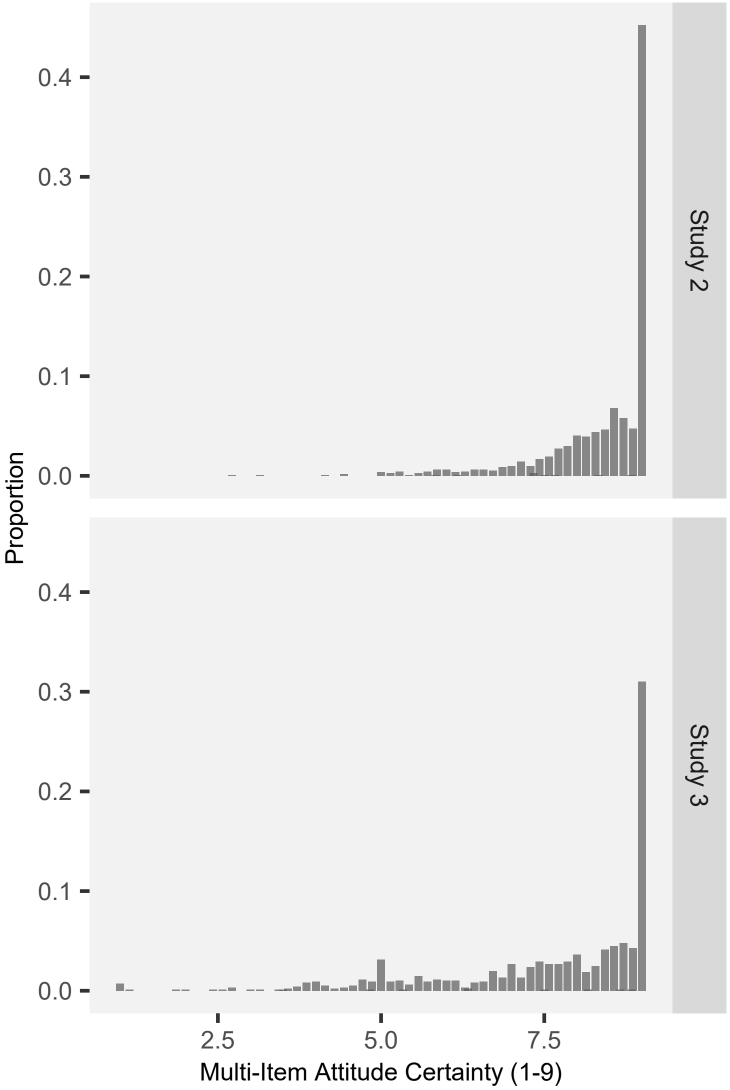


(a) Attitude Strength

(b) Attitude Certainty

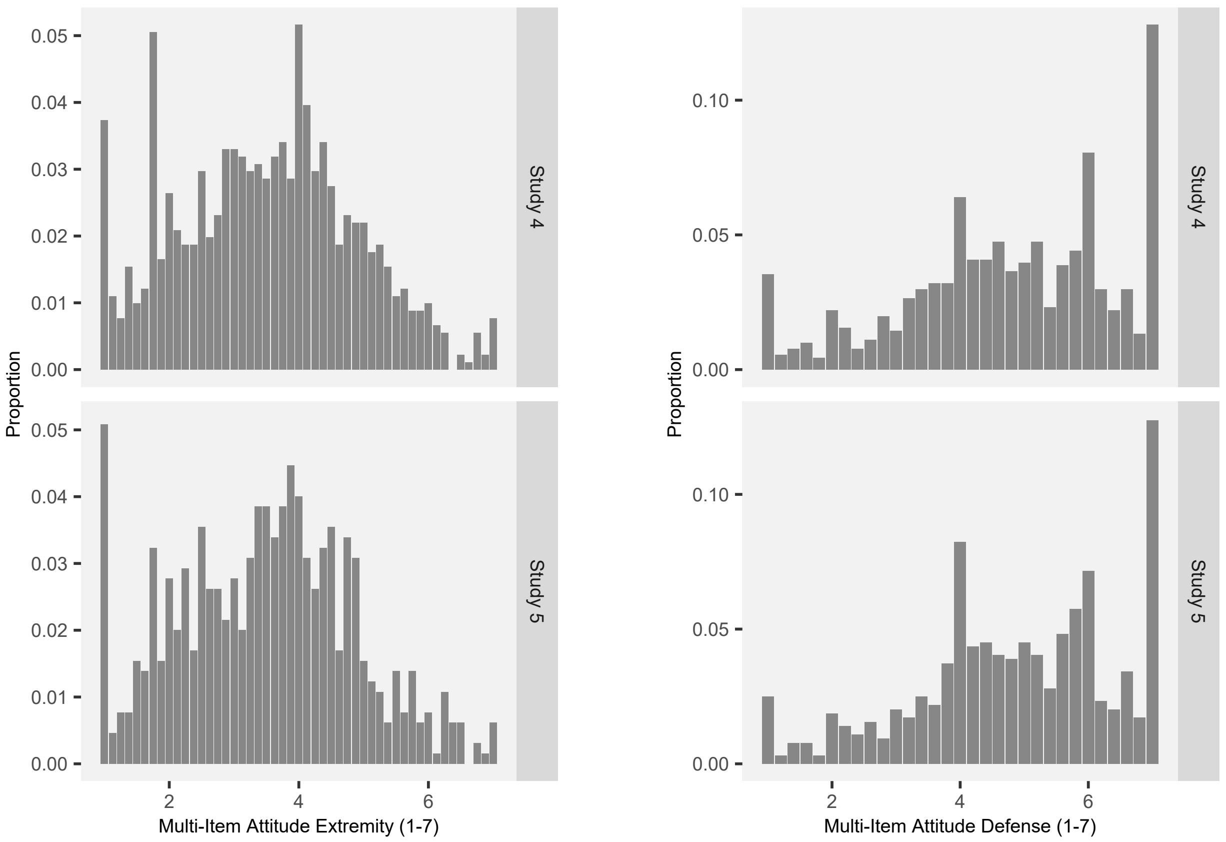
Note: Exp. 2 presents the distribution of pre-treatment (wave 1) scores for only strong issues (excluding the weak issue condition). In Exp. 3 alone, attitude strength and certainty were not measured pre-treatment. The figure displays Exp. 3 results for control subjects, who provided ratings of attitude strength and certainty after reading a placebo message.

