Supplemental Online Material

Overestimating the Social Costs of Political Belief Change

Table of Contents

Table S1: Preregistration Deviation Table

Adapted from Willroth & Atherton (2024)

Preregistration deviations						
#	**Details**		**Original wording**	**Deviation description**	**To what extent is this a deviation from the preregistered plan?**	**Judgment of impact**
1	Study	# Supp. S1	We will exclude participants from our final analysis who are unable to correctly answer both attention checks within two attempts.	Typo: This should read "comprehension check", rather than attention check.	Minor	None
	Type	Other (Please Explain)				
	Reason	Typo/Error				
	Timing	Other (Please Explain)				
2	Study	# 3 (phase 1)	We hypothesize that there will be a significant interaction between Role and the direction of belief change. Specifically, we hypothesize that the tendency for participants in the Predict condition to overestimate how much participants in the React condition will reject them will be stronger in the Weaken Belief-	After completing these analyses, we realized that they were conceptually flawed (although the results supported our hypotheses). Specifically, it would not be appropriate to compare predict condition participant scores across levels of belief change (weaken, no change, strengthen). We	Major	This represents a major empirical decision; however, it does not represent a major conceptual impact on the findings or their interpretation. Given that our manuscript focuses specifically on dissenting belief change, this change is in accordance with that scope.
	Type	Methods				

	Reason	New knowledge	Change condition than in the other two conditions.	originally intended to treat the different belief change outcomes as an independent variable; however, given the lack of random assignment among predictors to belief change conditions (i.e., one cannot experimentally assign people to experience a certain type of belief change), this approach is flawed. Specifically, it is possible and likely that some person-level variable simultaneously influenced predictors' a) belief change outcomes, and b) sensitivity to rejection. Thus, we shifted our approach to focus entirely on the subset of participants who experienced dissenting belief change.		
	Timing	After results known				
3	Study	# 3 (phase 2)	Wave 2: We will conduct an independent samples t-test of our dependent measure to examine whether participants in the salience vs control condition predict higher levels of rejection.	We pre-registered that we would conduct an independent samples t-test, but we realized later that a 3-cell ANOVA that also included Phase 1 react condition scores would be appropriate and more theoretically meaningful. The results of the independent samples t-tests as pre-registered were significant and in support of our hypotheses and we report them in the main text alongside the ANOVA results.	Minor	We judge that this deviation does not impact the interpretation of these findings.
	Type	Methods				
	Reason	New knowledge				
	Timing	After results known				

4	Study	# 2-5	We pre-registered a broad set of exploratory moderation analyses under Section 8 (“exploratory analyses”) in each pre-registration. Our intent was to ensure that we had a comprehensive plan that covered all possible avenues of inquiry, anticipating that these analyses might be relevant as the research evolved.	As the project progressed and our understanding deepened, it became clear that some of these exploratory analyses no longer aligned with the central aims and theoretical focus of the current research.	Minor	We judge that this deviation does not impact the interpretation of these findings.
	Type	Analyses				
	Reason	Plan not possible				
	Timing	After data access				
5	Study	# 3	N/A	We did not pre-register that we would combine samples for Study 3 because we did not foresee that there would be complications with achieving a sufficient sample size. Due to challenges in collecting a large sample of participants who reported dissenting belief change following the counter-attitudinal writing task, participants in Study 3 were collected from three separate studies with a similar experimental procedure. The methods and findings for each	Major	Minor impact. The key features of the studies were virtually identical, and the direction of the results for the key hypothesis tested was consistent across all samples included.

				of the three studies are nearly identical and are reported in full in the SOM.		
	Type	Analyses				
	Reason	Plan not possible				
	Timing	After data access				
6	Study	# 3	Wave 2 will be conducted two weeks after the conclusion of Wave 1 and will have two primary conditions: Salience vs Control.	This should read *2 months*, not 2 weeks.	Minor	Minor impact. Two months provides a much longer time period for participants to forget about the procedures of Wave 1 of the study, allowing for a better test of the hypothesis in question.
	Type	Methods				
	Reason	Typo/Error				
	Timing	Before data collection				
	Study	# 5	We plan to recruit 420 participants, with 120 participants per condition, anticipating some exclusions and dropouts to meet our per-topic cell minimums.	A far greater number of participants met our pre-registered sampling criteria, yielding a final sample of N = 620 in Study 5.	Minor	Minor impact, increased power.
	Type	Other (Please Explain)	Recruitment			
	Reason	Plan not possible				

	Timing	During data collection				
	Study	# S2	This pre-registration includes language about psychological distancing.	Helpful comments throughout the review process led us to re-interpret data from this study as a robustness check, consistent with the methods used by Savitsky et al., 2003.	Minor	Minor impract; shift in theoretical interpretation of data.
	Type	Interpretation				
	Reason	New knowledge				
	Timing	After results known				

Wilroth, E. C., & Atherton, O. E. (in press). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science.* Retrieved from https://osf.io/preprints/psyarxiv/dwx69

Supplemental Study 1

To explore the possibility that predictors' overestimation of social sanctions is driven by misperceptions about ingroup attitude extremity, Supplemental Study 1 tested whether providing predictors with accurate information about the attitudinal extremity of their react condition counterpart would lead to more accurate estimates of social sanctions for dissenting belief change. Additionally, Supplemental Study 1 included a measure of self-censorship among predictors to further investigate the relationship between anticipated social sanctions and self-censorship.

Methods

Design. Supplemental Study 1 employed a 2 (role: predict vs react) × 2 (extremity: extreme vs moderate) between-participants design.

Participants. Participants were recruited from Prolific. First, we conducted a pre-screen survey which served identify eligible participants to recruit for Supplemental Study 1. Per our pre-registration, we aimed to recruit a final sample of 400 participants. This sample size was determined by power analysis using G*power to estimate the minimum sample size required to observe an interaction between conditions for a two-way ANOVA with an effect size of *Cohen's d* = .03 at 95% power.

During the pre-screen survey, participants were asked to report the following: (1) Their partisan identity, (2) which of the following statements best represents their views on abortion access: "Abortion access in this country should be [**protected/restricted**]", and (3) how strongly they agree with the statement that they selected on a 0-100 scale (0 = *Completely Disagree;* 100 = *Completely Agree*). We used custom recruitment filters on Prolific to recruit a sample of 2,069

Democrat and Republican participants to complete the pre-screen with the goal of identifying at least 240 (120 Republicans; 120 Democrats) "extreme" participants who indicated that they agreed with the statement they selected on abortion between 95-100 on the agreement scale and at least 240 (120 Republicans; 120 Democrats) "moderate" participants who indicated they agree with the statement between 50-85 to re-recruit for to participate in Supplemental Study 1.

