Additional Study Details for Velez and Liu (2024)

A Pre-Registered Analyses **1**
- A.1 Full Model Results (Experiment 1) . . . 1
- A.2 Full Model Results (Experiment 2) . . . 2
- A.3 Full Model Results (Experiment 3) . . . 3
- A.4 Full Model Results (Experiment 4) . . . 4
- A.5 Full Model Results (Experiment 5) . . . 6
- A.6 Full Model Results (Prior Attitude Effect, Experiment 2) . . . 7
- A.7 Full Model Results (Disconfirmation Bias, Experiment 2) . . . 8

B Exploratory Analyses **9**
- B.1 Change Score Analysis (Experiment 2) . . . 9
- B.2 Are GPT-3 Arguments Persuasive? Full Results . . . 13
- B.3 GPT-3 Completion by Issue Topic (Experiment 2) . . . 14
- B.4 Certainty Analysis (Experiment 3) . . . 15
- B.5 Robustness of Effects (Experiment 3) . . . 16
- B.6 Consistency of GPT Output . . . 17
- B.7 Implications of GPT-3 Formatting Errors . . . 20
- B.8 CATEs by Subtopic . . . 25
- B.9 Assessing Post-Treatment Attrition (Experiments 4 and 5) . . . 35

C Experimental Materials **38**
- C.1 Question Wording . . . 38
- C.2 GPT-3 Details . . . 40
- C.3 Sample Composition . . . 42

D AsPredicted Pre-Registration Plans **43**
- D.1 Pre-Registration for Experiment 1 (AsPredicted #108182) . . . 43
- D.2 Pre-Registration for Experiment 2 (AsPredicted #109299) . . . 45
- D.3 Pre-Registration for Experiment 3 (AsPredicted #116238) . . . 47
- D.4 Pre-Registration for Experiment 4 (AsPredicted #146475) . . . 48
- D.5 Pre-Registration for Experiment 5 (AsPredicted #147883) . . . 50

A Pre-Registered Analyses

A.1 Full Model Results (Experiment 1)

TABLE A1. Effects of Arguments and Motivation Primes on Attitudes (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.845***	97.117***
	(0.024)	(0.164)
Pre-Treatment Strength	0.272***	0.043
	(0.064)	(0.189)
Pre-Treatment Certainty	0.182***	6.133***
	(0.033)	(0.466)
Con	-0.015	-0.025
	(0.038)	(0.359)
Mixed	-0.022	-0.004
	(0.035)	(0.230)
Directional Prime	0.012	0.053
	(0.033)	(0.213)
Con × Directional Prime	-0.007	-0.123
	(0.051)	(0.469)
Mixed × Directional Prime	-0.004	-0.900*
	(0.049)	(0.429)
N	1768	1772
R^2	0.438	0.718

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: OLS regression with HC2 robust standard errors.

A.2 Full Model Results (Experiment 2)

TABLE A2. Effects of Arguments and Motivation Primes on Attitudes (Experiment 2)

	Attitude Strength (1-7)	Certainty (0-100)	Certainty Score (1-9)
(Intercept)	5.951***	86.923***	7.703***
	(0.077)	(0.867)	(0.056)
Pre-Treatment Strength	0.804***	1.511**	0.132**
	(0.057)	(0.548)	(0.042)
Pre-Treatment Certainty	0.140**	9.973***	
	(0.046)	(0.651)	
Con	0.104	−0.396	−0.127+
	(0.103)	(0.990)	(0.067)
Mixed	0.077	−1.152	−0.184**
	(0.087)	(1.047)	(0.069)
Directional Prime	0.130	1.076	0.012
	(0.086)	(0.871)	(0.064)
Weak Attitude	−0.123	−1.613	0.041
	(0.133)	(1.709)	(0.117)
Con × Directional Prime	−0.190	−0.334	−0.023
	(0.128)	(1.236)	(0.098)
Mixed × Directional Prime	−0.113	−0.181	0.112
	(0.113)	(1.294)	(0.096)
Con × Weak Attitude	−0.321+	−3.357	−0.391*
	(0.165)	(2.193)	(0.165)
Mixed × Weak Attitude	−0.456**	−2.777	−0.199
	(0.155)	(2.162)	(0.160)
Directional Prime × Weak Attitude	−0.384*	−0.798	−0.253
	(0.161)	(2.169)	(0.159)
Con × Directional Prime × Weak Attitude	0.378+	−1.356	0.511*
	(0.223)	(3.071)	(0.230)
Mixed × Directional Prime × Weak Attitude	0.541*	1.721	0.155
	(0.216)	(2.947)	(0.230)
Pre-Treatment Certainty Scores			0.925***
			(0.050)
N	2338	2331	2357
R^2	0.485	0.413	0.456
Clustered Standard Errors	by: id	by: id	by: id

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: OLS regression with CR2 robust standard errors.

APSR Submission Template

A.3 Full Model Results (Experiment 3)

We present unadjusted and covariate-adjusted average treatment effects. In the models that estimate covariate-adjusted average treatment effects, we adjust for age (in years), education (Some high school or less, High school graduate, Other post high school vocational training, Completed some college, but no degree, Associate's degree, Bachelor's degree, Master's or professional degree, Doctorate degree), race (1 = non-Hispanic white; 0 otherwise), income ($14,999-$250,000 and above in $4,999 increments), partisanship (Strong Democrat, Not very strong Democrat, Independent Democrat, Independent - neither, Independent Republican, Other - leaning Democrat, Other - neither, Other - leaning Republican, Not very strong Republican, Strong Republican), and ideology (Very liberal, Somewhat liberal, Moderate, Somewhat conservative, Very conservative).

TABLE A3. Effects of Arguments on Attitudes (Experiment 3)

	Attitude Strength (1-7)	Attitude Strength (1-7)	Attitude Certainty (1-9)	Attitude Certainty (1-9)
(Intercept)	6.043***	6.038***	7.630***	7.630***
	(0.056)	(0.056)	(0.053)	(0.052)
Treatment	−0.487***	−0.510***	0.015	−0.005
	(0.089)	(0.089)	(0.073)	(0.071)
Age		0.152**		0.333***
		(0.049)		(0.037)
Education		0.056		0.033
		(0.050)		(0.040)
Income		0.054		0.122***
		(0.047)		(0.037)
Ideology		−0.072		−0.077
		(0.059)		(0.050)
White		0.104*		0.081*
		(0.048)		(0.040)
Partisanship		−0.074		0.005
		(0.057)		(0.048)
N	1891	1848	1895	1852
R^2	0.016	0.034	0.000 02	0.065

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: OLS regression with HC2 robust standard errors.

A.4 Full Model Results (Experiment 4)

As with Experiment 3, in the models that estimate covariate-adjusted average treatment effects, we adjust for age (in years), education (Some high school or less, High school graduate, Other post high school vocational training, Completed some college, but no degree, Associate's degree, Bachelor's degree, Master's or professional degree, Doctorate degree), income ($14,999-$250,000 and above in $4,999 increments), partisanship (Strong Democrat, Not very strong Democrat, Independent Democrat, Independent - neither, Independent Republican, Other - leaning Democrat, Other - neither, Other - leaning Republican, Not very strong Republican, Strong Republican), and ideology (Very liberal, Somewhat liberal, Moderate, Somewhat conservative, Very conservative).

In addition, we adjust for pre-treatment attitude strength (7-point Likert), pre-treatment attitude certainty (101-point scale), political sophistication, and a political participation scale (0-1). Following Pew Research Center's April 2018 report, "The Public, the Political System and American Democracy," our participation scale captures the frequency with which respondents have recently engaged in six political activities. Including this scale improves the precision of our new outcome measures in Experiments 4 and 5, which capture respondents' commitment to an attitude through their willingness to engage in public activities to advocate for or defend their issue position.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

TABLE A4. Effects of Arguments on Attitudes (Experiment 4)

	Attitude Extremity	Attitude Extremity	Attitude Defense	Attitude Defense
(Intercept)	2.20***	3.50***	−0.06	4.76***
	(0.37)	(0.05)	(0.48)	(0.05)
High Valence Counterargument	0.18**	0.15*	0.19**	0.17*
	(0.06)	(0.06)	(0.07)	(0.08)
Pre-treatment Likert	0.04		0.18*	
	(0.05)		(0.08)	
Pre-treatment Certainty	0.01**		0.03***	
	(0.00)		(0.00)	
Age	0.00		0.00*	
	(0.00)		(0.00)	
Education	−0.04		0.04	
	(0.02)		(0.03)	
Income	−0.01		0.02*	
	(0.01)		(0.01)	
Ideology	−0.04		0.07	
	(0.04)		(0.04)	
Pre-treatment Participation	2.44***		1.31***	
	(0.12)		(0.15)	
Political Knowledge	0.09		0.51**	
	(0.16)		(0.23)	
Partisanship	−0.05*		−0.02	
	(0.02)		(0.03)	
N	1814	1835	1808	1830
R^2	0.20	0.00	0.09	0.00

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: OLS regression with HC2 robust standard errors.

A.5 Full Model Results (Experiment 5)

TABLE A5. Effects of Arguments on Attitudes (Experiment 5)

	Attitude Extremity	Attitude Extremity	Attitude Defense	Attitude Defense
(Intercept)	2.14^{***}	3.49^{***}	0.69	4.85^{***}
	(0.36)	(0.05)	(0.43)	(0.06)
Low Valence Counterargument	0.09	0.07	-0.22^{*}	-0.21^{*}
	(0.07)	(0.07)	(0.09)	(0.09)
High Valence Counterargument	0.20^{**}	0.20^{*}	-0.06	-0.02
	(0.07)	(0.08)	(0.09)	(0.09)
Pre-treatment Likert	0.04		0.09	
	(0.05)		(0.06)	
Pre-treatment Certainty	0.01^{***}		0.03^{***}	
	(0.00)		(0.00)	
Age	-0.00		-0.00	
	(0.00)		(0.00)	
Education	-0.01		0.02	
	(0.02)		(0.03)	
Income	-0.00		0.01	
	(0.01)		(0.01)	
Ideology	-0.08^{*}		0.01	
	(0.04)		(0.05)	
Pre-treatment Participation	2.35^{***}		1.02^{***}	
	(0.12)		(0.15)	
Political Knowledge	-0.14		0.29	
	(0.14)		(0.17)	
Partisanship	-0.06^{*}		-0.01	
	(0.02)		(0.03)	
N	1931	1959	1901	1929
R^2	0.20	0.00	0.07	0.00

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: OLS regression with HC2 robust standard errors.