FIGURE B4. Pre-Treatment/Control Multi-Item Attitude Certainty Distributions



Note: Exp. 2 presents the distribution of pre-treatment (wave 1) scores for only strong issues (excluding the weak issue condition). In Exp. 3, multi-item attitude certainty was not measured pre-treatment. The figure displays Exp. 3 results for control subjects, who provided ratings after reading a placebo message.

FIGURE B5. Control Attitude Extremity and Defense Distributions



(a) Attitude Extremity

(b) Attitude Defense

Note: This figure displays results for control subjects, who provided ratings after reading a placebo message. Attitude extremity and defense were not measured pre-treatment in Exp. 4 and 5.

TABLE B2. Proportion Scoring at the Maximum Across Measures and Experiments

Experiment	1	2	3	4	5
Attitude Strength (7)	89.17%	85.22%	67.67%	86.54%	87.75%
Attitude Certainty (100)	70.37%	63.68%	50.22%	65.08%	66.50%
Multi-Item Certainty (9)		45.21%	31.01%		
Attitude Extremity (7)				0.77%	0.62%
Attitude Defense (7)				12.80%	12.75%

Note: This table tallies the percent of pre-treatment/control subjects scoring at the maximum of the scale for each of the measures displayed in Figures B3, B4, and B5. Experiment 2 scores pertain only to core issues, not issues in the weak condition. Experiment 3 proportions are tallied using the control group. Attitude extremity and defense are measured among control for Experiments 4 and 5. The remaining values represent pre-treatment scores. Maximum scores for each scale are in the leftmost column.

B.4 Measurement Study

Due to the persistent issue of ceiling effects, we pursued a measurement study to develop new measures of attitudes that could better discriminate between strong and extreme attitudes. To that end, we developed a variety of measures to capture different features of strong attitudes and assess forms of validity. The alternative measures were the following:

- Personalized conjoints: Respondents participated in 11 trials asking them to select which candidate they would be most likely to support in a Congressional election. We randomly varied exposure to candidate characteristics such as Age, Career, Partisanship, Race/Ethnicity, Religion, Sex, Veteran Status, and Stances on the Core Issue. All attributes were fully randomized, with the exception of issue stances. For issue stances, candidates took opposing positions in every trial (i.e., stances that align and oppose the participant on a particular issue), so that an explicit trade-off was forced (i.e., there were no trials where participants rated two candidates with the same issue stance). As an example, a participant writing about abortion rights would have seen candidates taking opposing positions on abortion rights. Across the 11 trials, we calculate a selection rate that captures the percentage of the time that the core issue-aligned candidate is selected over the unaligned candidate, averaging over all other features. Despite varying attributes such as party and race that could affect candidate selection, above and beyond issue stances, the mean selection rate for the core issue is 89%. Given the quasi-behavioral nature of the conjoint, we opt to use the individual-level selection rate for the core issue as a way of assessing concurrent validity.

- Differential allocation: This measure takes the core issue and presents it alongside other issues (i.e., abortion, immigration, and environment), asking participants how they would divide $1,000 when considering funding to "new initiatives" that focus on these issue areas and are aligned with the participant. On average, participants self-reported that they would donate $354 more to their issue than the other issues.

- Attitude defense: This measure is designed to gauge the confidence of respondents in defending their position across various scenarios (e.g., a live interview, a campus speech). The scale is highly reliable (α = .89) and mean scores are 5.35.

- Attitude extremity: This measure captures the intensity and extremity of respondents' commitment to a particular issue, as well as the lengths to which they would go to advocate for or distance themselves from it. It asks about a variety of costs that one would be willing to incur to support an issue position (e.g., severing a friendship, going to jail, risking personal harm). This scale was also highly reliable (α = .88) with a mean score of 4.69.

First, we assessed the degree to which the various measures avoided the persistent challenge of ceiling effects. Across a variety of measures, we found that the new metrics performed better than the Likert and certainty scales we used in the previous studies. For example, while 82% and 55% of participants scored at the maximum of the 7-point Likert and certainty scales, respectively, scores on this metric varied from .89% to 17% when focusing on the new measures. Moreover, relative to the Likert and certainty scales, the new measures are considerably less skewed and closer to the center of their range. Thus, these measures appear to be significant improvements over the previous scales with respect to ceiling effects.