Participants who reported that they held the typical view of their party on the topic of abortion and agreed with that view between either 50-85 or 95-100 on the attitude agreement measure during the pre-screen were eligible to participate in Supplemental Study 1. Among these, we randomly selected an equal number of Democrats and Republicans who met our recruitment criteria in each category (n = 515) and invited them to participate in the study twenty-four hours later[1]. Of these, 435[2] accepted our invitation to participate in the study and 401 participants met all criteria to be included in our final sample (221 Democrats, 190 Republicans; 213 females, 185 males, 3 self-identify; M_{age} = 43.05; SD_{age} = 13.92).

Our sample consisted of near-equal numbers of participants who agreed with the typical view of their party either between 50-85 (moderates; n = 200) or between 95-100 (extreme; n = 201) on the 0-100 agreement scale. We specified these attitude-based grouping criteria based on participant responses to previous studies and feasibility. Regarding the latter, we observed a high concentration of participants who reported that they strongly agreed with the typical view of their party in previous studies, whereas participants reporting more moderate views were distributed

[1] There were no significant differences between the randomly selected subset of eligible participants from the pre-screen survey and those who were not randomly selected for participation in Supplemental Study 1 in terms of their agreement with the statement or their political identity centrality. We report these analyses in the in the Additional Analyses section later in the OSM.

[2] There are no significant differences between the 435 participants who accepted the Supplemental Study 1 invitation and the 80 participants who did not. These analyses are also reported later in the OSM.

across a wider range of agreement. After reviewing the distribution of responses to this item in the pre-screen data, we realized that it would not be feasible to recruit a sufficient sample size of "moderate" participants if we narrowly defined the range by five scale points (e.g., 75-80), as in the "extreme" category. Therefore, we created a relatively narrower grouping category for "extreme" agreement and a broader category for "moderate" agreement. We report all data on the sample statistics for the pre-screen and Supplemental Study 1 in the SOM.

Procedure

Employing the same paradigm from previous studies, we randomly assigned participants to either a predict condition or a react condition at the beginning of Supplemental Study 1. We told predictors that they would be paired with another participant who identified with the same political party and held the same party-typical view as themselves on the issue of abortion access in the United States. Then, we randomly assigned predictors to either a "moderate" partner or an "extreme" partner condition and gave them information about their partner's attitudes, which experimentally varied whether their ingroup partner agreed with the typical view of the party between 50-85 on the agreement scale (which we refer to as the "moderate partner" condition) or between 95-100 on the agreement scale (which we refer to as the "extreme partner" condition).

We informed predictors that their ingroup partner would learn the following information about them: (1) which political party they support, (2) that they were assigned to write a persuasive message in favor of the opposing party's view on abortion access, and (3) that after they completed this writing task, they reported that their agreement with their party's view on abortion access decreased. We then asked predictors to predict how their partner in the react condition would respond upon learning about their dissenting belief change using the same nine-item scale measure of social sanctions from Study 1b.

Participants who were randomly assigned to the react role condition were categorized as either “extreme” (those who responded between 95-100) or “moderate” (those who responded between 50-85) based on their pre-screen survey responses to the 0-100 measure of their agreement with the party-typical view on abortion. This participant-level quasi-independent variable served as the “extremity” condition factor in the react condition. We told reactors the exact information that we informed predictors we would tell them. Specifically, we told reactors that they were paired with another participant taking the study who identified with the same political party and agreed with the same party-typical view on the issue of abortion as themselves. We then told reactors that this person was assigned to write a persuasive message in favor of the opposing party’s point of view on the issue of abortion as part of the study, after which they reported that their agreement with the typical view of the party decreased. Reactors then reported social sanctions towards their partner on the same nine-item composite measures as the predictors (with modified wording so that reactions were captured instead of predictions).

We also included a behavioral measure of economic punishment in the form of a dictator game. We told participants that they would be randomly assigned to the distributor (dictator) versus receiver role for a financial bonus; however, in reality, reactors were always assigned to the distributor role. This paradigm created a financial incentive for reactors to punish or reward the dissenting ingroup member they were evaluating, wherein any amount they took from their partner represented a gain for themselves. For this item, predictors estimated how much of a $1.00 bonus their partner would give them versus keep for themselves and reactors distributed the bonus between themselves and their partner.

Next, after predictors completed all other dependent measures, they responded to a single item measuring how likely they would be to self-censor if their beliefs on abortion actually

changed in the way they were asked to consider in the study. Specifically, predictors were asked, "If your agreement with the statement [piped text showing the typical view of their party] **actually DECREASED**, would you tell that to the person you've been paired with in this study if the topic came up in conversation?" (1 = *I definitely would **NOT** tell them*; 7 = *I definitely **WOULD** tell them*; emphasis original).

Finally, predictors responded to a manipulation comprehension check that asked them to estimate the attitude strength of their partner in the study, two attention checks, and basic demographic questions. All participants were debriefed at the end of the study.

Results

Manipulation Check. We found that predictors in the moderate partner condition estimated that their partner held relatively more moderate attitudes (M = 72.01; SD = 13.93) compared to predictors' estimates in the extreme partner condition (M = 89.34; SD = 18.84), $t(197) = 7.37$, $p < .001$, $d = 1.05$. We note that a small number of predictors failed this manipulation check in both the moderate condition (n = 10) and the extreme condition (n = 26). We re-ran our key analyses with these participants excluded and replicated an identical pattern of results, therefore the following analyses are reported with these participants included.

Social Sanctions. As the main test of our hypothesis, we conducted a 2 (role: predict vs. react) × 2 (extremity: moderate vs. extreme) ANOVA on our composite measure of social sanctions ($\alpha = .97$) to detect whether a perception gap was present between predictors and reactors, and, if so, whether the magnitude of the perception gap was greater for participants in the moderate condition versus extreme condition. In this analysis, the "extremity" variable took on the value of "partner extremity" for predictors and "participant extremity" for reactors. With

this yoked design, we were able to match the predictions that predictors made about moderate and extreme ingroup members, respectively, with the social sanction scores that moderate and extreme reactors reported.

This analysis provided further support for the perception gap, revealing a main effect of role wherein predictors ($M = 4.12$; $SD = 1.36$) significantly over-estimated sanctions relative to reactors' reports ($M = 2.47$; $SD = 1.55$), $F(1, 397) = 133.22$, $p < .001$, $\eta^2 = 0.25$. We also observed a main effect of extremity, $F(1, 397) = 16.18$, $p < .001$, $\eta^2 = 0.04$; however, the interaction of role × extremity was not statistically significant, $F(1, 397) = 1.81$, $p = .18$, $\eta^2 = 0.01$ (see figure below). The lack of interaction in this model suggests that the effect size of role (i.e., the perception gap) was of comparable magnitude in the moderate and extreme conditions, suggesting that the main effect of overestimation by role was not driven by partner extremity.