APSR Submission Template

A.6 Full Model Results (Prior Attitude Effect, Experiment 2)

TABLE A6. Estimated Marginal Means of Argument Strength and Accuracy Ratings

Outcome	Type	Issue Strength	Motivation	Mean	SE	df	*p*
Strength	Con Args	Core	Accuracy Prime	3.50	0.07	1131	0.00
			Directional Prime	3.24	0.07	1131	0.00
		Peripheral	Accuracy Prime	4.69	0.06	1220	0.00
			Directional Prime	4.53	0.05	1220	0.00
	Pro Args	Core	Accuracy Prime	6.04	0.05	1131	0.00
			Directional Prime	6.09	0.05	1131	0.00
		Peripheral	Accuracy Prime	5.05	0.05	1220	0.00
			Directional Prime	5.10	0.05	1220	0.00
Accuracy	Con Args	Core	Accuracy Prime	2.42	0.04	1130	0.00
			Directional Prime	2.31	0.03	1130	0.00
		Peripheral	Accuracy Prime	2.92	0.03	1219	0.00
			Directional Prime	2.89	0.03	1219	0.00
	Pro Args	Core	Accuracy Prime	3.58	0.02	1131	0.00
			Directional Prime	3.60	0.02	1131	0.00
		Peripheral	Accuracy Prime	3.08	0.03	1220	0.00
			Directional Prime	3.09	0.02	1220	0.00

Note: OLS regression with robust standard errors clustered at the participant level. See Appendix A.3 for discussion.

A.7 Full Model Results (Disconfirmation Bias, Experiment 2)

TABLE A7. Share of Denigrating Comments and Response Times Across Information and Issue Strength Conditions

	Denigrating Share	Time (Logged)
(Intercept)	0.053	1.500
	(0.008)	(0.008)
Con	0.120	0.049
	(0.018)	(0.011)
Mixed	0.063	0.035
	(0.014)	(0.011)
Weak	0.033	0.026
	(0.013)	(0.011)
Con × Weak	−0.072	−0.038
	(0.025)	(0.016)
Mixed × Weak	−0.053	−0.038
	(0.020)	(0.016)
N	2361	2362
R^2	0.025	0.009

Note: OLS regression with HC2 robust standard errors. See Appendix A.4 for discussion.

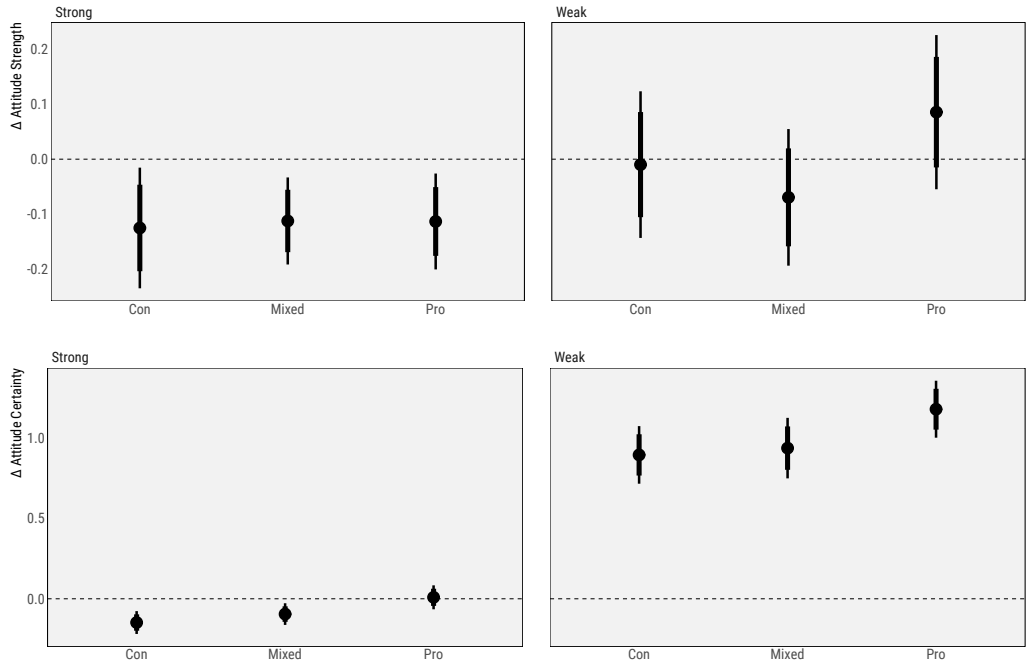
APSR Submission Template

B Exploratory Analyses

B.1 Change Score Analysis (Experiment 2)

As shown in Figure B1, we fail to detect relative increases in attitude strength or certainty for those in the counter-attitudinal conditions. Focusing on attitude strength, we generally detect moderation across the board for the "strong attitude," with those exposed to "Con," "Mixed," and "Pro" information reporting weaker attitudes after the informational and motivational interventions. Estimates for the "Con," "Mixed," and "Pro" condition are -.089 (SE = .049), -.118 (SE = .0427), and -.11 (SE = .056) scale points, respectively. Changes in attitude strength for the weak condition are muted and not statistically discernible from zero, ranging from -.006 to .052 scale points. Turning now to attitude certainty, certainty shifts by -.146 in the "Con" condition, -.08 in the "Mixed" condition, and .023 in the "Pro" condition when assessing the "strong attitude." The difference in Δ attitude certainty between the "Pro" and "Con" condition (d = .16; SE = .06; p = .003) is significant. Finally, turning to the "weak attitude," we observe significant increases in certainty for all informational conditions. Certainty scores increase by .94, .91, and 1.17 scale points in the "Con," "Mixed," and "Pro" conditions, respectively. The change in the "Pro" condition is larger than the counterattitudinal conditions. However, these differences are not statistically significant. Taken together, we detect evidence of moderation with respect to strong attitudes and increases in certainty for weak attitudes. For the latter, increases in certainty occur across the board, which is consistent with a possible learning process whereby exposure to any new information causes an increase in certainty. An additional analysis that examines conditional means across the motivational conditions fails to recover evidence of attitude polarization.

FIGURE B1. Change Score Analysis of Attitude Strength and Certainty Across Information and Issue Strength Conditions



Note: This figure presents point estimates and confidence intervals for changes in attitude strength and certainty across information 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. See Table B1 for full model results.

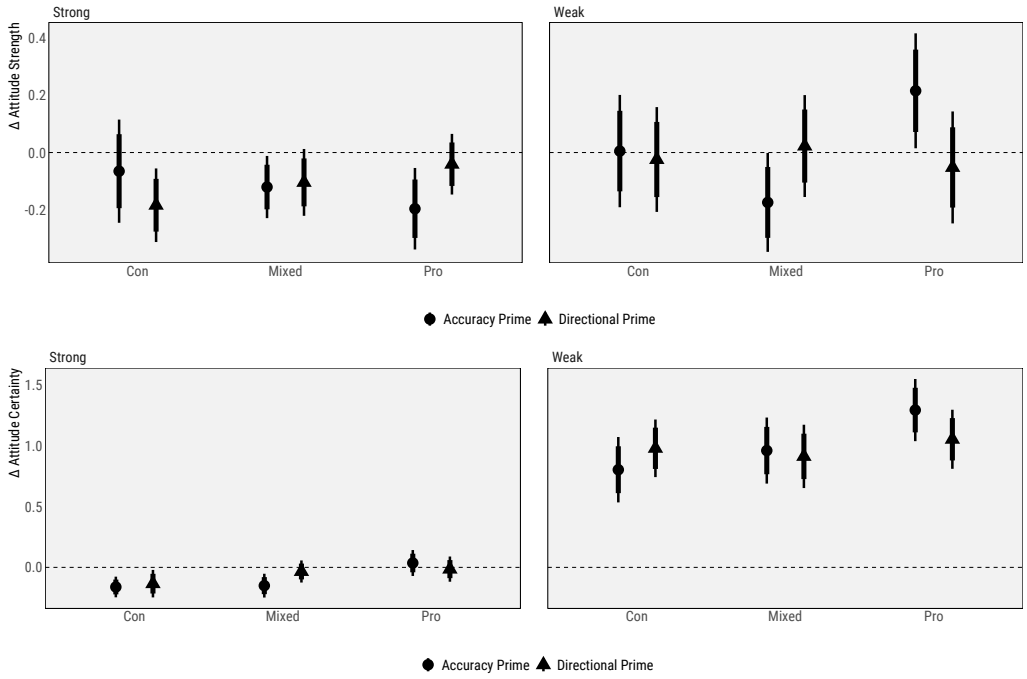
TABLE B1. Estimated Marginal Means of Change Scores Across Information and Issue Strength Conditions

Outcome	Issue Strength	Information	Mean	SE	df	*p*
Δ Strength	Strong	Pro	-0.11	0.04	1127	0.01
		Con	-0.13	0.06	1127	0.03
		Mixed	-0.11	0.04	1127	0.01
	Weak	Pro	0.09	0.07	1218	0.23
		Con	-0.01	0.07	1218	0.88
		Mixed	-0.07	0.06	1218	0.27
Δ Certainty	Strong	Pro	0.01	0.04	1134	0.81
		Con	-0.15	0.04	1134	0.00
		Mixed	-0.09	0.03	1134	0.01
	Weak	Pro	1.18	0.09	1222	0.00
		Con	0.89	0.09	1222	0.00
		Mixed	0.93	0.10	1222	0.00

Note: OLS regression with robust standard errors clustered at the participant level.

APSR Submission Template

FIGURE B2. Change Score Analysis of Attitude Strength and Certainty Across Information, Motivation, and Issue Strength Conditions



Note: This figure presents point estimates and confidence intervals for changes in attitude strength and certainty across information conditions with facets defined by issue attitude strength and motivation 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 Table B2 for full model results.

TABLE B2. Estimated Marginal Means of Change Scores Across Information, Motivation, and Issue Strength Conditions

Outcome	Motivation	Issue Strength	Information	Mean	SE	df	*p*
Δ Strength	Accuracy Prime	Strong	Pro	-0.20	0.07	556	0.01
			Con	-0.06	0.09	556	0.48
			Mixed	-0.12	0.06	556	0.03
		Weak	Pro	0.21	0.10	597	0.04
			Con	0.01	0.10	597	0.96
			Mixed	-0.17	0.09	597	0.05
	Directional Prime	Strong	Pro	-0.04	0.05	568	0.45
			Con	-0.18	0.07	568	0.01
			Mixed	-0.10	0.06	568	0.08
		Weak	Pro	-0.05	0.10	618	0.60
			Con	-0.02	0.09	618	0.79
			Mixed	0.02	0.09	618	0.80
Δ Certainty	Accuracy Prime	Strong	Pro	0.04	0.05	560	0.51
			Con	-0.16	0.04	560	0.00
			Mixed	-0.15	0.05	560	0.00
		Weak	Pro	1.29	0.13	597	0.00
			Con	0.80	0.14	597	0.00
			Mixed	0.96	0.14	597	0.00
	Directional Prime	Strong	Pro	-0.01	0.05	571	0.79
			Con	-0.13	0.06	571	0.02
			Mixed	-0.03	0.05	571	0.46
		Weak	Pro	1.05	0.12	622	0.00
			Con	0.98	0.12	622	0.00
			Mixed	0.91	0.13	622	0.00

Note: OLS regression with robust standard errors clustered at the participant level.