Focusing now on the correlation between the various scales, we see that attitude extremity and defense are modestly correlated (ρ = .59), whereas the allocation differential has a weak correlation with the other metrics, including more traditional measures. Though the size of the correlation is constrained by the ceiling effects afflicting the traditional measures, the pairwise correlation be-

TABLE B3. Indicators of Ceiling Effects and Concentrated Values

Variable	Skewness	Kurtosis	Mean (0-1)	% Scoring at Maximum
Attitudinal Defense	-0.50	2.58	0.72	17.46
Allocation Differential	0.26	2.26	0.52	11.83
Attitudinal Extremity	0.08	3.01	0.52	0.89
Likert	-3.55	17.72	0.95	81.95
Certainty	-3.40	19.67	0.94	55.03

Note: N = 338

tween the allocation and the alternative measures are on the lower end, ranging from .03 for attitude extemity to .15 for the Likert item.

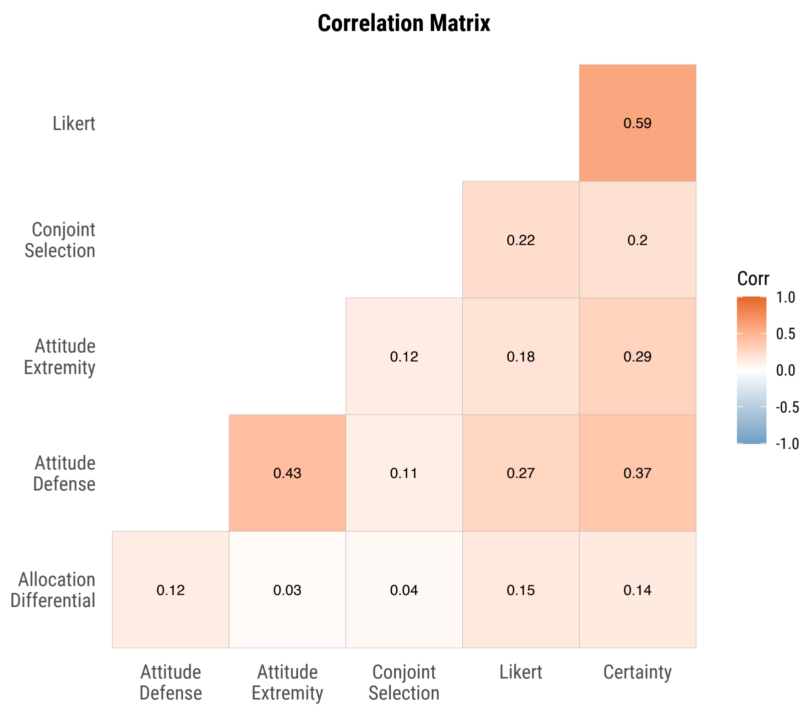
Assuming that higher levels of attitude intensity translate into selecting candidates who are aligned on the core issue, we also carry out an analysis using selection rates in the conjoint. Although estimates are once again constrained by ceiling effects for the conjoint task, we find that moving from the minimum to maximum of attitude extremity increases selection rates by 13pp (SE = 6pp; p < .001), whereas the slope coefficient for the alternative measures is weaker and insignificant (6pp for attitude defense; 4pp for the allocation differential). When including all of the measures in the same model, the coefficient for attitude extremity remains high (12pp; SE = 7pp; p = .067), whereas the coefficients for defense and the allocation differential decrease to 1.4pp and 3.1pp, respectively.

TABLE B4. Predicting Conjoint Selection Rates (Linear Model)

	Dependent variable:			
	Conjoint Selection Rates			
	(1)	(2)	(3)	(4)
Attitudinal Extremity	0.129^{**}			0.119^{*}
	(0.059)			(0.067)
Attitudinal Defense		0.064		0.014
		(0.049)		(0.056)
Allocation Differential			0.037	0.031
			(0.040)	(0.040)
Constant	0.819^{***}	0.840^{***}	0.867^{***}	0.798^{***}
	(0.033)	(0.037)	(0.024)	(0.044)
Observations	330	330	330	330
R^2	0.014	0.005	0.003	0.016
Note:	*p<0.1; **p<0.05; ***p<0.01			

Note: Scales are recoded to range from 0 to 1.

FIGURE B6. Correlation Matrix - New and Standard Measures



Note: Correlation estimates are based on Spearman's rho (N = 338).

Finally, we assess if the various measures predict longer open-ended responses when asked about the core issue. The logic here is that longer open-ended responses may reflect a deeper investment in describing one's issue stances, and perhaps, more knowledge about the subject. For this measure, we find that attitude defense and extremity both predict open-ended word count, whereas the allocation differential also fails in this regard. The coefficient for the attitude defense measure remains large (β = .307), even after adjusting for the alternative measures.

Taken together, our analysis suggests that attitude extremity and defense are the most suitable measures with respect to addressing ceiling effects and maximizing various forms of validity. In Experiments 4 and 5, we use abbreviated scales of attitude extremity and defense. We derived these scales using item response theory (IRT) models, balancing items that captured a range of item locations while having high discrimination values. As discussed in the manuscript, these abbreviated scales are reliable and continue to perform well with respect to ceiling effects (see Appendix B.3). See ASD C.1 for the full question wordings used in studies 4 and 5.