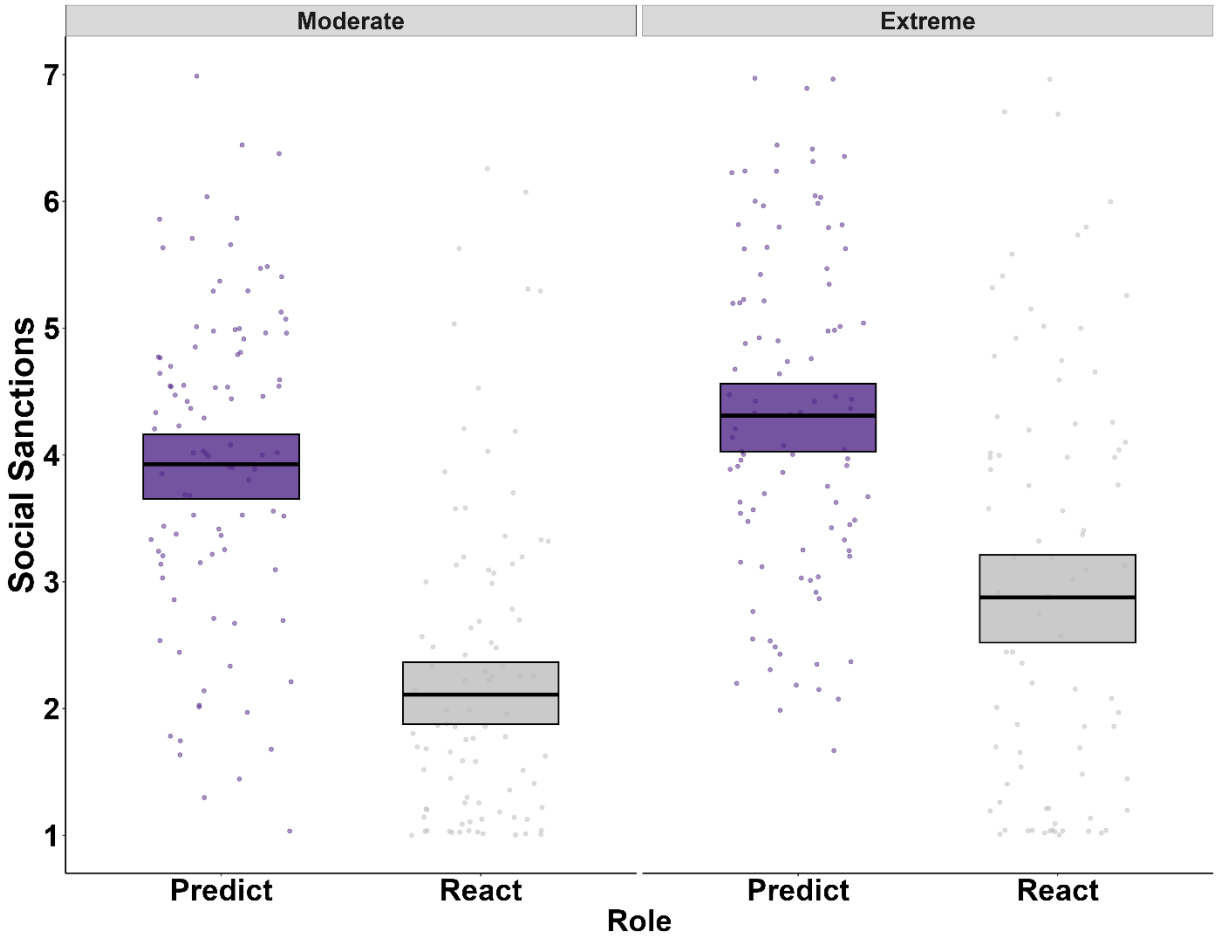


Figure. *Predictors report higher social sanction scores than reactors across extremity conditions. Note.* The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval. Anticipated and actual social rejection were measured on a seven-point response scale (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*).

One plausible explanation for these results is that predictors in the moderate-partner condition may have anchored on the upper-bound (e.g., '85') of the range of attitude strength provided, which, in turn, would have inflated estimates of ingroup member attitude strength relative to the true distribution in the moderate-react subset of participants. To address the concern that predictors in the moderate condition may have anchored to this upper bound, we analyzed predictors' estimates of their partners' attitude strength, which were measured on a 1–100 scale (see Table below). These estimates allow us to test whether predictors in the moderate

condition were anchoring on the upper bound of the assigned range (i.e., 85), thereby conceptualizing their partner as more extreme than intended.

Supplementary Table. Predictor Estimates of Partner Extremity Against Actual Partner Extremity.

Extremity Condition	Mean (SD) Predicted Extremity	Mean (SD) Actual Extremity
Moderate (50–85)	72.0 (13.9)	73.8 (9.7)
Extreme (95–100)	92.5 (8.4)	95.9 (4.7)

On average, predictors in the moderate condition estimated their partner's attitude strength to be 72.01 (*SD* = 13.93). This estimate closely matched the actual average extremity score of reactors in the moderate condition (*M* = 73.81; *SD* = 9.74) and was not significantly different from it, $t(204) = 1.07$, $p = .285$, $d = .15$. Moreover, only eight predictors in this condition estimated their partner's extremity to be 85 or higher, suggesting that most participants did not anchor on the upper end of the range. These findings do not support the alternative explanation that the observed overestimation effects were driven by predictors calling to mind more extreme partners than those they were actually paired with.

Behavioral Punishment. We conducted a 2 (role: predict vs. react) × 2 (extremity: moderate vs. extreme) ANOVA to examine whether predictors in our sample overestimated how much reactors would punish them for dissenting belief change, and, if so, whether the magnitude of this overestimation was greater based on the extremity condition. This analysis revealed a significant main effect of role, $F(1, 397) = 14.10$, $p < .001$, $\eta p^2 = 0.034$, indicating that predictors anticipated that their react condition partner would keep significantly more of the bonus money (*M* = 72.31*; SD* = 22.17) than reactors actually kept (*M* = 63.73; *SD* = 23.53). The main effect of extremity was not significant, $F(1, 397) = 1.61$, $p = .205$, $\eta^2 = 0.004$, suggesting no overall

difference between the moderate and extreme conditions. Additionally, interaction of role × extremity was also not statistically significant, $F(1, 397) = 0.09$, $p = .760$, $\eta^2 < 0.001$.

Self-Censorship. Participants in the predict condition were asked whether they would share information about their belief change with their partner or self-censor if they actually updated their beliefs on the topic. This item was reverse-coded for a more intuitive interpretation in our analysis, with higher scores representing a greater likelihood of self-censorship ($M = 3.96$; $SD = 1.93$). We conducted Pearson's correlational tests to examine the relationship between the likelihood of self-censorship, anticipated social sanctions, and anticipated economic punishment in the dictator game. These analyses revealed that predictors reported a greater likelihood of self-censorship when they anticipated harsher social sanctions ($r = .22$, $p = .002$) and greater economic punishment in the dictator game ($r = .26$, $p < .001$). These findings provide further evidence in support of the relationship between fear of social sanctions and self-censorship observed in Pilot Study 1.

As an exploratory analysis, we conducted an independent samples t-test to examine whether predictor self-censorship varied by partner-extremity condition. Predictors who were paired with a moderate partner ($M = 4.15$, $SD = 1.83$) reported directionally lower self-censorship than those paired with an extreme partner ($M = 3.76$, $SD = 2.01$); however, this difference was not statistically significant, $t(196) = 1.44$, $p = .152$, $d = 0.20$.