APSR Submission Template

B.2 Are GPT-3 Arguments Persuasive? Full Results

The results below correspond to Figure B4 in the Appendix. See Appendix B.2 for a full discussion of the Kialo study.

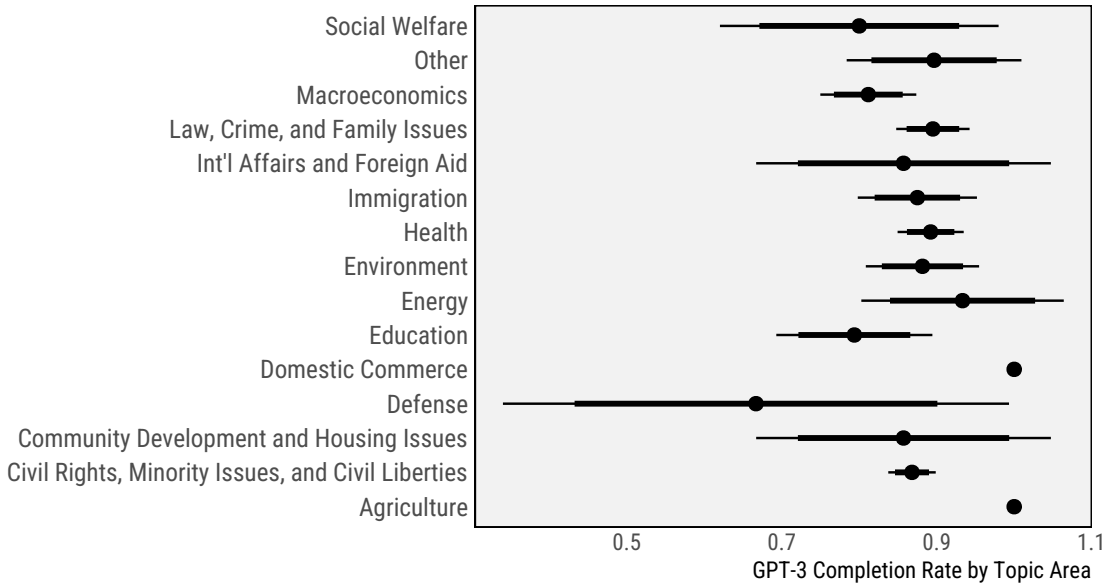
TABLE B3. GPT versus Kialo Argument Ratings by Subtopic

Subtopic	Mean	SE	p	95% C.I.
Big multinational companies should stop manufacturing in developing countries.	0.39	0.05	0.00	(0.30, 0.49)
COVID-19 vaccines should not be mandatory.	0.39	0.06	0.00	(0.28, 0.50)
Cannabis should be federally legalized in the United States.	0.76	0.05	0.00	(0.67, 0.86)
Death penalty should not be abolished.	0.41	0.05	0.00	(0.31, 0.51)
Governments should make an effort to reduce the gender pay gap.	0.50	0.05	0.00	(0.40, 0.61)
Hate speech should be legally protected.	0.37	0.06	0.00	(0.25, 0.49)
Humans should act to fight climate change.	0.27	0.05	0.00	(0.18, 0.37)
K-12 teachers are not paid enough in America.	0.46	0.05	0.00	(0.36, 0.56)
Nuclear energy should be the most prominent source of energy.	0.50	0.05	0.00	(0.41, 0.60)
Pregnant people should have the right to choose abortion.	0.61	0.05	0.00	(0.51, 0.71)
Student loan debt in the US should be forgiven for all.	0.59	0.05	0.00	(0.50, 0.69)
The US should adopt stricter gun control legislation.	0.38	0.05	0.00	(0.28, 0.48)
The US should build a wall on its Mexican border.	0.23	0.04	0.00	(0.14, 0.32)
The US should introduce a carbon tax.	0.66	0.05	0.00	(0.56, 0.76)
The US should pay reparations for slavery.	0.52	0.06	0.00	(0.40, 0.63)
The United States needs a strong political party for moderates.	0.39	0.05	0.00	(0.29, 0.49)
The West should give less development aid.	0.37	0.04	0.00	(0.28, 0.46)
The internet will be better off with Net Neutrality regulations.	0.24	0.05	0.00	(0.15, 0.34)
U.S. military spending should not be lowered in favor of other programs.	0.49	0.05	0.00	(0.39, 0.58)
Undocumented immigration in the United States is a problem.	0.45	0.05	0.00	(0.35, 0.55)
Wealthy countries should provide citizens with a universal basic income (UBI).	0.44	0.05	0.00	(0.34, 0.54)

B.3 GPT-3 Completion by Issue Topic (Experiment 2)

Although GPT-3 failures occurred a small percentage of the time, these errors occurred before the treatment was administered. As discussed in the paper, participants could be assigned to either (1) four pros, (2) four cons, or (3) two pros and two cons from a balanced list of eight GPT-3 generated pros and cons. This list was produced before treatment assignment, and in cases where GPT-3 could not produce a list of pros and cons, these respondents were flagged in the survey software (i.e., generic_flag = 1) and provided with a generic set of pros and cons discussing government intervention. As we note in our pre-registration plans, we exclude these participants because they did not receive a tailored response. Since all of this occurs before treatment assignment, there is no risk of confounding. However, these errors could affect the representativeness of our estimated treatment effects if certain topics are less likely to receive a tailored response. Though there is variation across issue topics with respect to successful GPT-3 output, differences are small in magnitude, and we fail to reject the null hypothesis in an F-test assessing equality of issue topics with respect to GPT-3 completions ($F(14, 1298) = 1.12$, $p = .33$).

FIGURE B3. Proportion of Valid GPT Completions by Issue Topic



Note: This figure presents point estimates for the proportion of valid GPT-3 responses by issue topic. 84% (thick bar) and 95% (thin bar) CIs are presented alongside point estimates.

TABLE B4. F-test assessing equality of issue topics

	df	Sum Sq.	Mean Sq.	F	p
Issue Topic	14	1.83	0.13	1.12	0.3308
Residuals	1298	150.58	0.12		

APSR Submission Template

B.4 Certainty Analysis (Experiment 3)

To minimize the possibility of multiple testing, we pre-registered an analysis of only two outcomes (strength and multi-item certainty). However, we also included a single-item measure of certainty in our survey. As an exploratory analysis, we assess if this item was also moved by the intervention. We find evidence that exposure to counter-attitudinal information reduced certainty on this scale by about 1.7 scale points on a 0-100 scale (p = .05, adjusting for the pre-registered set of covariates used in the main models). This corresponds to a shift of approximately .09 control-group standard deviation units.

	Model 1	Model 2
(Intercept)	88.012***	87.945***
	(0.613)	(0.605)
Treatment	-1.527+	-1.708*
	(0.884)	(0.870)
Age		3.195***
		(0.465)
Education		0.276
		(0.487)
Income		0.951*
		(0.465)
Ideology		-0.872
		(0.589)
White		0.281
		(0.476)
Partisanship		0.277
		(0.604)
N	1815	1773
R^2	0.002	0.038

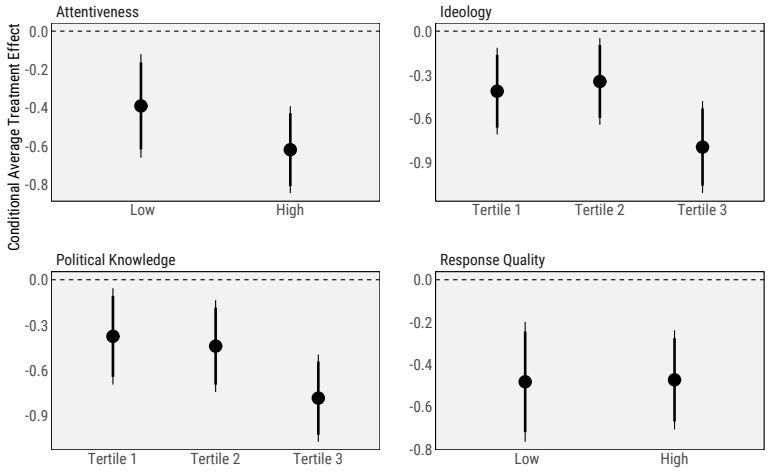
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Covariates are mean centered.

B.5 Robustness of Effects (Experiment 3)

We assess conditional effects across levels of attentiveness, ideology, political knowledge, and response quality. Attentiveness is measured using a question that assesses the retention of information provided in a paragraph-long news vignette. As **?** show, this method tends to predict longer survey duration, higher factual manipulation check passage, and larger treatment effects. Ideology is measured using a five-point scale. Political knowledge is measured using four questions about American politics (i.e., majority party in the House, most conservative party, size of veto-proof majority, knowledge of Supreme Court's function, former Vice President identification). Finally, response quality is measured using a GPT-3 script ("low" represents the medium quality response, given that we filtered out participants who provided low quality responses). We break the continuous and ordinal variables into tertiles to avoid model dependence issues with linear multiplicative interaction models (**?**) and estimate CATEs using the pre-registered models presented in the manuscript (i.e., covariate-adjusted average treatment effects with HC2 robust standard errors). We detect limited treatment effect heterogeneity across subgroups, but conditional effects generally increase with attentiveness, political knowledge, and conservative self-identification. In all cases, we detect negative point estimates, such that individuals moderate when exposed to counter-attitudinal information.

FIGURE B4. Individual-level CATEs



Note: This figure presents point estimates for conditional average treatment effects by attentiveness, response quality, ideology, and political knowledge scores.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

B.6 Consistency of GPT Output

Our design relies on GPT-3 and GPT-3.5 to generate both Likert items and arguments based on respondents' open-ended input. In other sections, we explore the implications of formatting errors in GPT-generated Likert items (ASD B.7) or instances in which the model fails to generate any arguments (ASD B.3). In this section, we examine a variety of errors that may occur when respondents' open-ended input is difficult for the model to interpret. Prompts that were written unclearly or were excessively long may lower the precision of our measurement (i.e., by causing the model to generate double-barreled Likert items) or affect the intensity of our treatment intervention (i.e., by leading the model to write off-topic arguments). Over the course of our five experiments, we refined our prompts to improve the consistency of GPT output.