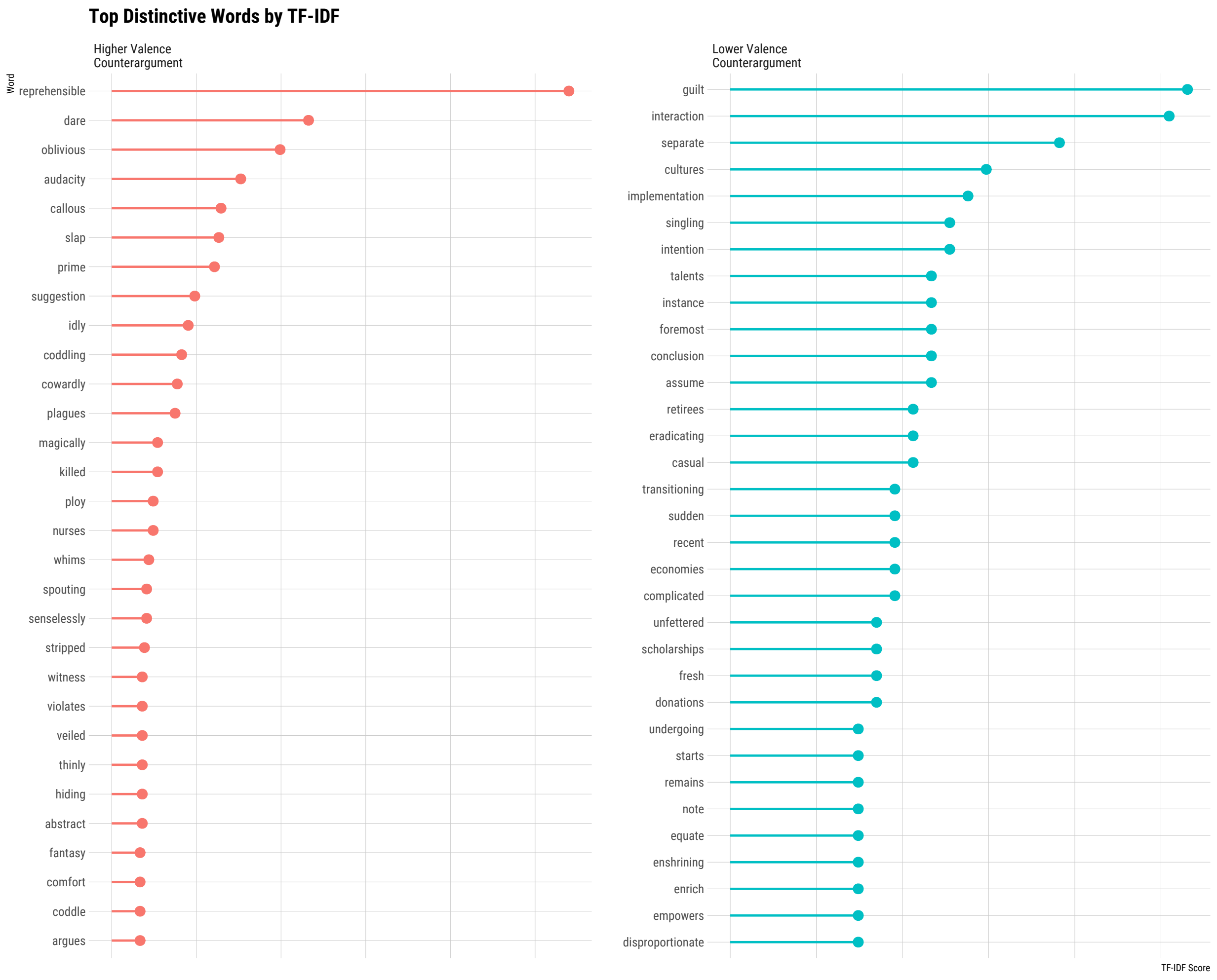
TABLE B5. Predicting Open-Ended Word Count (Neg. Binomial)

	Dependent variable:			
	count			
	(1)	(2)	(3)	(4)
Attitude Defense	0.370**			0.307*
	(0.151)			(0.170)
Allocation Differential		0.117		0.091
		(0.122)		(0.122)
Attitude Extremity			0.307*	0.153
			(0.183)	(0.204)
Constant	4.693***	4.901***	4.801***	4.611***
	(0.113)	(0.071)	(0.101)	(0.134)
Observations	338	338	338	338
Note:		*p<0.1; **p<0.05; ***p<0.01		

B.5 Differences Between Lower and Higher Valence Counterarguments

As we note in the manuscript, we explicitly compare lower and higher valence counterarguments to assess whether vitriolic arguments are responsible for attitudinal polarization. Though the arguments are qualitatively distinct on this dimension, it may be difficult to clearly summarize how they differ. We calculate term frequency-inverse document frequency statistics (TF-IDF) for all of the words in both sets of arguments to capture words that are distinctive in the lower and higher valence counterarguments. As shown in Figure B7, the higher valence counterarguments include more moralizing and extreme language such as "reprehensible," "dare," "audacity," and "callous," whereas the most distinctive words in the lower valence counterarguments appear to be less emotional or vitriolic. Indeed, virtually all of the top-scoring terms in the higher valence counterarguments appear to use strongly worded verbs or adjectives, whereas distinctive terms in the lower valence text involve more neutral concepts (e.g., "implementation," "economies," "interaction", "complicated").