Discussion

Providing predictors with accurate information about their partner's attitudes did not close the perception gap. Instead, predictors in both extremity conditions significantly overestimated how much ingroup reactors would socially sanction them for dissenting belief

change. We found a similar pattern of results for our behavioral measure of punishment. Correlational findings showed that greater self-censorship likelihood was associated with increased estimates on both survey and behavioral rejection measures. Taken together, these findings suggest that the overestimation of social sanctions is not driven by an overestimation of ingroup extremity but by other psychological factors.

Supplemental Study 2

As an additional test of the robustness of the overestimation effect observed in Studies 1-5 in the main text, we conducted a pre-registered study in which participants were randomly assigned to either a predict condition, a react condition, or a third-party perspective-taking condition. In this third-party perspective taking condition, participants were asked to imagine a partisan ingroup member who had changed their view on a controversial political topic and to predict how *another ingroup member* (i.e., a third-party) would react to that person. This methodological approach has been employed in previous self-other perception gap studies (e.g., Savitsky et al., 2001) to mitigate the potential for self-presentational concerns among reactors (e.g., social desirability bias), as reactors should be more forthcoming about reporting a harsh evaluation of a dissenting ingroup member when self-presentation concerns (i.e., the desire to appear open-minded, tolerant) are reduced.

The core overestimation effect replicated in this study: Predictors anticipated significantly more social rejection than reactors reported, even when reactors did so from a third-party perspective. The methods and results of this study are described in greater detail below.

Method

Participants. Supplemental Study 2 used a 3-cell between-participants design. We recruited 700 participants to yield at least 200 participants per cell after making exclusions for failed attention and comprehension checks per our pre-registered exclusionary criteria. This sample size was determined based on a priori power analysis that indicated 200 participants per cell would provide sufficient power to detect our smallest hypothesized effect size ($f = .15$) at 90% power with alpha set to .05. Far fewer participants failed attention and comprehension checks than we anticipated, yielding a final sample of $N = 693$ after making such exclusions (510

Democrats, 183 Republicans; 426 females, 260 males, 7 self-identify; M_{age} = 41.03; SD_{age} = 13.60).

Procedure. Similar to previous studies, Supplemental Study 2 began with a brief set of political attitude questions wherein participants reported their partisan identity and their attitudes on three central political topics (abortion, gun control, and immigration). Using the same procedure as Study 1, we then randomly assigned participants to consider one of the topics upon which they reported a view that was consistent with the typical view of their political party as the focal topic in the study.

We then randomly assigned participants to either a predict condition, a third-party condition, or a react condition. Participants in the predict and react conditions completed the same study procedure as predictors and reactors in Study 1. Participants in the third-party condition were asked to predict how an ingroup member would react if another ingroup member (named Sam) adopted the opposing party's view on the randomly assigned topic. An example of the third-party condition for a Republican participant assigned to the gun control condition is shown below (emphasis original).

> Another **Republican** participant taking this survey named Sam indicated that they believe that access to private gun ownership in this country should be **protected**.
>
> In the next part of the survey, we are going to ask you questions about how you think **another Republican on Prolific** would react if they found out that Sam changed their mind on this issue and now believes that access to private gun ownership in this country should be **restricted.**

After reading condition-specific instructions, all participants responded to the five-item social rejection composite measure that was used in Study 1. Participants also completed two attention checks. The first attention check asked participants about the political topic they were asked to consider during the study. The second attention check asked participants about the party identity of the person they read about during the study[3].

We hypothesized that we would detect a significant main effect of condition (predict vs. third-party vs. react) on the social rejection composite such that participant scores would be highest in the predict condition, second highest in the third-party condition, and lowest in the react condition.

Results

To explore our main hypothesis, we conducted an ANOVA to determine the effect of condition on social rejection composite scores. In support of our central hypothesis, we detected a significant effect of condition, $F(2, 690) = 43.45$, $p < .001$, $\eta^2 = 0.11$. Next, we conducted planned orthogonal contrasts to test our hypotheses that participants in the predict condition would report higher scores than participants in the react and third-party conditions. In a replication of the primary overestimation finding from the previous studies, the first planned contrast revealed that scores in the predict condition ($M = 4.06$, $SD = 1.52$) were significantly higher than scores in the react condition ($M = 2.75$, $SD = 1.57$), $t(690) = -3.35$, $p < .001$, $d = .85$. The second planned contrast showed that scores in the predict condition were significantly higher than scores in the third-party condition ($M = 3.58$, $SD = 1.52$), $t(690) = 5.82$, $p < .001$, $d =$

[3] We note that we included additional exploratory measures that are not relevant to the present research.

.32. The third planned contrast indicated that scores in the third-party condition were significantly higher than scores in the react condition, $t(690) = -9.21$, $p < .001$, $d = .53$.

Discussion

Supplemental Study 2 provided further evidence for the robustness of the overestimation effect. Even when reactors reported evaluations from a third-party perspective, predictors still overestimated the extent of social rejection. This finding provides further evidence that the overestimation effect is not merely a product of social desirability bias or reputational concerns among reactors. While the magnitude of the perception gap was relatively smaller than the predict versus react comparison, the direction and significance of the effect were consistent. These results add complementary support to our broader conclusion that individuals systematically overestimate the social costs of partisan belief change. Alongside the financially incentivized behavioral data reported in Study 2, these findings add further evidence to suggest that reactors' social desirability concerns do not account for the perception gap.

Pilot Study 1

In Pilot Study 1, we recruited U.S. partisans who reported a dissenting shift in their political beliefs in the past year and asked them (1) how they think an ingroup member would react if they found out about their dissenting belief change, and (2) whether they have become more or less likely to share their true political beliefs when discussing that topic with other members of their political ingroup in the time since their beliefs on the topic changed. We hypothesized that we would detect a significant correlation between the anticipation of negative social consequences for disclosure and the increased likelihood of self-censorship.

Method

Participants

We recruited 150 Democrat and Republican participants from Amazon's Mechanical Turk using CloudResearch, an online survey recruitment platform and tool, who answered 'yes' to brief screening question which asked participants if their political views on any topic have become less aligned with their party's views on that topic in the past year. We excluded participants who failed a comprehension check at the end of the study, which asked participants to select a statement that describes a partisan becoming less aligned with the majority views of their political group (i.e., a dissenting shift in beliefs). We were left with a final sample of N = 131 participants (69% Democrat; 31% Republican; 47% male; 53% female; M_{age} = 43.69; SD_{age} = 12.27).

Procedure

Participants provided informed consent and began the experimental procedure. Participants were asked to write the name of the issue upon which they became less aligned with

the majority view of their group[4]. Next, participants were asked to report how they think another member of their political ingroup would react if this person found out that their views on the topic became less aligned with the majority views of their group using a composite measure of anticipated social sanctions.