We sampled 100 observations from each experiment and hand-coded each instance of an error. The types of errors and their implications for our analysis are discussed below. Note that one Likert item or argument could be coded as exhibiting more than one error.

Likert items:

- Double-barreled: Likert item describes more than one attitude object or more than one distinct evaluation of the same object. In theory, respondents may find it more difficult to rate their level of agreement, causing greater variance. In practice, double-barreled items most often contained two complementary evaluations of a single attitude object and simply reflected the respondents' own reasoning or nuance – e.g., "I believe that climate change is real and money should be set aside to stop it." Though this error may not represent a significant threat to precision in our outcome measures, we refined our prompts in Experiments 4 and 5, improving the model's capacity to distill respondents' answers into a single evaluation. Table B5 shows that whereas 20-26% of sampled Likert items in Experiments 1-3 were double-barreled, only 5-6% of items were coded as such in Experiments 4-5.

- Justified: Likert item includes one or more reasons to support the attitude – e.g., "I believe that retired elderly folks should receive higher social security benefits in order to live comfortable lives." Similar to double-barreled items, Likert items that contain additional reasoning could harm measurement precision but were unlikely to confuse respondents in most cases because they mirrored the form of respondents' own answers. Table B5 shows that about 12-14% of sampled Likert items were coded as "justified" in Experiments 1-3. We refined our prompts in Experiments 4 and 5 and found no instances of this error among the sampled observations we coded.

- Empty or Unclear: Likert item is uninterpretable, either because no Likert item was generated or because the statement is incomplete or vague – e.g., "I believe that there needs to be some general wealth distribution away from the top .0" or "of meritI believe that it is too political and not work on basis of merit." This error occurred rarely but was somewhat more frequent in Experiment 3, primarily due to lower quality responses by respondents. In spite of the noise that this error might have created, we find evidence of moderation in Experiment 3.

- Mischaracterized: Likert item mischaracterizes the attitude shared by the respondent. In these cases, respondents were shown an issue slightly different from the one they wrote about, and for which they may not hold a strong prior attitude. For instance, one respondent in Experiment 3 who wrote "I believe we should stop funding other wars" (expressing non-interventionism) was asked to rate their agreement with the statement "I believe we should stop funding wars" (anti-war sentiment). Among sampled Likert items, we found no instances

in the first two experiments and a very small number of instances in Experiments 3-5. Given the theoretical expectation that strong prior attitudes are necessary for backfire to occur, occasional instances of mischaracterized Likert items would in theory limit our capacity to detect attitude polarization, yet we find evidence of polarization in Experiments 4 and 5.

- Non-Attitude: Likert items that fail to capture a core attitude, either because the respondent did not offer an issue (e.g., "I'm not sure of how to respond") or because their open-ended response merely mimicked the example sentence we provide concerning farm subsidies (e.g., "I believe farm subsides should be increased to help farmers."). Though we were mostly successful in filtering out of the experiment respondents who wrote about farm subsidies, a handful of respondents who misspelled keywords in their response were able to complete the experiment. These instances were rare, however, and concentrated in Experiment 3.

TABLE B5. Proportion of Sampled Likert Items Containing Errors

Exp	Double-barreled	Justified	Empty or Unclear	Mischaracterized	Non-Attitude
1	26%	14%	2%	0%	0%
2	20%	14%	0%	0%	0%
3	22%	12%	7%	6%	5%
4	6%	0%	0%	4%	0%
5	5%	0%	0%	3%	1%

Note: N = 100 subjects were sampled from each experiment.

Arguments: In Experiments 1-2, GPT-3 generated four pro-attitudinal and four counter-attitudinal arguments based on respondents' open-ended answers. Respondents were then assigned to read all four pro statements, all four con statements, or two pros and two cons. To examine the frequency of errors in these arguments, we sampled 100 respondents from each experiment, then randomly selected one pro and one con argument from each respondent. For each of Experiments 3-5, we sampled counterarguments shown to 100 participants assigned to treatment.

- Includes Con/Pro Info: Treatment text in the counter-attitudinal (pro-attitudinal) condition includes any pro (con) information or arguments. In a handful of cases, this error occurred because the model mistook which position the respondent adopted. In most instances, the model understood the respondent's stance but countered from a more impassioned position rather than from the opposite side of the ideological spectrum (e.g., "Gay rights are not something to be in favor of, they are something to be demanded! ..." in reply to "Gay rights in favor of"). OpenAI's content filter may play a role here, preventing the model from outputting toxic content on issues such as inclusivity or mental health.

  This error raises two potential concerns. The first is that we may spuriously detect attitude polarization when, in truth, respondents are updating in the direction of information and arguments aligned with their priors. The attitude polarization we observe in Experiments 4 and 5 is unlikely to be explained away by this error given the small number of occurrences.

  A second concern is that respondents do not update in the direction of the pro arguments, but these errors leave us underpowered to detect attitude polarization. Given that approximately one-third of the counterarguments in Experiment 3 were generated from an aligned point of view, we conducted a follow-up coding to identify cases where the counterarguments clearly disputed the participant's issue position from a different point of view. Conditioning on this argument feature, we recover an average treatment effect of -0.47 (SE = 0.11; $p < 0.001$;

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

see Table B7), which is only slightly smaller in magnitude than the estimate we recover in the full sample (-0.487; SE = 0.089; p < 0.001; see Table A3). Thus, the attitude moderation we observe in Experiment 3 cannot be explained by "aligned counterarguments," where GPT-3 implores the participant to take a stronger position or grants their argument but suggests a slight modification.

- Weak: Treatment text is off-topic, lacking argumentative content, or (on occasion in the Pro-Attitudinal condition) a restatement of the respondent's attitude. These errors were rare in Experiments 3-5. A higher proportion of sampled arguments were coded as weak in Experiments 1 and 2 in part becase these arguments were a single sentence and in part because GPT-3 was tasked with generating four distinct pro or con arguments at a time. Though instances of weak arguments may weaken the intensity of the treatment, this concern is mitigated by the fact that respondents were shown a set of four arguments in total.

TABLE B6. Proportion of Sampled Arguments Containing Errors

	Pro-Attitudinal		Counter-Attitudinal	
Exp	Includes Con Info	Weak	Includes Pro Info	Weak
1	3%	18%	3%	9%
2	3%	7%	5%	6%
3			29%	3%
4			5%	1%
5			7%	0%

Note: N = 100 subjects were sampled from each experiment. Pro-attitudinal arguments were used only in experiments 1 and 2.

TABLE B7. Robustness Check (Experiment 3, Removing Aligned Counterarguments)

	Dependent variable:
	Argument Strength (1-7)
Intercept	6.05^{***}
	(0.06)
Treatment	-0.47^{***}
	(0.11)
Observations	1,389
Note:	*p<0.1; **p<0.05; ***p<0.01

B.7 Implications of GPT-3 Formatting Errors

In a small minority of cases, GPT-3 autocompletes sentences before summarizing them. Therefore, a respondent writing "gun control" may receive a Likert item stating "is importantI believe gun control is an important issue." This was relatively rare across the studies (7% in Experiment 1; 5% in Experiment 2; 19% in Experiment 3), and the gist of the survey question is still discernible in most cases. However, we still think it is worth considering whether (1) these errors are distributed evenly across conditions and (2) these errors condition treatment effects. Given that the API is called before treatment assignment, we expect minimal differences across conditions. This is exactly what we find. Errors are balanced across conditions, with one marginally significant exception. In the vast majority of cases, differences in errors are small and indistinguishable from zero.

Moving on to conditional effects, we find marginally significant evidence of greater attitude moderation among those in the directional and mixed information condition if they receive a properly formatted GPT-3 Likert summary in Experiment 1 ($\hat{\beta}$ = -.455; SE = .245; p < .10). In Experiment 2, we find marginally significant evidence that the "Con" effect on attitude certainty for the strong attitude is muted in the accuracy condition if participants receive a Likert item without this completion error ($\hat{\beta}$ = .590; SE = .352; p < .10). Finally, in Experiment 3, we find that receiving a "correct Likert item" decreases the effect of counter-attitudinal information on certainty ($\hat{\beta}$ = -.789; SE = .201; p < .001). The average treatment effect is negative and statistically significant among those with properly formatted Likert items ($\hat{\beta}$ = -.14; SE = .07; p = .05).

Taken together, the evidence suggests minimal effects of formatting errors on our findings. However, future iterations of this design should be mindful of creating prompts that minimize these errors. Encouraging longer open-ended responses that have a similar format might be one solution (i.e., "I believe that X"). Relative to the CloudResearch studies, the median number of characters in the Lucid open-ended question was 75 whereas the median number of characters in Experiments 1 and 2 was 100 and 93, respectively. In the Lucid study, approximately 36% of responses began with "I think" or "I believe," whereas this number was 68% and 67% in Experiments 1 and 2, respectively. Future studies could encourage longer and more consistently formatted open-ended responses to reduce error rates. Fine-tuning GPT-3 can also help: using correctly formatted data from Experiments 1-3 as training data, we have found that error rates substantially decrease.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

TABLE B8. Balance in GPT-3 Errors

	DV: Correct Likert (No Formatting Error, 0/1)		
	Experiment 1	Experiment 2	Experiment 3
(Intercept)	0.933***	0.955***	0.807***
	(0.013)	(0.009)	(0.013)
Con	0.003	-0.008	
	(0.015)	(0.011)	
Mixed	0.010	-0.016	
	(0.015)	(0.011)	
Directional	-0.021+	0.005	
	(0.012)	(0.009)	
Counter-Attitudinal Argument			−0.025
			(0.019)
N	1782	2362	1901
R^2	0.002	0.001	0.001

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

TABLE B9. Conditional Effects of GPT-3 Errors (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.924***	97.317***
	(0.076)	(0.706)
Pre-Treatment Attitudes	0.270***	0.039
	(0.063)	(0.191)
Pre-Treatment Certainty	0.181***	6.133***
	(0.034)	(0.473)
Correct Likert	-0.086	-0.217
	(0.081)	(0.733)
Con	-0.056	-0.217
	(0.079)	(0.884)
Mixed	-0.268	-0.542
	(0.198)	(0.880)
Directional	-0.190	-0.525
	(0.168)	(0.835)
Correct Likert × Con	0.045	0.209
	(0.089)	(0.941)
Correct Likert × Mixed	0.264	0.579
	(0.199)	(0.910)
Correct Likert × Directional	0.220	0.627
	(0.171)	(0.868)
Con × Directional	0.106	0.660
	(0.196)	(1.832)
Mixed × Directional	0.419+	0.666
	(0.243)	(1.073)
Correct Likert × Con × Directional	-0.123	-0.857
	(0.204)	(1.926)
Correct Likert × Mixed × Directional	-0.455+	-1.687
	(0.245)	(1.179)
N	1768	1772
R^2	0.440	0.719