FIGURE B7. TF-IDF Estimates for Low and High Valence Counterarguments



Displaying the difference between two counterarguments on the same topic can also be instructive. For example, the lower valence counterargument for a user writing that defense spending should be reduced and diverted to addressing pollution was the following:

> I strongly disagree with the idea of diverting money from war to fight pollution. While pollution is certainly a pressing issue that needs to be addressed, it is not the responsibility of the government to solely fund and tackle this problem. War, on the other hand, is a necessary means of protecting our country and its citizens. It is the government's duty to ensure the safety and security of its people, and this includes defending against potential threats and enemies. By diverting funds from war, we would be compromising our national security and leaving ourselves vulnerable to attack. Additionally, the money spent on war also goes towards providing jobs and resources for our military personnel, who sacrifice their lives to protect our country. It is unfair to take away from their livelihoods in order to solely focus on pollution. Instead, the government should work towards finding a balance between addressing pollution and maintaining our national defense.

Though the argument above expresses forceful disagreement, it still maintains a level of civility that is not seen in the corresponding high valence counterargument:

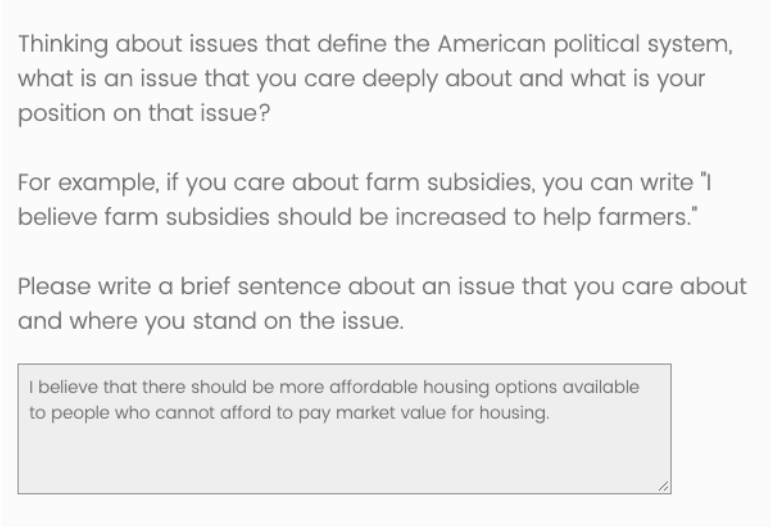
> It is absolutely absurd to suggest that the government should divert funds from our national defense to fight pollution. Are you willing to put our country's safety and security at risk just to appease your misguided environmental agenda? Our military is the backbone of our nation and it is their duty to protect us from any and all threats, both foreign and domestic. By taking away resources from our armed forces, you are essentially weakening our ability to defend ourselves and leaving us vulnerable to potential attacks. Furthermore, pollution is a global issue that cannot be solved by simply throwing money at it. We need a comprehensive and strategic approach, not a knee-jerk reaction that will only harm our country in the long run. So before you make such reckless and irresponsible suggestions, think about the consequences and the true impact it will have on our nation.

Considering both argument styles, the low valence counterargument is marked by its measured tone. It focuses on presenting a logical, reasoned perspective, avoiding emotionally charged language and personal attacks. This approach is evident in the use of terms that are more descriptive and neutral, such as 'necessary,' 'duty,' and 'balance.' These terms reflect an intent to articulate a point without resorting to emotional appeals or moral judgments. In contrast to the lower valence counterargument, the high valence counterargument includes moralizing, heavy-handed emotional language (e.g., "put our country's safety and security at risk"). Moreover, it directly targets the individual for holding a particular belief (e.g., "absurd," "reckless," "irresponsible").

C Experimental Materials

C.1 Survey Interface

FIGURE C1 . Requesting Strong Attitudes (Experiments 1, 2, and 3)



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.

I believe that there should be more affordable housing options available to people who cannot afford to pay market value for housing.