Anticipated Social Sanctions

We used the same nine-item scale to measure anticipated social sanctions (α = .97) that was used in Study 2, which consisted of five measures of social rejection (adapted from Cavazza, Pagilaro, & Guidetti, 2014) and four additional items to measure reputation damage (adapted from Levine & Schweitzer, 2015). We chose to use these scale items because social rejection and reputation damage represent social sanctions that most people have a strong desire to avoid (Williams, 2007) and are especially relevant in the realm of intragroup dynamics and attitude deviance (Jetten et al., 2014). The items included, *"If another [**Democrat / Republican**] learned that you changed your beliefs on the topic of [piped-text of the topic listed] in a way that is **less aligned** with the majority view of the [Democrat / Republican] party, to what extent do you think this person would… (1) exclude you, (2) ignore your input, (3) reject you, (4) disrespect you, (5) criticize you (6) dislike you, (7) be upset with you, (8) lose respect for you, (9) lose trust in you?"* (emphasis original; order randomized). All items were answered on a Likert-type scale (1 = *Not at all;* 4 = *A moderate amount;* 7 = *Very much).*

Self-Censorship

To measure self-censorship, we asked participants if they have become more or less likely to share their true beliefs on the topic when discussing it with other ingroup members in

[4]The top three topics listed were immigration (21%), abortion (19%), and the climate change (12%); see Table S1 below for a list of the topics participants listed.

the time since their beliefs of this topic changed. Participants responded to a 1-7 Likert scale (1 = *I am LESS likely to share my views about this topic than I was before I changed my mind;* 4 = *I am equally as likely to share my views about this topic as I was before I changed my mind;* 7 = *I am MORE likely to share my views about this topic than I was before I changed my mind*). Participants reported basic demographic information to conclude the study.

Results

First, we tested normality in the distribution of participant responses to the social sanctions composite items and the self-censorship item. Neither item was normally distributed, so we employed Spearman's correlation test to examine the relationship between these variables. We found that participant responses to the social sanctions composite measure ($\alpha = .97$; $M = 3.61$; $SD = 1.63$) were significantly correlated with responses to the self-censorship measure ($M = 3.88$; $SD = 1.47$), Spearman's $r = -.22$; two-tailed $p = .014$ (see Figure S1 below).

In line with previous research, these findings showed that participants who expected harsher social sanctions from members of their ingroup were more likely to self-censor when discussing a political topic upon which their beliefs drifted away from the majority views of the group. We believe that this study represents a conservative test of the relationship between anticipated rejection and self-censorship behaviors for two reasons. First, participants in this sample were willing to admit to us (and themselves) that their beliefs have shifted away from the majority view of their party. Participants who are willing to admit such a shift are likely to be less fearful of rejection compared to participants who are not willing to admit such changes. Second, people are often unaware of social influences on their behaviors and are unlikely to admit that they are influenced by peer pressure (Cialdini & Goldstein, 2004; Asch, 1956). Thus,

participants in this survey likely underreported the extent to which they self-censor due to social concerns.

Table S1: Topics that participants reported they have changed their minds on and the percentage of participants who listed each topic.

Topic	Count	Percentage
Immigration	31	21%
Abortion	28	19%
Climate Change	17	12%
Middle East	15	10%
Economic Issues	11	7%
Diversity	10	7%
Gun Ownership	9	6%
Foreign Affairs	4	3%
Trump	3	2%
Economy	3	2%
All other topics	16	14%

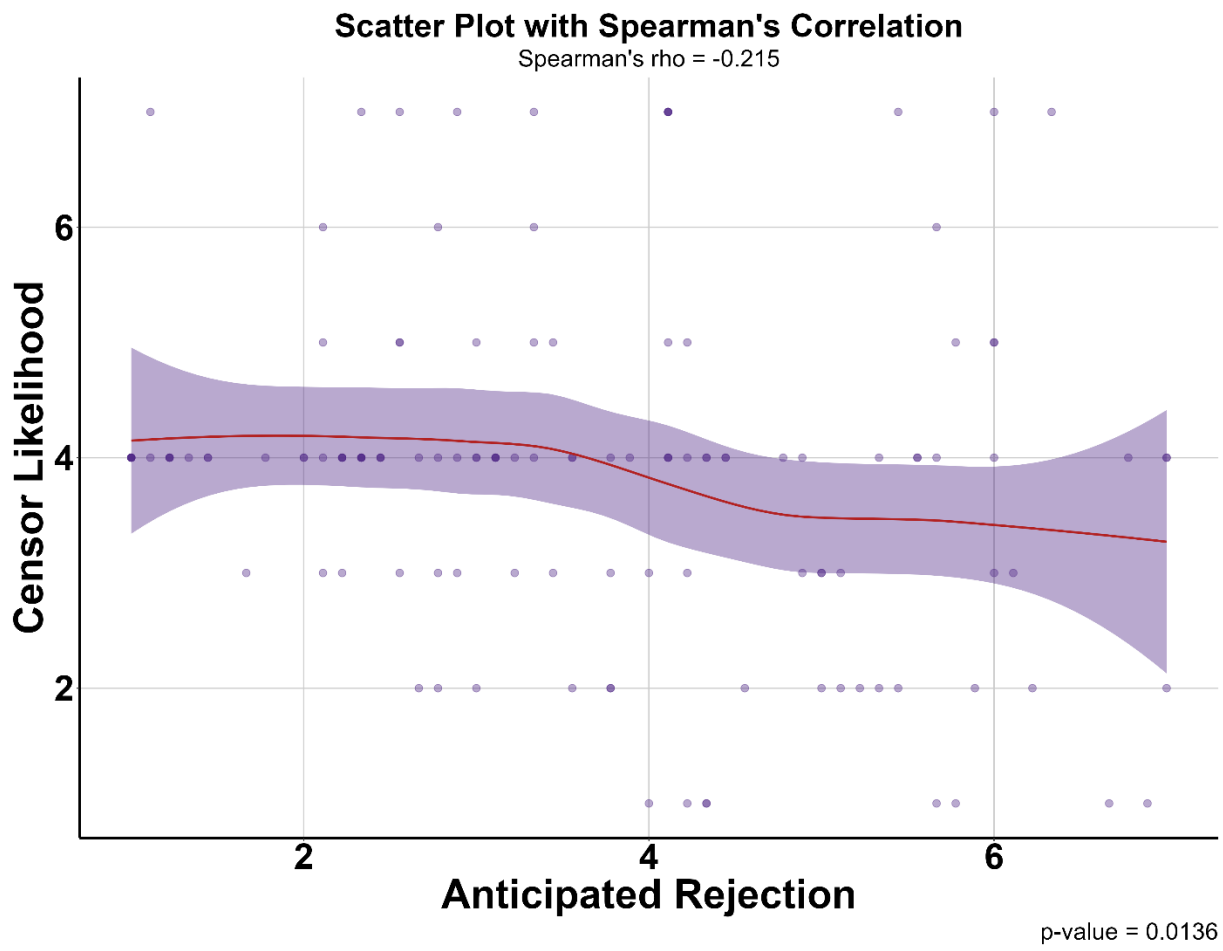


Figure S1. Scatter plot with Spearman's Correlation showing the relationship between anticipated social sanctions and the frequency of self-censorship. Anticipated social rejection was measured on a seven-point response scale (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*). Self-censorship was also measured on a seven-point response scale, which asked participants if they are more or less likely to share their updated views on this topic in conversation with other ingroup members (*1 = I am less likely to share my views, 4 = I am equally likely to share my views on this topic as I was before I changed my mind, 7 = I am MORE likely to share my views).*

Pilot Study 2

Study 1 (reported in the main text) documented that US partisans overestimated the social sanctions they would incur following a complete change in their beliefs, but does this effect generalize to incremental belief change? Pilot Study 2 explored this question by asking participants about an incremental shift in beliefs that came after considering arguments from both sides of the issue. In this study, we define an incremental shift in one's beliefs as becoming slightly less supportive of the typical view of one's group.