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

APSR Submission Template

TABLE B10. Conditional Effects of GPT-3 Errors (Experiment 2)

	Attitude Strength (1-7)	Attitude Certainty (1-9)
(Intercept)	5.796***	7.473***
	(0.154)	(0.183)
Con	0.248	−0.650+
	(0.205)	(0.344)
Mixed	0.282	−0.244
	(0.192)	(0.286)
Correct Likert	0.101	−0.030
	(0.162)	(0.173)
Directional Prime	−0.076	−0.232
	(0.265)	(0.261)
Pre-Treatment Strength	0.480***	0.230**
	(0.123)	(0.083)
Pre-Treatment Certainty	0.550***	1.173***
	(0.116)	(0.126)
Con × Correct Likert	−0.185	0.590+
	(0.231)	(0.352)
Mixed × Correct Likert	−0.228	0.101
	(0.213)	(0.295)
Con × Directional Prime	0.107	0.594
	(0.309)	(0.436)
Mixed × Directional Prime	−0.025	0.263
	(0.311)	(0.398)
Correct Likert × Directional Prime	0.198	0.289
	(0.280)	(0.268)
Con × Correct Likert × Directional Prime	−0.275	−0.729
	(0.336)	(0.449)
Mixed × Correct Likert × Directional Prime	−0.070	−0.230
	(0.333)	(0.411)
N	1122	1126
R^2	0.214	0.348

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

TABLE B11. Conditional Effects of GPT-3 Errors (Experiment 3)

	Attitude Strength (1-7)	Attitude Certainty (1-9)
(Intercept)	5.255***	6.751***
	(0.138)	(0.150)
Treatment	−0.596**	0.643***
	(0.216)	(0.187)
Correct Likert	0.971***	1.084***
	(0.150)	(0.159)
Age	0.116*	0.316***
	(0.048)	(0.037)
Education	0.072	0.037
	(0.049)	(0.039)
Income	0.040	0.115**
	(0.046)	(0.036)
Ideology	−0.051	−0.065
	(0.057)	(0.049)
White	0.111*	0.092*
	(0.047)	(0.039)
Partisanship	−0.096+	−0.016
	(0.055)	(0.046)
Treatment × Correct Likert	0.152	−0.789***
	(0.236)	(0.201)
N	1854	1858
R2	0.080	0.105

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

APSR Submission Template

B.8 CATEs by Subtopic

We investigate whether treatment effects vary across subtopics. To define issue topics, we rely on the Comparative Agendas Project coding scheme, which comprises 21 major political topics and 220 subtopics (see Appendix B.1 for more detail). For each of the three experiments, we subsetted the data separately on the three subtopics most frequently invoked by respondents (abortion, health care reform, and gun control) and re-ran the main regressions to find conditional average treatment effects. We chose to stop at the third most common subtopic due to the small number of observations for all other subtopics. Results from these regressions are displayed in the nine tables below. We do not find evidence that attitude polarization occurs for some issues but not others.

TABLE B12. CATEs for Subtopic 1: Abortion (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.829***	97.114***
	(0.028)	(0.420)
Pre-Treatment Strength	0.474**	0.356
	(0.155)	(0.277)
Pre-Treatment Certainty	0.063	5.840***
	(0.059)	(1.188)
Con	0.035	0.401
	(0.049)	(0.395)
Mixed	−0.058	−0.008
	(0.070)	(0.396)
Directional Prime	−0.053	0.490
	(0.045)	(0.406)
Con × Directional Prime	0.021	−0.397
	(0.050)	(0.470)
Mixed × Directional Prime	0.136+	−0.361
	(0.081)	(0.474)
Num.Obs.	476	477
R2	0.504	0.734
R2 Adj.	0.497	0.730
AIC	521.2	2075.6
BIC	558.6	2113.1
RMSE	0.41	2.09

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APSR Submission Template

TABLE B13. CATEs for Subtopic 2: Comprehensive Health Care Reform (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.925***	97.031***
	(0.050)	(0.350)
Pre-Treatment Strength	0.148	0.534
	(0.192)	(0.776)
Pre-Treatment Certainty	0.200*	6.696***
	(0.101)	(0.561)
Con	−0.190	−0.045
	(0.163)	(0.453)
Mixed	−0.085	−0.216
	(0.055)	(0.619)
Directional Prime	−0.055	0.103
	(0.056)	(0.399)
Con × Directional Prime	0.220	0.036
	(0.174)	(0.795)
Mixed × Directional Prime	0.085	0.177
	(0.077)	(0.674)
Num.Obs.	219	219
R2	0.196	0.854
R2 Adj.	0.170	0.850
AIC	301.2	973.1
BIC	331.7	1003.6
RMSE	0.46	2.14

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B14. CATEs for Subtopic 3: Gun Control (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.669***	96.862***
	(0.145)	(0.132)
Pre-Treatment Strength	0.741***	0.155
	(0.126)	(0.485)
Pre-Treatment Certainty	0.012	7.360***
	(0.071)	(0.755)
Con	0.125	−0.882
	(0.153)	(0.568)
Mixed	0.106	0.142
	(0.150)	(0.251)
Directional Prime	0.151	−1.280
	(0.144)	(0.841)
Con × Directional Prime	−0.147	1.686
	(0.158)	(1.188)
Mixed × Directional Prime	−0.094	1.733
	(0.185)	(1.258)
Num.Obs.	103	104
R2	0.814	0.794
R2 Adj.	0.801	0.779
AIC	5.8	510.9
BIC	29.5	534.7
RMSE	0.23	2.59

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APSR Submission Template

TABLE B15. CATEs for Subtopic 1: Abortion (Experiment 2)

	Attitude Strength (1-7)	Certainty (0-100)	Certainty Score (1-9)
(Intercept)	5.805***	89.613***	7.973***
	(0.254)	(1.810)	(0.114)
Pre-Treatment Strength	0.723***	0.248	−0.018
	(0.118)	(1.097)	(0.084)
Pre-Treatment Certainty	0.218*	9.417***	
	(0.101)	(1.409)	
Con	0.428	−0.799	−0.254*
	(0.310)	(2.054)	(0.126)
Mixed	0.223	1.079	−0.113
	(0.281)	(1.620)	(0.103)
Directional Prime	0.299	−0.390	−0.118
	(0.250)	(1.744)	(0.128)
Weak Attitude	0.039	−4.634	−0.465+
	(0.314)	(3.379)	(0.250)
Con × Directional Prime	−0.589+	1.778	0.196
	(0.333)	(2.379)	(0.194)
Mixed × Directional Prime	−0.366	−1.967	0.041
	(0.318)	(2.172)	(0.179)
Con × Weak Attitude	−0.592	−8.241	−0.320
	(0.442)	(5.242)	(0.379)
Mixed × Weak Attitude	−0.914*	−5.943	−0.213
	(0.402)	(4.626)	(0.343)
Directional Prime × Weak Attitude	−0.652+	−3.911	−0.178
	(0.379)	(4.571)	(0.332)
Con × Directional Prime × Weak Attitude	0.777	6.469	0.577
	(0.529)	(7.116)	(0.506)
Mixed × Directional Prime × Weak Attitude	1.376**	10.291+	0.221
	(0.508)	(6.133)	(0.497)
Pre-Treatment Certainty Scores			0.913***
			(0.110)
Num.Obs.	538	536	540
R2	0.476	0.445	0.473
R2 Adj.	0.463	0.431	0.460
AIC	1701.7	4453.4	1792.3
BIC	1766.1	4517.6	1856.7
RMSE	1.14	14.99	1.24
Std.Errors	by: id	by: id	by: id

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B16. CATEs for Subtopic 2: Comprehensive Health Care Reform (Experiment 2)

	Attitude Strength (1-7)	Certainty (0-100)	Certainty Score (1-9)
(Intercept)	5.874***	85.362***	7.662***
	(0.163)	(2.801)	(0.161)
Pre-Treatment Strength	0.893***	1.935	0.115
	(0.155)	(1.597)	(0.102)
Pre-Treatment Certainty	0.064	9.036***	
	(0.107)	(1.706)	
Con	0.381	1.666	−0.081
	(0.234)	(2.937)	(0.200)
Mixed	0.117	1.544	−0.125
	(0.170)	(2.960)	(0.176)
Directional Prime	0.173	3.746	−0.160
	(0.163)	(2.559)	(0.189)
Weak Attitude	0.014	1.615	0.391
	(0.377)	(4.689)	(0.328)
Con × Directional Prime	−0.360	−2.444	−0.001
	(0.265)	(3.243)	(0.262)
Mixed × Directional Prime	0.104	−1.044	0.269
	(0.202)	(3.183)	(0.243)
Con × Weak Attitude	−0.975*	−9.898+	−0.971*
	(0.401)	(5.630)	(0.450)
Mixed × Weak Attitude	−0.093	−6.866	−0.019
	(0.407)	(6.196)	(0.495)
Directional Prime × Weak Attitude	−0.212	0.151	−0.139
	(0.400)	(5.834)	(0.429)
Con × Directional Prime × Weak Attitude	0.762	−2.755	0.772
	(0.576)	(8.729)	(0.626)
Mixed × Directional Prime × Weak Attitude	−0.403	−6.553	−0.901
	(0.537)	(8.495)	(0.690)
Pre-Treatment Certainty Scores			1.077***
			(0.125)
Num.Obs.	326	329	331
R2	0.545	0.422	0.511
R2 Adj.	0.526	0.398	0.491
AIC	961.5	2767.5	1103.4
BIC	1018.3	2824.4	1160.4
RMSE	1.01	15.51	1.22
Std.Errors	by: id	by: id	by: id

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APSR Submission Template

TABLE B17. CATEs for Subtopic 3: Gun Control (Experiment 2)