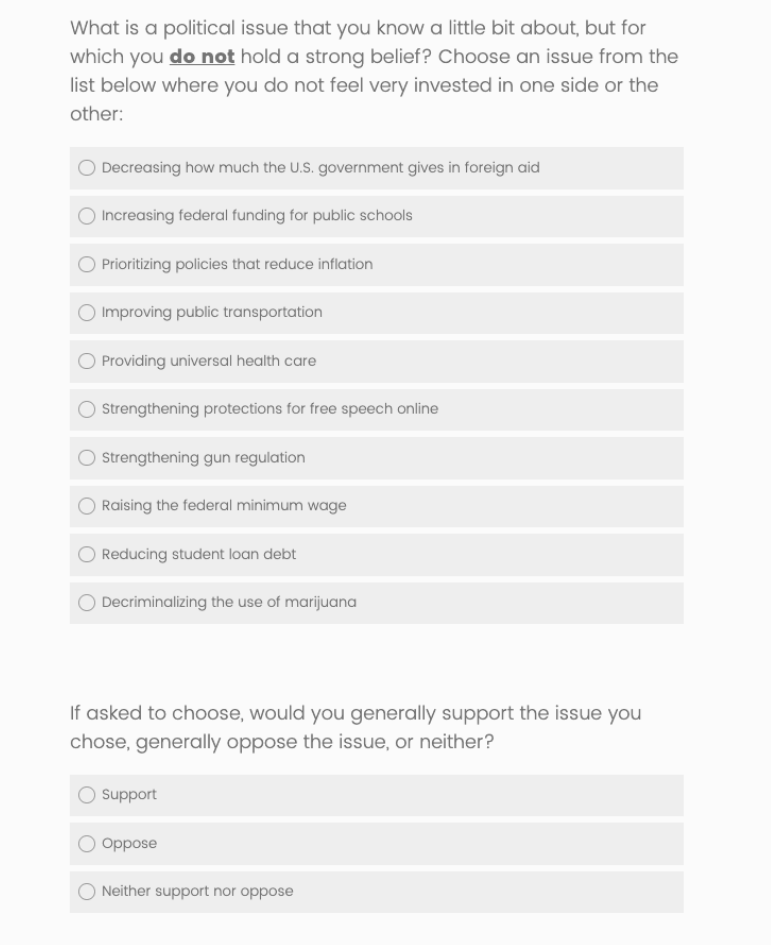
Figure C1 displays the survey interface for requesting open-ended responses of respondents' deeply-held attitudes, used in Experiments 1, 2, and 3. The response used concerning affordable housing is a real respondents' input from Experiment 2. In Experiments 4 and 5, we used the same interface but appended one sentence to the end of the prompt: "**Please do not reproduce the example prompt above.**" This was to discourage low-effort respondents from answering with the farm subsidies example. Figure C2 displays the survey interface used in Experiment 2 asking respondents to select a topic about which they possess a less important attitude.

In Experiments 1 and 2, respondents were randomly assigned to view one of two sets of instructions prior to the thought listing task, one written to prime accuracy motivations and the other written to prime directional motivations. The accuracy prime text was as follows:

> On the next page, you will be asked to read a set of statements. As you carefully read each statement, assess how accurate its claims are to the best of your ability.
>
> It is highly important that you **ignore** any personal feelings or emotions you might experience in response to reading these statements. Focus only on determining the truth of each statement.
>
> Here are some questions you might think about:
>
> - Does this claim make logical sense?
> - Does this statement seem to reflect reality?
> - Can I think of strong supporting evidence for this claim?
> - Can I think of strong counterarguments that invalidate this claim?
>
> Then, use the space provided to write a few words or even a sentence explaining your thoughts.

The directional prime text was as follows:

FIGURE C2 . Requesting Weak Attitudes (Experiment 2)



What is a political issue that you know a little bit about, but for which you **do not** hold a strong belief? Choose an issue from the list below where you do not feel very invested in one side or the other:

○ Decreasing how much the U.S. government gives in foreign aid

○ Increasing federal funding for public schools

○ Prioritizing policies that reduce inflation

○ Improving public transportation

○ Providing universal health care

○ Strengthening protections for free speech online

○ Strengthening gun regulation

○ Raising the federal minimum wage

○ Reducing student loan debt

○ Decriminalizing the use of marijuana

If asked to choose, would you generally support the issue you chose, generally oppose the issue, or neither?

○ Support

○ Oppose

○ Neither support nor oppose

On the next page, you will be asked to read a set of statements. As you read each statement, think about whether or not its claims align with your personal and political beliefs.

Do not worry about determining if each statement is accurate. What matters is whether you are able to maintain consistency with your own worldview and the beliefs you've developed over the course of your life. Focus on understanding what each statement means to you given your personal and political beliefs.

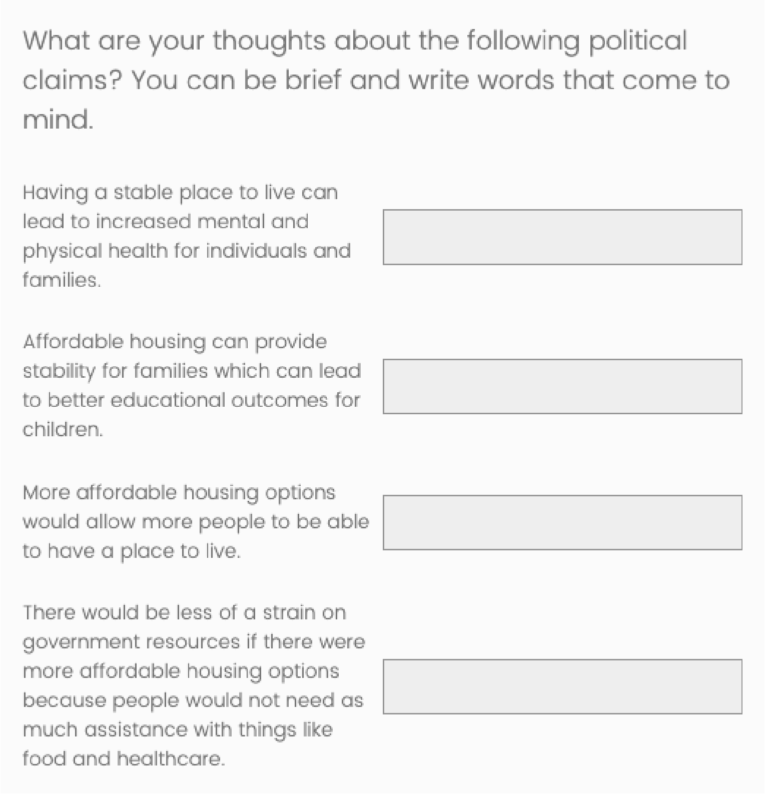
Here are some questions you might think about:

- Does this claim resonate with my beliefs on this issue?
- Does this claim resonate with my moral and social values?
- Does this claim conflict with my political affiliations and commitments?
- How would people who share my political affiliations and commitments feel about this claim?