Method

Participants

Participants were recruited from Amazon's Mechanical Turk using CloudResearch. We applied the same approach as Study 2 to recruit an equal number of Democrats and Republicans in our sample. The final sample consisted of $N = 101$ participants ($M_{age} = 40.61$; $SD_{age} = 14.76$; 52% female; 48% male; 77% Caucasian; 51% Democrat; 49% Republican).

Procedure

Pilot Study 2 utilized a 2-cell between-subjects design where participants were randomly assigned to either a predict condition or a react condition. First, participants responded to the same political opinion intake survey as in Study 1; however, we also included items asking participants how strongly they agreed with the statement choices they selected (1 = *strongly disagree;* 7 = *strongly agree)*. We used the same procedure that we used in Study 1 to randomly assign one of the three political topics (abortion, gun control, immigration) that participants reported they held their party's view on as the focal topic in the study.

Next, participants received instructions based on their condition. Predictors were told that they would be shown a 5-minute video titled, "The debate on [gun

control/immigration/abortion]: an unbiased introduction to both sides of the issue" from a non-partisan online website called ProCon.org (a real, non-partisan informational website). Predictors were asked to imagine that watching this video would lead them to change their agreement with the statement they selected by one point to the left on the agreement scale that they responded to earlier in the study (e.g., from "*agree*" to "*somewhat agree*"), which was shown on the survey screen for reference. For example, a Democrat predictor who indicated that they strongly agree with the liberal view on abortion would have seen the following text (emphasis original):

> Earlier in this study, you said that you **strongly agree** with this statement on abortion.
>
> **Statement:** "Abortion access in this country should be **protected**."
>
> We'd like you to imagine that watching this video will cause you to change your agreement with this statement from **strongly agree** to **agree** on the scale below.

On the same screen, predictors were shown the same 1-7 Likert-scale measure of agreement that they responded to in the pre-screen survey. Given this description, the shift in beliefs that participants were asked to consider represented the smallest possible amount of dissent from the typical view of the group.

Reactors completed a similar procedure. Reactors were told that they would be shown how an ingroup participant from a previous research study responded to survey items about their beliefs on the randomly selected topic after watching a video. Reactors were given the same description of this video as predictors. Reactors were told that this participant identified with the same political party and held the same view on the randomly selected topic as themselves. Reactors were told that after watching this video, the participant selected the same binary choice option that was aligned with the party, but reported that their agreement with this statement decreased by one point to the left on the 1-7 likert measure of agreement (1 = *strongly disagree;*

7 = *strongly agree*), which was displayed on the screen for reference, similar to the predict condition.

After reading the condition-specific instructions, participants reported anticipated (predictors) and actual (reactors) social sanctions on the same nine-item composite measure from Pilot Study 1 ($\alpha = .97$).

Results

Replicating the pattern of results observed in Study 1, we found that predictors ($M = 3.46$; $SD = 1.33$) overestimated how much ingroup reactors would socially sanction them for incremental belief change compared to reactors' actual reports ($M = 2.14$; $SD = 1.10$), $t(99) = 5.41$, $p < .001$, $d = 1.08$. Similar to Study 1, a follow-up 2 (role: predict vs. react) × 3 (topic: abortion vs. gun control vs. immigration) ANOVA revealed no significant moderation of this effect by topic condition, $p = .315$.

Pilot Study 5

To validate the effect of our loyalty affirmation intervention employed in Study 5, we conducted a pilot study prior to Study 5. This pilot used the same design and measures but included only predictors—no reactor condition—allowing us to isolate the effect of a loyalty affirmation intervention on predictors' loyalty meta-perceptions. The goal was to determine whether reflecting on one's past (dis)loyalty to the political ingroup could shift expectations about how others would evaluate them following dissenting belief change. We note procedural differences from Study 5 where relevant in the sections that follow.

Method

Participants. Our sample size for Pilot Study 5 was determined by an a priori power analysis, which indicated that a sample size of 140 participants would be required to detect an effect size of cohen's d = .50 in an independent samples t test at ninety-percent power. Eligibility was restricted to Democrats and Republicans who reported, via a pre-screen, that they had changed their mind on a central political issue in the past year in a way that placed them at odds with the typical view of their party. Our final sample consisted of 146 U.S. adults (50 Republicans, 96 Democrats; 73 males, 69 females, 4 self-identified participants; M_{age} = 46.97*;* SD_{age} = 12.90) who met our recruitment criteria to participate in the study and passed an attention check, which asked a true/false question about the study procedure.

Whereas Study 5 asked participants about belief change on a set list of five topics, Pilot Study 5 asked (a) a binary yes/no question, and then (b) an open-ended question that asked them to list the topic. This procedure was employed for two reasons. First, it allowed us to generate a list of the most frequently occurring topics of belief change for Democrats and Republicans, respectively, to be used in Study 5. Second, because this study design did not employ a react

condition, we did not need to worry about "matching" predict-react pairs based on topics and therefore did not need to manage the possibility of a wide variety of topics being listed.

*Procedure and Design***.** Participants were randomly assigned to one of two conditions. In the high-loyalty condition, participants listed three actions they had taken in support of their political party. In the low-loyalty condition, participants listed three actions they had taken that went against their party. This was same intervention employed in Study 5.

All participants were then asked to imagine that another ingroup member would evaluate them after learning that they had recently changed their political views. Participants completed the following measures: (a) Loyalty meta-perceptions (e.g., "This person will think I am loyal to my party"; 2 items, $r = .79$), (b) Anticipated social rejection (6-item composite used in Study 5; $\alpha = .94$), (c) Exploratory measures: self-perceived loyalty, perceived alignment, political identity strength, and anticipated liking (bi-polar).

Results

Manipulation Checks and Main Effects. Participants in the high-loyalty condition reported that they felt more loyal to their party ($M = 4.84$; $SD = 1.51$) than the low-loyalty condition ($M = 4.33$; $SD = 1.47$), $t(144) = 2.04$, $p = .043$, $d = 0.34$. This item was administered at the end of the study, making it a conservative test of the manipulation. In Study 5, we placed the item immediately after the intervention to better reflect the theorized causal sequence.