	Attitude Strength (1-7)	Certainty (0-100)	Certainty Score (1-9)
(Intercept)	6.315***	87.300***	7.466***
	(0.110)	(2.113)	(0.197)
Pre-Treatment Strength	0.845***	3.019+	0.084
	(0.173)	(1.784)	(0.146)
Pre-Treatment Certainty	−0.029	10.119***	
	(0.154)	(2.748)	
Con	−0.216*	−1.995	0.138
	(0.103)	(2.440)	(0.206)
Mixed	−0.369*	−4.404	0.087
	(0.162)	(2.763)	(0.241)
Directional Prime	0.253+	1.397	0.163
	(0.153)	(1.965)	(0.191)
Weak Attitude	−1.508**	−9.590	−0.099
	(0.496)	(6.989)	(0.450)
Con × Directional Prime	−0.119	−1.709	−0.237
	(0.254)	(3.767)	(0.292)
Mixed × Directional Prime	−0.182	0.151	−0.069
	(0.258)	(3.484)	(0.324)
Con × Weak Attitude	1.323**	15.564*	−0.001
	(0.505)	(7.862)	(0.613)
Mixed × Weak Attitude	0.967+	5.936	−0.331
	(0.526)	(7.396)	(0.555)
Directional Prime × Weak Attitude	0.276	1.056	−0.198
	(0.537)	(8.770)	(0.534)
Con × Directional Prime × Weak Attitude	−0.590	−11.742	0.300
	(0.692)	(11.034)	(0.764)
Mixed × Directional Prime × Weak Attitude	−0.158	8.124	1.048
	(0.753)	(10.705)	(0.796)
Pre-Treatment Certainty Scores			1.000***
			(0.189)
Num.Obs.	194	193	196
R2	0.573	0.475	0.515
R2 Adj.	0.542	0.437	0.480
AIC	550.1	1607.2	616.2
BIC	599.1	1656.1	665.4
RMSE	0.92	14.40	1.08
Std.Errors	by: id	by: id	by: id

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B18. CATEs for Subtopic 1: Abortion (Experiment 3)

	Attitude Strength (1-7)	Attitude Strength (1-7)	Certainty Score (1-9)	Certainty Score (1-9)
(Intercept)	6.196***	6.247***	7.920***	7.943***
	(0.141)	(0.255)	(0.117)	(0.223)
Treatment	0.247	0.283	0.086	0.040
	(0.190)	(0.189)	(0.168)	(0.164)
Age		0.138		0.321***
		(0.093)		(0.094)
Education		0.062		0.035
		(0.102)		(0.099)
Income		0.044		0.091
		(0.101)		(0.085)
Ideology		−0.105		−0.091
		(0.152)		(0.151)
White		−0.054		−0.043
		(0.099)		(0.096)
Partisanship		−0.019		0.002
		(0.044)		(0.044)
Num.Obs.	279	271	280	272
R2	0.006	0.031	0.001	0.064
R2 Adj.	0.002	0.005	−0.003	0.040
AIC	3935.3	1027.9	3661.3	956.7
BIC	3946.2	1060.3	3672.2	989.1
RMSE	1.60	1.56	1.40	1.36

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APSR Submission Template

TABLE B19. CATEs for Subtopic 2: Comprehensive Health Care Reform (Experiment 3)

	Attitude Strength (1-7)	Attitude Strength (1-7)	Certainty Score (1-9)	Certainty Score (1-9)
(Intercept)	5.929***	5.927***	7.796***	7.950***
	(0.371)	(0.575)	(0.282)	(0.461)
Treatment	−0.244	−0.509	−0.193	−0.334
	(0.501)	(0.521)	(0.370)	(0.347)
Age		0.047		−0.042
		(0.304)		(0.198)
Education		0.069		−0.138
		(0.261)		(0.190)
Income		0.464+		0.067
		(0.250)		(0.195)
Ideology		−0.424		−0.340
		(0.329)		(0.297)
White		0.277		−0.118
		(0.270)		(0.173)
Partisanship		0.002		−0.034
		(0.105)		(0.094)
Num.Obs.	66	65	66	65
R2	0.004	0.106	0.004	0.111
R2 Adj.	−0.012	−0.004	−0.011	0.002
AIC	4416.2	286.2	3752.1	244.3
BIC	4422.8	305.8	3758.7	263.9
RMSE	2.02	1.90	1.47	1.38

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B20. CATEs for Subtopic 3: Gun Control (Experiment 3)

	Attitude Strength (1-7)	Attitude Strength (1-7)	Certainty Score (1-9)	Certainty Score (1-9)
(Intercept)	5.879***	5.435***	8.004***	8.320***
	(0.379)	(0.805)	(0.205)	(0.442)
Treatment	0.121	−0.067	−0.359	−0.416
	(0.526)	(0.576)	(0.421)	(0.450)
Age		0.271		0.453+
		(0.244)		(0.233)
Education		−0.360		0.135
		(0.359)		(0.231)
Income		0.357		−0.086
		(0.295)		(0.206)
Ideology		−0.360		−0.132
		(0.502)		(0.219)
White		0.254		0.047
		(0.363)		(0.196)
Partisanship		0.115		−0.056
		(0.126)		(0.085)
Num.Obs.	58	57	58	57
R2	0.001	0.123	0.014	0.146
R2 Adj.	−0.017	−0.002	−0.003	0.024
AIC	4397.2	251.9	3785.6	216.2
BIC	4403.4	270.3	3791.8	234.6
RMSE	2.00	1.88	1.49	1.38

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APSR Submission Template

B.9 Assessing Post-Treatment Attrition (Experiments 4 and 5)

Given the vitriolic quality of treatment text in Experiments 4 and 5, a potential concern is that respondents could have responded poorly to treatment exposure and chosen to exit the survey before providing post-treatment outcome measures, resulting in differential attrition across treatment conditions. Such attrition could, in turn, drive our findings of attitude polarization if respondents who attrited were disproportionately unlikely to react in a defensive manner to vitriolic text—such as if attriters were less extreme partisans or exhibited lower pre-treatment attitude strength.

To assess whether differential attrition explains our findings in Experiments 4 and 5, we first illustrate the rates of post-treatment attrition for our two outcome scales. Attitude extremity and attitude defense questions were presented on separate pages. Both the order of the pages and order of questions on each page were randomized. The extremity scale comprises eight questions, while the defense scale comprises five questions. We code outcomes as missing if the respondent failed to provide an answer to any question. Attrition could therefore constitute skipping an entire page of outcomes or refusing to answer one question. Table B21 shows that attrition so defined was fairly low across both experiments but slightly higher in our fifth experiment than in our fourth.

TABLE B21. Rates of Post-Treatment Attrition

	Experiment 4 (N = 1,850)		*Experiment 5 (N = 1,985)*	
	Extremity	Defense	Extremity	Defense
Number missing	15	20	26	56
Percent missing	0.81%	1.08%	1.31%	2.82%

We now present estimates from linear models assessing attrition rates across experimental conditions. These models allow us to evaluate if those exposed to vitriolic counter-arguments, for example, were more likely to have missing values on outcome measures than those exposed to placebo content, indicating potential attrition problems.

TABLE B22. Predicting Missingness

	Experiment 4		*Experiment 5*	
	Extremity NA (1)	Defense NA (2)	Extremity NA (3)	Defense NA (4)
(Intercept)	0.009**	0.013***	0.011**	0.020***
	(0.003)	(0.004)	(0.004)	(0.005)
High Valence	−0.001	−0.004	0.000	0.006
	(0.004)	(0.005)	(0.006)	(0.008)
Low Valence			0.007	0.019*
			(0.006)	(0.009)
Num.Obs.	1850	1850	1985	1985
Note:			+p<0.1; *p<0.05; **p<0.01; ***p<0.001	

Table B22 suggests that respondents exposed to high valence, vitriolic counterarguments were no more likely than respondents exposed to placebo text to attrit in either experiment. Column 4 suggests that respondents exposed to low valence counterarguments were somewhat more likely

than control respondents to refrain from answering attitude defense questions in Experiment 5. We lack an intuition about why respondents in this treatment condition would attrit at higher rates, but significant differences between the low valence and control conditions would not explain why vitriolic counterarguments precipitated attitude polarization in any case.

Although we fail to find evidence of differential attrition rates by condition, it could still be the case that participants who possessed weaker attitudes exited the survey early after being exposed to vitriolic content. If this were the case, our effect estimate for the vitriolic condition would not reflect a causal effect, but would instead capture pre-treatment differences in attitude strength among participants who remained in the vitriol condition versus the placebo. To assess this, we estimate an additional set of models where we regress our attrition measures on (1) treatment status indicators, (2) pre-treatment measures of attitude strength, attitude certainty, and party ID strength (4-point scale), and (3) treatment by covariate interactions. To code party ID strength, we folded the standard 7-point party ID measure. True independents were coded as a 1, respondents who lean Democrat or lean Republican were coded as a 2, and so on.

Based on the results in Table B23, we fail to find evidence that those with weaker attitudes were more likely to exit the survey if exposed to vitriolic content. In all cases, point estimates for the treatment status by covariate interactions are close to zero and insignificant, indicating similar attrition patterns across subgroups varying in attitude strength, certainty, and partisan strength.

Though we fail to find evidence of differential attrition here, we advise that scholars remain vigilant about the potential for differential attrition when employing online convenience samples and exposing respondents to treatment texts containing vitriolic language. Attrition may be pernicious when exiting the survey is low-stakes and respondents vary in their prior exposure to or familiarity with distressing political content.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

TABLE B23. Predicting Missingness with Treatment by Covariate Interactions

	Experiment 4		*Experiment 5*	
	Extremity NA (1)	Defense NA (2)	Extremity NA (3)	Defense NA (4)
(Intercept)	0.047	0.026	−0.022	0.103
	(0.052)	(0.044)	(0.020)	(0.116)
High Valence	−0.041	0.062	0.056	−0.080
	(0.055)	(0.081)	(0.056)	(0.130)
Low Valence			0.049	0.048
			(0.051)	(0.182)
Pre-Treatment Strength	0.004	−0.001	0.007	−0.004
	(0.004)	(0.004)	(0.004)	(0.011)
Pre-Treatment Certainty	−0.001	0.000	0.000	−0.001
	(0.001)	(0.000)	(0.000)	(0.001)
Party ID Strength	0.001	0.006	−0.001	0.000
	(0.003)	(0.004)	(0.005)	(0.006)
High Valence X Strength	−0.002	−0.002	−0.006	0.000
	(0.004)	(0.005)	(0.007)	(0.020)
High Valence X Certainty	0.001	0.000	0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
High Valence X PID	−0.003	−0.004	0.006	0.002
	(0.004)	(0.005)	(0.006)	(0.008)
Low Valence X Strength			−0.013	−0.004
			(0.008)	(0.015)
Low Valence X Certainty			0.000	0.000
			(0.000)	(0.001)
Low Valence X PID			0.002	−0.003
			(0.007)	(0.009)
Num.Obs.	1843	1843	1975	1975