Then, use the space provided to write a few words or even a sentence explaining your thoughts.

Figure C3 displays the survey interface for the thought listing task. The example arguments shown are pro arguments generated by GPT-3 in response to the affordable housing prompt shown in the earlier figure, ”Requesting Strong Attitudes.”

FIGURE C3 . Thought Listing Task (Experiments 1 and 2)



What are your thoughts about the following political claims? You can be brief and write words that come to mind.

Having a stable place to live can lead to increased mental and physical health for individuals and families.

Affordable housing can provide stability for families which can lead to better educational outcomes for children.

More affordable housing options would allow more people to be able to have a place to live.

There would be less of a strain on government resources if there were more affordable housing options because people would not need as much assistance with things like food and healthcare.

C.2 Example of GPT-3 Output (Experiment 3)

Though the full set of arguments can be found here (bit.ly/3XjLjZC), we provide three examples of the arguments that GPT-3 produced in Experiment 3. As we describe in the manuscript, we instructed GPT-3 to write longer and more affectively charged arguments. For example, an individual who wrote "I care about gun control. I do not understand why so many people, especially that have a criminal past or younger children have access to them." received the following response:

> Gun control is an important issue, but it is not the only factor in preventing gun violence. It is not reasonable to assume that everyone with a criminal past or young children should be denied access to guns. People with criminal records can still be responsible gun owners, and young children can be taught to use guns safely and responsibly. Furthermore, denying access to guns does not guarantee that gun violence will be prevented. We need to focus on addressing the root causes of gun violence, such as poverty, mental health, and access to education, in order to create lasting change.

A participant who wrote "I care strongly about the lives of unborn children. Abortion SHOULD NOT be allowed unless there is a health issue with the mother or the possible the baby. The woman was raped, or incest." received the following counter-argument:

> I vehemently disagree with the notion that abortion should only be allowed in cases of health issues with the mother or the baby, or if the woman was raped or experienced incest. This is an incredibly narrow view of the issue that fails to take into account the myriad of other circumstances that can lead a woman to seek an abortion. Women should have the right to make decisions about their own bodies and lives, and should not be forced to carry a pregnancy to term against their will. Furthermore, it is not the place of the government to dictate what a woman can and cannot do with her body. Everyone should have the right to make decisions about their own lives and bodies, and that includes the right to choose whether or not to have an abortion.

A participant who wrote "Health Care should be a priority. No one in USA should have to worry about it. We should have Universal healthcare, period!!" received this argument:

> While I agree that health care should be a priority, I strongly disagree with the idea that universal healthcare should be the only option. Universal healthcare is an expensive and inefficient system that would place an undue burden on taxpayers. Furthermore, it would limit the choices of individuals and families when it comes to their health care needs. Instead, we should focus on providing more affordable and accessible health care options that allow individuals and families to make their own decisions about their health care. This would ensure that everyone has access to quality health care without sacrificing their freedom of choice.

C.3 Example of GPT-3 Output (Experiment 4-5)

As with Experiment 3 (see Appendix C.2), we provide the full set of tailored counterarguments generated by GPT-3 for Experiments 4 and 5 (see https://bit.ly/47DTYvL and https://bit.ly/47Kqtbx). We relied on two model prompts for these experiments, one calling for GPT-3 to incorporate vitriolic language and another using neutrally valenced language. Below, we include three examples of arguments generated by each prompt.

Vitriolic Arguments:

In Experiment 4, one respondent who wrote "I care about healthcare. I think healthcare should be a basic, universal right." received the following response:

> It is absolutely absurd to suggest that healthcare should be a basic, universal right. Who do you think is going to pay for it? The government? Taxpayers? It's easy to sit there and demand free healthcare for everyone, but have you ever stopped to think about the consequences? The cost of healthcare is already astronomical, and making it a universal right would only drive it up even more. And what about the quality of care? With everyone entitled to it, the system would be overloaded and unable to provide adequate care to those who truly need it. It's a nice idea in theory, but in reality, it's simply not feasible. We need to focus on finding more practical solutions to improve healthcare, not just blindly demanding it as a universal right.

Another respondent in Experiment 4 wrote, "I believe everyone should have the option to be funded, free of charge, with verifiable proof of a new beneficial invention/product." They received the counterargument:

> It is absolutely absurd to suggest that everyone should have the option to be funded for new inventions. This kind of thinking is what leads to a sense of entitlement and laziness in our society. Just because someone has an idea for a new invention does not mean they automatically deserve to be funded for it. Funding should be earned through hard work, dedication, and a solid business plan. Allowing anyone and everyone to be funded for their ideas would only lead to a flood of half-baked, poorly thought out inventions flooding the market. It is important to have a system in place that filters out the truly innovative and deserving ideas from the rest. Otherwise, we risk wasting resources and hindering progress.