As in Study 5, participants in the high-loyalty condition reported significantly higher loyalty meta-perceptions ($M = 4.73$, $SD = 1.27$) than those in the low-loyalty condition ($M = 3.90$, $SD = 1.27$), $t(144) = 4.17$, $p < .001$, $d = 0.69$. They also anticipated significantly less social rejection ($M = 3.14$, $SD = 1.52$) than those in the low-loyalty condition ($M = 3.92$, $SD = 1.52$), $t(144) = 3.21$, $p = .002$, $d = 0.53$.

Exploratory Measures. Replicating the main effect of social sanctions, scores on the bi-polar measure of anticipated liking were significantly higher in the high-loyalty condition (M = 4.23) than the low-loyalty condition (M = 3.54), $t(144) = 3.77$, $p < .001$, $d = 0.62$. Meta-perceptions about the extent to which a reactor would assume their views on the topic are misaligned with the typical view of the ingroup were not significantly different between the high-loyalty condition (M = 2.05) than the low-loyalty condition (M = 1.81), $t(144) = 1.72$, $p = .087$, $d = 0.29$.

Robustness Checks. The effect of condition on anticipated rejection remained significant when controlling for perceived alignment ($b = 0.34$, $p = .005$) and political identity strength ($b = 0.36$, $p = .004$).

Mediation Analysis. We conducted a mediation analysis using PROCESS Model 4 with 5,000 bootstrap samples with condition as the IV, loyalty meta-perceptions as the mediator, and anticipated social sanctions as the dependent measure. Loyalty meta-perceptions significantly mediated the effect of condition on anticipated rejection:

- Indirect effect via loyalty meta-perceptions: $b = –0.538$, 95% CI [–0.894, –0.284]
- Direct effect: $b = –0.255$, $p = .25$ (non-significant)

Discussion

This pilot study validated the effects of our loyalty affirmation intervention and offered initial evidence supporting the role of loyalty meta-perceptions in shaping predictors' expectations of social rejection. Reflecting on past loyalty-affirming (vs. disloyal) behaviors increased perceived loyalty and, in turn, reduced anticipated social sanctions. The mediation analyses revealed that this effect operated through loyalty meta-perceptions, underscoring the centrality of loyalty-based reputational concerns in driving the perception gap. These findings

informed the design of Study 5 by validating the intervention's efficacy in a controlled setting before testing its effects on actual prediction error in the full paradigm and generated the list of topics that were used in Study 5 (reported in the table below).

Table. Topics of Belief Change Listed in Pilot Study 5. *Note.* The five most frequently occurring topics for Democrats and Republicans, respectively, were used in Study 5.

	Democrat Count	Dem %	Republican Count	Rep %
Immigration	20	28%	7	16%
Gender Issues	16	22%	1	2%
Gaza/Israel/Palestine	14	19%	1	2%
Abortion	9	13%	13	30%
Gun Ownership	6	8%	6	14%
Climate Change	4	6%	10	23%
Economic Policy	3	4%	6	14%

Pre-Test: Studies 4b and 4c

This pre-test aimed to identify political topics that are perceived as moral but do not strongly signal a person's partisan identity. Unlike canonical partisan issues (e.g., abortion), where attitude positions offer strong cues about political affiliation, the goal here was to select topics for use in Studies 4b and 4c that allowed us to manipulate perceived group polarization without activating entrenched partisan priors. Although additional data from this study are used in a separate research program, here we report only the results relevant to topic selection and believability of the polarization manipulation.

Method

Participants. A sample of 301 U.S. Democrats and Republicans was recruited from Prolific to complete a survey on moral and political beliefs. After excluding participants who failed one of three attention checks, our final sample consisted of $N = 217$ participants (111 Republicans, 106 Democrats; 84 males, 131 females, 2 self-identifying participants).

Procedure. Participants were randomly assigned to either a *high polarization* or *low polarization* condition. All participants completed the same topic evaluation survey; the experimental condition only influenced the final portion of the study.

At the start of the survey, participants were presented with five candidate political topics: AI regulation, gambling, government surveillance, cloning, and organ transplantation. For each topic, participants (a) indicated their own view by selecting one of two opposing statements (e.g., "The U.S. government should impose strict regulations on AI surveillance programs" vs. "The U.S. government should *not* impose strict regulations"), then (b) evaluated both someone who held the **opposing** view and someone who held the **same** view on two dimensions: Moral

judgment (−3 = *Very immoral* to 3 = *Very moral*); Perceived party affiliation (−3 = *Definitely a Democrat* to 3 = *Definitely a Republican*). The order of topics was randomized. Participants completed all measures for all five issues.

In the final stage of the survey, participants were shown fabricated "public opinion data" indicating that one of the topics (randomly selected) was either highly polarized or not polarized along party lines. Participants then rated how surprising they found this information (1 = *Not at all surprising*, 7 = *Very surprising*). This served as a proxy for manipulation believability, with lower surprise scores interpreted as evidence that the information was perceived as credible.

This design allowed us to identify issues that were moralized but not strongly associated with partisan identity and test the credibility of the type of polarization framing intervention that would be employed in Studies 4b and 4c.

Results

We identified gambling legalization and AI surveillance as the two most suitable topics for use in Studies 4b and 4c based on three criteria: (1) moralization, (2) weak partisan signaling, and (3) believability of the false feedback manipulation.

Partisan signaling ambiguity. Participants evaluated how strongly political attitudes on each topic signaled partisan identity by reporting to what extent they would think a person who holds the same view as the participant (aligned view) and the opposite view as the participant (opposite view) is either an ingroup member or an outgroup member (i.e., Democrat or a Republican) (-3 = ingroup member; 0 = no judgment; 3 = outgroup member). Among the five topics, participants had the weakest partisan priors for gambling (aligned view: $M = 0.57$, $SD = 1.59$; opposing view: $M = -0.34$, $SD = 1.67$) and surveillance (aligned view: $M = 0.59$, $SD = 1.62$;

opposing view: M = -0.57, SD = 1.77). Across both issues, participants tended to project their own views onto their party—reporting that someone who agreed with them was more likely to be a co-partisan, and someone who disagreed was more likely to be an out-partisan. This pattern suggests that attitudes on these issues were not anchored to party stereotypes; instead, participants appeared to infer partisanship based on agreement, indicating ambiguity in partisan signaling. These findings suggested that these topics would be suitable candidates for manipulating perceived polarization without confounding effects of entrenched partisan identity cues.

Moralization. To assess whether these issues were sufficiently moralized, we conducted one-sample *t*-tests on the absolute-value of participant moralization scores. Participants rated both gambling (M = 0.87, SD = 0.95) and surveillance (M = 1.05, SD = 1.03) as significantly moralized, with values well above zero, gambling: $t(216) = 13.53$, $p < .001$; surveillance: $t(216) = 14.99$, $p < .001$. These findings confirmed that both issues carried moral weight for participants, satisfying the second selection criterion.