Note: +p<0.1; *p<0.05; **p<0.01; ***p<0.001

C Experimental Materials

C.1 Question Wording

- **Attitude strength (1-7)**: To what extent do you agree or disagree with the following statement? [GPT-3 summary]
  (*Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree*)
- **Attitude certainty (0-100)**: How certain are you regarding this position? [GPT-3 summary]
  When giving a score, keep in mind:
  - 100 = Complete certainty. You could not be more sure of where you stand.
  - 75 ≈ Very sure about where you stand.
  - 50 ≈ Pretty sure about where you stand.
  - 25 ≈ Mildly sure about where you stand.
  - 0 = Complete uncertainty. You have a belief, but you are very unsure if you should hold it.
- **Attitude certainty scale (1-9)**: On a scale of 1 to 9, where 1 = "Not at all certain" and 9 = "Very certain", how certain do you feel in response to the following questions?
  (*Each item is measured using a 9-point scale ranging from 1 (Not at all certain) to 9 (Very certain). Only endpoints of the scale are labeled.*)
  - How certain are you that your attitude on this issue is the correct attitude to have?
  - How certain are you that other people should have the same attitude as you on this issue?
  - How certain are you that of all the possible attitudes one might have on this issue, your attitude reflects the right way to think and feel about the issue?
  - How certain are you that you know what your true attitude on this issue really is?
  - How certain are you that the attitude you expressed on this issue really reflects your true thoughts and feelings?
  - To what extent is your true attitude on this issue clear in your mind?
  - How certain are you that the attitude you just expressed on this issue is really the attitude you have?
- **Attitude Duration**: How long have you held this belief with this level of certainty?
  - Since beginning this survey
  - In the last week
  - In the last month
  - In the last 2 to 6 months
  - In the last 7 to 12 months
  - In the last 2 to 3 years
  - Longer than 4 years
- **Discussion Frequency**: In the past 6 months, how often have you debated with family members, peers, or people online about this issue?
  - More than once a week
  - Once a week
  - Once or twice a week
  - Once a month

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

- A few times
- Have not debated this issue in past 6 months

- **Attitude extremity (1-7)**: To what extent do you agree or disagree with the following statements about your level of commitment to the following issue position: [GPT-3 summary]. Please indicate your level of agreement ranging from 'Strongly Disagree' to 'Strongly Agree.' (*Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree*)
  - I would phone state representatives to advocate for this issue.
  - I would risk personal harm or even death to advocate for this cause.
  - I would initiate grassroots campaigns to mobilize supporters for this issue.
  - I would walk five miles to bring attention to this issue.
  - I would join a protest, despite a risk of arrest, to support this issue.
  - I wouldn't cross the street to participate in a rally concerning this issue. (reverse-keyed)
  - I would go on a hunger strike to amplify awareness of this issue.
  - I would willingly get fired from my job to advance this issue.
- **Attitude defense (1-7)**: To what extent do you agree or disagree with the following statements about your level of commitment to the following issue position: [GPT-3 summary]. Please indicate your level of agreement ranging from 'Strongly Disagree' to 'Strongly Agree.' (*Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree*)
  - I would not be confident defending this position in a telephone interview that will end up in the local newspaper. (reverse-keyed)
  - I could confidently defend this position in a filmed interview for live national television.
  - I would not be confident defending this position in a press conference in front of news reporters. (reverse-keyed)
  - I could confidently defend this position in a public speech on a college campus.
  - I would not be confident defending this position at a city council meeting. (reverse-keyed)

C.2 GPT-3 Details

We provide the details of our various API calls here. The temperature parameter varies the extent to which the output is deterministic versus stochastic. Lower scores tend to generate similar output across different API calls. Max tokens captures the maximum amount of text tokens produced by GPT-3. The frequency penalty is a parameter that limits the tendency for GPT-3 to repeat output, whereas the presence penalty is a parameter that varies whether the model should produce more "novel" text predictions. OpenAI has released several models that vary in complexity and cost. In all of our experiments, we choose the "davinci" class of GPT-3 models or more recent models that mirror the quality of "davinci" (GPT-3.5 Turbo Instruct).

Likert API call:

- Model: Davinci-002
- Prompt: "Summarize this person's belief in the first person in one sentence (always begin with the word I believe and stick with what the person says):"
- Temperature: 0.7
- Max Tokens: 50
- Frequency Penalty: 0
- Presence Penalty: 0

Pros and Cons API call:

- Model: Davinci-002
- Prompt: "Write four pros and four cons for the following argument in a JSON format. Each pro and con should be labeled pro1, pro2, pro3, pro4, con1, con2, con3, con4."
- Temperature: 0.7
- Max Tokens: 250
- Frequency Penalty: 1
- Presence Penalty: 1

Counterattitudinal Argument API call:

- Model: Davinci-003
- Prompt: "Write a paragraph-long passionate rebuttal of the following argument. The author should strongly disagree with the statement."
- Temperature: 0
- Max Tokens: 256
- Frequency Penalty: 0
- Presence Penalty: 0

Open-Ended Response Quality Check API call:

- Model: Davinci-003
- Prompt: "Is the following response relevant to the prompt? Always generate a number (0 = low quality, .5 = medium quality, 1 = high quality).

APSR Submission Template

Examples: Prompt: 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? Response: I believe in abortion. {quality: ".5"}
Prompt: 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? Response: I believe if California stops handing out free money, they will go broke {quality: ".5"}
Prompt: 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? Response: I believe foods free help a lot of people {quality: "0"}
Prompt: 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? Response: Inflation has caused me to lose my house and car {quality: ".5"}
Prompt: 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? Response: Itś a issue about money {quality: "0"}
Prompt: 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? Response: I believe that healthcare is a human right and that everyone deserves access to quality care. {quality: "1"}
Prompt: 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? Response:"

- Temperature: 0
- Max Tokens: 50
- Frequency Penalty: 0
- Presence Penalty: 0

Vitriolic Counterargument API Call:

- Model: GPT-3.5 Turbo Instruct
- Prompt: "Given the user's argument, create a paragraph-long affectively charged counterargument. Use these guidelines: - Make sure the argument directly attacks what the person said. - Must be a paragraph and no longer. - Make sure the argument is emotionally charged (angry; passionate; strongly worded) - Eliminate the use of personal attacks (e.g., How dare you) and refer to the person in the abstract (do not use you) - Present a forceful and passionate argument that does not grant anything (while X might be true, I believe Y). Instead just passionately argue for Y. Argument:[text] Counterargument:"
- Temperature: 0
- Max Tokens: 500
- Frequency Penalty: Default
- Presence Penalty: Default

C.3 Sample Composition

TABLE C1 . Sample Composition

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Gender (F)	0.58	0.54	0.52	0.49	0.53
Age	43.11	43.40	45.02	42.16	53.29
Income	5.37	5.66	8.31	5.85	5.71
Education	4.30	4.38	4.39	4.38	4.35
Ideology (7-point scale)	2.68	2.79	3.18	2.79	2.75
Democrat	0.46	0.48	0.34	0.48	0.49
Republican	0.21	0.24	0.25	0.24	0.23
Independent	0.33	0.29	0.41	0.29	0.28
White	0.79	0.79	0.66	0.74	0.64
Latino	0.07	0.07	0.14	0.12	0.15
Black	0.07	0.07	0.11	0.12	0.15
AAPI	0.05	0.05	0.04	0.02	0.05

Note: Mean estimates for demographic categories. Racial and ethnic demographic estimates for Experiment 1 and 2 were obtained by recontacting a random sample of participants from each study (N = 517). Sample composition for Experiments 4-5 was taken directly from the CloudResearch Connect platform.

APSR Submission Template

D AsPredicted Pre-Registration Plans

D.1 Pre-Registration for Experiment 1 (AsPredicted #108182)

URL: https://aspredicted.org/3Z5_F2C

Created: 09/28/2022 02:19 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Q: Do respondents who have been primed for directional motives update their attitudes less than respondents who have been primed for accuracy motives when both groups are presented with arguments incongruent with their strongly held political beliefs?
H0: Priming for directional motives does not affect the magnitude of respondents' belief strength and certainty.
H1: Respondents primed for directional motives will differ in average post-treatment belief strength and certainty from respondents primed for accuracy motives.

3) Describe the key dependent variable(s) specifying how they will be measured.
Post-treatment attitudes: 7-point Likert scale rating agreement to a one-sentence political belief based on respondent input
Post-treatment certainty: 0-100 sliding scale rating certainty about a one-sentence political belief based on respondent input
Seconds spent on thought listing task: Qualtrics timer question measuring how long respondents spent on the thought listing page before moving onto the next page
Number of thoughts: code from respondents' open-ended responses to the thought listing task
Share of denigrating arguments: (number of thoughts in thought listing responses that critique the statements presented) / (number of thoughts in total)
Difference-in-variances for attitudes and confidence: use Levene's Test to compare variances between directional and accuracy conditions for the post-treatment attitude measure and post-treatment certainty measure

4) How many and which conditions will participants be assigned to?
6 conditions total = 2 priming conditions (directional vs. accuracy motives) x 3 levels of information (4 pro arguments vs. 2 pro and 2 con vs. 4 con)

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
1. Estimate main effects
Regress post-treatment attitudes on pre-treatment attitudes and information condition indicators
Regress post-treatment attitudes on pre-treatment attitudes and motivation condition indicator
Regress post-treatment certainty on pre-treatment attitudes and information condition indicators
Regress post-treatment certainty on pre-treatment attitudes and motivation condition indicator
Regress seconds spent on thought listing task (logged) on information condition indicators
Regress seconds spent on thought listing task (logged) on motivation condition indicator
Regress number of thoughts on pre-treatment attitudes and information condition indicators

Regress number of thoughts on pre-treatment attitudes and motivation condition indicator
Regress share of "denigrating" arguments in thought listing task on pre-treatment attitudes and information condition indicators
Regress share of "denigrating" arguments in thought listing task on pre-treatment attitudes and motivation condition indicator
Estimate difference-in-variances for attitudes and confidence (use Levene's Test) comparing directional to accuracy conditions (i.e., two variance estimates)

2. Estimate interaction effects
Regress post-treatment attitudes on pre-treatment attitudes, information condition indicators, motivation condition indicator, and information x motivation
Regress seconds spent on thought listing task (logged) on pre-treatment attitudes, information condition indicators, motivation condition indicator, and information x motivation
Regress post-treatment certainty on pre-treatment attitudes, information condition indicators, motivation condition indicator, and information x motivation
Regress number of thoughts on pre-treatment attitudes, information condition indicators, motivation condition indicator, and information x motivation
Regress share of "denigrating" arguments on pre-treatment attitudes, information condition indicators, motivation condition indicator, and information x motivation

All models above:
Split prior attitudes, certainty, and political sophistication into terciles

All models will be estimated using OLS with HC2 robust standard errors.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. We will analyze cases where generic_flag does not equal 1.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Target enrollment: 2,000 respondents.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We plan to use structural topic models to assess how the use of topics varies as a function of the motivation condition. We also plan to assess conditional effects using text analysis models that interact the information and motivation conditions. We will perform an exploratory analysis on observations flagged with generic_flag = 1, looking for differences in post-treatment beliefs between respondents shown tailored and generic responses to their prompts.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

D.2 Pre-Registration for Experiment 2 (AsPredicted #109299)

URL: https://aspredicted.org/C4T_FG2

Created: 10/12/2022 11:14 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Q1: Do respondents update their attitudes and certainty of attitudes less when presented arguments incongruent with their strongly held political beliefs than when they are presented arguments incongruent with political beliefs they feel mostly ambivalent about?
H0: Strength of prior attitude does not affect the magnitude of change in respondents' attitude strength and certainty.
H1: On average, respondents will differ in how much they update post-treatment attitude strength and certainty between strong prior attitudes and weak prior attitudes.