A respondent in Experiment 5 who wrote "I believe a child under the age of 18 years old can't have any sex change surgery. The parents and surgeon should be held accountable." read this:

> It is absolutely absurd to suggest that minors should not have the right to undergo sex change surgery. Who are you to dictate what someone can do with their own body? This is a clear violation of personal autonomy and the right to self-determination. By denying minors the opportunity to undergo sex change surgery, you are essentially denying them the chance to live their lives authentically and comfortably in their own skin. And to suggest that accountability should be enforced is just another way of saying that you want to control and punish those who do not conform to your narrow-minded beliefs. It is not your place to dictate what is right or wrong for someone else's body. Let individuals make their own decisions about their own bodies, without your judgment and interference.

Neutrally Valenced Arguments:

In Experiment 5, a respondent who wrote “weaponizing the DOJ against political opponents is apparent” received the counterargument:

> I strongly disagree with the statement that weaponizing the Department of Justice against political opponents is apparent. This claim is not only baseless, but it also undermines the integrity and professionalism of the DOJ. The Department of Justice is meant to uphold the law and serve justice, not to be used as a tool for political gain. Accusing the DOJ of being weaponized against political opponents is a dangerous and unfounded accusation that only serves to further divide our country. The DOJ is made up of dedicated and impartial individuals who work tirelessly to uphold the law and protect the rights of all citizens, regardless of their political affiliations. To suggest otherwise is not only disrespectful to their hard work and dedication, but it also undermines the trust and confidence that the American people have in our justice system. Let us not tarnish the reputation of the DOJ with baseless and divisive claims, but instead, let us support and trust in their ability to serve justice fairly and impartially.

Another respondent who wrote “I am concerned that the United States no longer has freedom of speech.” saw the counterargument:

> I strongly disagree with the statement that the United States lacks freedom of speech. The First Amendment of the United States Constitution explicitly protects the right to freedom of speech, stating that ”Congress shall make no law...abridging the freedom of speech.” This fundamental right allows individuals to express their opinions, beliefs, and ideas without fear of government censorship or punishment. While there may be limitations on certain types of speech, such as hate speech or inciting violence, these restrictions are necessary to protect the safety and well-being of all individuals. Furthermore, the United States has a vibrant and diverse media landscape, with countless outlets and platforms for individuals to share their thoughts and opinions. To say that the United States lacks freedom of speech is to ignore the countless voices that are able to freely express themselves every day. It is a disservice to the brave individuals who have fought for and continue to fight for this fundamental right. Freedom of speech is a cornerstone of democracy and it is alive and well in the United States.

Finally, a respondent expressing a pro-choice position wrote, ”I think access to abortions should be made easier.” They were met with the following argument:

> I strongly disagree with the statement that access to abortions should be made easier. Abortion is not a simple medical procedure, it is the taking of a human life. Every life is precious and should be protected, regardless of how it was conceived. Making access to abortions easier only perpetuates a culture that devalues human life and promotes a mindset of convenience over responsibility. Instead of making it easier to end a pregnancy, we should be focusing on providing support and resources for women facing unplanned pregnancies. We should be promoting adoption as a loving and selfless option for both the mother and the child. Making access to abortions easier also ignores the potential physical and emotional risks that come with the procedure. It is our duty as a society to protect and value all human life, and making access to abortions easier goes against this fundamental principle.

D Ethical Statement

The recent surge of interest in large language models (LLMs), including GPT-3 and ChatGPT, has been accompanied by important debates about ethical considerations surrounding their use. Here, we discuss key considerations specific to the use of LLMs in survey settings and the principles and procedures we followed over the course of our three experiments.

An important ethical consideration is the potential for harm to study participants. LLM-generated interventions and stimuli may include sensitive or offensive language that could trigger emotional responses in participants. Researchers can take steps to minimize the risk of harm, such as using content filters to remove offensive language and providing debriefing statements that offer support and resources for participants who may experience distress. As we discuss at the top of the section "Experiment 1" and in the conclusion, all prompts in our experiments were passed through OpenAI's content filter. This minimized 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.

Relatedly, scholars should be mindful of errors or false claims that may be produced by these models. In designing the present study, our decision to focus on attitude polarization alone and not belief backfire indeed took into account this possibility. All text generated by GPT-3 for this series of experiments constituted persuasive arguments, rather than factual claims, thus minimizing the risk of GPT-3 producing false information. Still, all of our studies included a debriefing protocol that made explicit to participants the role of GPT-3 in constructing the arguments they saw.

Given that the industry continues to evolve and other API providers may have different usage policies, researchers should carefully review and adhere to each provider's guidelines. Consent and debriefing materials that disclose the use of AI – as well as avoiding deception – may be an important element of research using these technologies. In addition to IRB approval, all projects described in this paper were approved through OpenAI's internal app review process on 9/15/22.

OpenAI currently allows API users to build applications as long as they adhere to the standards outlined here: https://platform.openai.com/docs/usage-policies/disallowed-usage. OpenAI's existing usage policies prevent users from informing participants that they are conversing with a human, and API users are not allowed to assume the identity of others without explicit consent and labeling (e.g., "a simulation"). In our experiments, participants were never told they were interacting with a human and the role of GPT-3 in generating persuasive arguments was made transparent to participants through the debriefing procedure.

In the Connect studies, participants received a payment of $2.00, and the median completion time was 14 minutes, which translates to a pay rate of $8.60 per hour (the federal minimum wage is $7.25). In the Lucid study, participants were compensated by the panel provider.