Q2: Do respondents who have been primed for directional motives update their attitudes less than respondents who have been primed for accuracy motives when both groups are presented with arguments incongruent with their weakly held political beliefs?
H0: Priming for directional motives does not affect the magnitude of respondents' belief strength and certainty.
H1: Respondents primed for directional motives will differ in average post-treatment belief strength and certainty from respondents primed for accuracy motives.

3) Describe the key dependent variable(s) specifying how they will be measured.
Post-treatment attitudes: 7-point Likert scale rating agreement to a one-sentence political belief
Post-treatment certainty: 0-100 sliding scale rating certainty about a one-sentence political belief
Post-treatment certainty (multi-item): index from seven 9-point Likert scales rating certainty about a one-sentence political belief
Post-treatment combined attitude measure: measure combined from (1) and (2) to produce a -100 to 100 201-point scale
Seconds spent on thought listing tasks: Qualtrics timer question measuring how long respondents spent on the thought listing page before moving onto the next page
Number of thoughts: code from respondents' open-ended responses to the thought listing task
Share of denigrating arguments: (number of thoughts in thought listing responses that critique the statements presented) / (number of thoughts in total)
Argument strength rating: 7-point Likert scale rating strength of a set of 4 pro and 4 con arguments
Argument accuracy rating: 4-point Likert scale rating accuracy of a set of 4 pro and 4 con arguments
Difference-in-variances for attitudes and confidence: use Levene's Test to compare variances between directional and accuracy conditions for the post-treatment attitude measure and post-treatment certainty measure

4) How many and which conditions will participants be assigned to?
12 conditions total = 2 priming conditions (directional vs. accuracy motives) x
3 levels of information (4 pro arguments vs. 2 pro and 2 con vs. 4 con) x

2 levels of belief strength (strong belief and weak belief (within-subjects))

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
To assess the first hypothesis, we plan to estimate the following models for the single-item attitude strength (outcome 1) and multi-item certainty scale (outcome 3): we will regress both outcomes on pre-treatment measures of each outcome, an information condition indicator, an attitude strength condition indicator, and their interaction.
To assess the second hypothesis, we plan to estimate the following model for the attitude strength and multi-item certainty outcomes: we will regress both outcomes on pre-treatment measures of each outcome, a motivation condition indicator, an information condition indicator, an attitude strength condition indicator, and lower and higher-order interactions for all experimental factors.
To validate that the information and motivation conditions were effective, we will regress outcomes 5-7 on pre-treatment attitudes, pre-treatment certainty (single item), a motivation condition indicator, an information condition indicator, an attitude strength condition indicator, and lower and higher-order interactions for all experimental factors.
Outcomes 8-9 will be analyzed by examining the mean ratings for pro arguments, the mean ratings for con arguments, and the mean difference between pro and con ratings. We will regress this outcome on pre-treatment attitudes, pre-treatment certainty (single item), a motivation condition indicator, an information condition indicator, an attitude strength condition indicator, and lower and higher-order interactions for all experimental factors.
All models will be estimated using OLS with CR2 robust standard errors (clustered by respondent).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. We will analyze cases where generic_flag does not equal 1.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Target enrollment sample: 1,200 respondents.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We plan to use structural topic models to assess how the use of topics varies as a function of the motivation and attitude strength conditions. We also plan to assess conditional effects using text analysis models that interact the information, motivation, and attitude strength conditions. We will perform an exploratory analysis on observations flagged with generic_flag = 1, looking for differences in post-treatment beliefs between respondents shown tailored and generic responses to their prompts. We may conduct an argument-level analysis that regresses each argument individually on the experimental conditions above. We may also use outcomes (2) and (4) to examine the robustness of our findings to alternative measures of certainty. We also plan to estimate difference-in-variances for attitudes and confidence (using Levene's Test), comparing directional to accuracy conditions (i.e., two variance estimates) for weak and strong conditions separately. We also plan to assess subgroup differences based on prior attitudes, certainty, and political sophistication. We will discretize these variables into tertiles and compare the lower to upper tertile.

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

D.3 Pre-Registration for Experiment 3 (AsPredicted #116238)

URL: https://aspredicted.org/QRV_BYR

Created: 12/12/2022 11:59 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Do people exposed to counter-attitudinal information polarize or moderate their attitudes? We use GPT-3 to generate tailored counterarguments for each respondent to assess if strong attitudes are susceptible to attitude polarization.

3) Describe the key dependent variable(s) specifying how they will be measured.
Attitude strength (7-point Likert item ranging from "strongly disagree" to "strongly agree")
Attitude certainty (multi-item): index from seven 9-point Likert items rating certainty

4) How many and which conditions will participants be assigned to?
Two conditions: placebo condition (random pseudo-news article drawn from placebo corpus) and counter-attitudinal condition

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Main analysis: We plan to regress both outcomes on age, education, income, gender, ideology, political party, and race (white =1) with HC2 robust standard errors (using OLS regression). If ATE > 0 (ATE < 0), this is evidence of attitude polarization (moderation).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. In addition, GPT-3 will generate a rating score indicating the quality of the open-ended response (0 = low quality; .5 = medium quality; 1 = high quality). We will analyze cases where generic_flag does not equal one and quality does not equal zero.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We plan to analyze a sample of 2000 participants.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We also plan to conduct tests of heterogeneous treatment effects by major topic (subsetting on topic and estimate CATEs within subsets) and subtopic (using hierarchical models).

D.4 Pre-Registration for Experiment 4 (AsPredicted #146475)

URL: https://aspredicted.org/NB2_DDP

Created: 10/09/2023 11:57 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Do people exposed to counter-attitudinal information polarize or moderate their attitudes? We use GPT-3.5 to generate tailored counterarguments for each respondent to assess if strong attitudes enhance attitude polarization in the face of counter-attitudinal content.

3) Describe the key dependent variable(s) specifying how they will be measured.
We plan to measure attitude strength using two new measures we developed: an attitude defense scale capturing one's willingness to defend an attitude across different social contexts and an attitude extremity scale that assesses the level of commitment individuals are willing to invest in supporting their attitude. The attitude defense scale is a five-item summative scale (α = .9), whereas the attitude extremity scale is an 8-item summative scale (α = .8).

4) How many and which conditions will participants be assigned to?
Two conditions: placebo condition (random pseudo-news article drawn from placebo corpus) and counter-attitudinal condition.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Main analysis: We plan to regress all outcomes on pre-treatment measures of age, political sophistication, gender, education, ideology, political party, attitude certainty, 7-point attitude strength, self-reported political behavior, and race (white =1) with HC2 robust standard errors (using OLS regression). If ATE > 0 (ATE < 0), this is evidence of attitude polarization (moderation).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. We will analyze cases where generic_flag does not equal one.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We plan to analyze a sample of 2000 participants.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will use the 7-point Likert scale and 0-100 certainty scale to verify that we have recovered strong attitudes and as pre-treatment covariates, rather than analyzing them as outcome variables. We specifically developed the two scales above – attitude defense and attitude extremity – to address

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

potential ceiling effects affecting the more traditional measures.

We will also perform a conditional test and perform the main analysis while limiting our sample to people who score a 7 on the 7-point attitude strength measure. This removes potential cases in which the AI model made a random error while summarizing the subject's strongly held attitude.

D.5 Pre-Registration for Experiment 5 (AsPredicted #147883)

URL: https://aspredicted.org/PMY_DTH

Created: 10/19/2023 03:18 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Do people exposed to counter-attitudinal arguments polarize or moderate their attitudes? Is polarization's occurrence contingent upon how strongly worded the counterarguments are? We use GPT-3.5 to generate tailored counterarguments for each respondent to assess if strong attitudes enhance attitude polarization in the face of counter-attitudinal content.

3) Describe the key dependent variable(s) specifying how they will be measured.
As in Study 4, we plan to measure attitude strength using two new measures we developed: an attitude defense scale capturing one's willingness to defend an attitude across different social contexts and an attitude extremity scale that assesses the level of commitment individuals are willing to invest in supporting their attitude.

4) How many and which conditions will participants be assigned to?
Three conditions: placebo condition (random pseudo-news article drawn from placebo corpus), weakly worded counter-attitudinal condition, and strongly worded counterattitudinal condition. The third condition differs from the second in that the counterarguments will directly and harshly attack the respondent's issue position (as in Study 4).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Main analysis: We plan to regress all outcomes on pre-treatment measures of age, political sophistication, gender, education, ideology, political party, attitude certainty, 7-point attitude strength, self-reported political behavior, and race (white =1) with HC2 robust standard errors (using OLS regression). If ATE > 0 (ATE < 0), this is evidence of attitude polarization (moderation).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We plan to analyze the subset of respondents who provided a valid prompt and obtained tailored responses. Qualtrics will flag if the respondent is shown generic responses (if GPT-3 could not produce output in response to their prompt), returning generic_flag = 1. We will analyze cases where generic_flag does not equal one.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
2000

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will use the 7-point Likert scale and 0-100 certainty scale to verify that we have recovered strong

APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template APSR Submission Template

attitudes and as pre-treatment covariates, rather than analyzing them as outcome variables. We specifically developed the two scales above – attitude defense and attitude extremity – to address potential ceiling effects affecting the more traditional measures.

We will also perform a conditional test and perform the main analysis while limiting our sample to people who score a 7 on the 7-point attitude strength measure. This removes potential cases in which the AI model made an error while summarizing the subject's strongly held attitude.

After answering outcome questions, respondents will be shown both the weakly worded and strongly worded counterarguments and will be asked to provide open-ended thought listings in response to each. To explore the mechanisms by which harsh wording may contribute to attitude polarization, we will conduct exploratory analyses of thought listings, comparing respondents' thought listing text on various measures: level of valence/affect (using VADER sentiment analysis), amount of defensive language (coded by GPT), effort (time spent and length), and share of words in thought listing denoting a feeling of being attacked by the counterarguments ("insult", "attack", "hurt", "offend", "disrespect").