Believability. To evaluate whether participants believed the fabricated public opinion data presented at the end of the survey (i.e., that a given issue was either highly polarized or not polarized), we analyzed surprise ratings as a proxy for plausibility. Surprise was rated on a 7-point scale (1 = *Not at all surprising*, 7 = *Very surprising*). Participants reported surprise levels below the midpoint for both issues, indicating that the information was seen as plausible and not especially unexpected: surveillance, M = 3.41, SD = 1.91, $t(215) = –4.52$, $p < .001$; gambling, M = 3.20, SD = 1.79, $t(216) = –6.61$, $p < .001$. Moreover, surprise ratings did not differ between participants in the polarized vs. non-polarized framing conditions, surveillance: $t(215) = 0.06$, p

= .95; gambling: $t(215) = 0.06$, $p = .95$. These results suggest that the manipulation was equally believable across conditions.

Finally, although partisan distribution was not a primary selection criterion, we report here the breakdown of position statements by party for descriptive purposes. For digitized gambling, 46% of Republicans and 40% of Democrats favored banning it except in designated areas, whereas 54% of Republicans and 60% of Democrats supported permitting it nationwide. For AI surveillance, 60% of Republicans and 40% of Democrats favored banning the use of AI for surveillance, while 40% of Republicans and 60% of Democrats supported permitting it. These near-split distributions further support the ambiguity of partisan opinion on these issues (see Table below).

Table. *Distribution of Attitudes Toward Gambling and AI Surveillance by Political Party.*

Issue	Party	Ban (%)	Permit (%)	N
Gambling	Republican	45.9	54.1	111
	Democrat	40.6	59.4	106
AI Surveillance	Republican	60.4	39.6	111
	Democrat	40.6	59.4	106

Conclusion

Together, these findings supported the selection of gambling legalization and AI surveillance as focal topics in Studies 4b and 4c. Both were perceived as moralized issues with ambiguous partisan signals, and the false feedback manipulation used to vary perceived polarization was judged as believable to a similar extent across polarization conditions.

Study 4A

Study 4a tested two core hypotheses. First, we tested whether the perception gap is larger for partisan (vs. non-partisan) topics—helping to distinguish whether it reflects loyalty-specific reputational concern or a more general tendency to expect more negative social evaluations from others than is warranted (e.g., Epley et al., 2022; Kardas et al., 2024; Savitsky et al., 2001). To do so, we manipulated whether participants evaluated dissenting belief change on a partisan or non-partisan issue, holding constant the direction, magnitude, and political nature of the belief change.

Second, we tested whether the perception gap is explained by miscalibrated meta-perceptions of ingroup loyalty. Specifically, we predicted that predictors would expect to be seen as more disloyal than reactors actually perceived—particularly when the belief change occurred on a partisan issue—thereby demonstrating a type of group-specific signal amplification bias (Vorauer et al., 2003). To test this account, we conducted a moderated mediation analysis, hypothesizing that the effect of role (predict vs. react) on anticipated social sanctions would be explained by an indirect effect via loyalty meta-perceptions, and that this indirect effect would be stronger for partisan (vs. non-partisan) topics[5].

Method

Participants

We recruited 400 participants to yield approximately 100 participants per cell after exclusions for failed attention and comprehension checks. The required sample size was

[5] We also tested and found evidence that predictors in the partisan (vs. non-partisan) condition reported greater self-censorship, and this effect was sequentially mediated by loyalty meta-perceptions and anticipated social sanctions. These results are reported in the OSM.

determined using a priori power analysis conducted in G*Power 3.1 (Faul et al., 2009), targeting the 2 × 2 interaction between role (predictor vs. reactor) and topic condition (partisan vs. non-partisan) on the social-sanctions composite. We specified a small-to-moderate interaction effect size of $f = 0.17$ (partial $\eta^2 \approx .03$), based on pilot testing, theoretical expectations, and conventional benchmarks for interaction effects in social psychological research (Cohen, 1988). This analysis indicated that a sample size of 92 participants per cell ($N = 368$) would provide 90% power to detect the hypothesized interaction.

We note that although our predicted pattern reflected a non-crossover ("attenuation") interaction, G*Power does not account for interaction *shape* when estimating power. As such, this approach may underestimate the sample size required to detect this form of interaction (Sommet et al., 2023). After applying all pre-registered exclusion criteria, the final sample consisted of $N = 393$ participants (240 Democrats, 153 Republicans; 253 females, 134 males, 6 self-identifying participants).

Table. Participant demographics in Studies 4a-c.

Study	N	Democrats n (%)	Republicans n (%)	Female n (%)	Male n (%)	Self-identifying n (%)	Age M (SD)
4a	393	240 (61.1)	153 (38.9)	253 (64.4)	134 (34.1)	6 (1.5)	—
4b	282	175 (62)	107 (38)	—	—	—	—
4c	596	324 (54.4)	272 (45.6)	343 (57.6)	247 (41.4)	6 (1.0)	41.46 (13.34)

Procedure

Study 4a used a 2 × 2 between-participants design wherein participants were randomly assigned to a role (predict vs. react) and a topic (partisan vs. non-partisan) condition. To begin the study, participants responded to three binary-choice questions about their attitudes on either partisan or non-partisan political topics.

Following broad definitions of "political topics" as matters relevant to governance, institutional decision-making, or the exercise of public authority (Lasswell, 1936), we categorize all issues used in the study as political. Within this framework, we distinguish partisan (versus non-partisan) issues as those marked by clear divisions between Democrats' and Republicans' attitudes (Stokes, 1963).

The partisan political topics were consistent with those used in Study 1: abortion, gun control, and immigration. The non-partisan political topics, by contrast, included government- and policy-adjacent questions unlikely to evoke strong partisan divisions: USPS postal delivery service policy ("The USPS should deliver mail on FIVE [SIX] days a week"), National Weather Service alert guidelines ("National Weather Service alerts should [should NOT] include emojis"), and the display of presidential portraits in the White House ("Presidential portraits should be displayed in chronological [reverse-chronological] order in the White House"). Although these topics fall within the broader political domain, they lack strong partisan division and thus served as a non-partisan political comparison set.

Participants in the partisan topic condition were randomly assigned to consider one of the three partisan issues (abortion, gun control, or immigration). Consistent with the procedure used in Study 1, participants in this condition could only be assigned to a topic if they reported a view that was consistent with the typical view of their political party on the issue. Given that the non-partisan issues were intentionally non-partisan, participants in this condition were randomly assigned to consider one of these three topics as the focal topic in the study.

As in Study 1, predictors were asked to imagine changing their view to the opposite position on the assigned topic, and to predict how an ingroup member on Prolific would react. In both topic conditions, we told predictors that their ingroup partner held the same original view as

they did—ensuring that the imagined belief change would always represent dissent from the partner's perspective. This design controlled for perceived disagreement across conditions. In the partisan condition, disagreement was likely assumed based on the issue itself. In the non-partisan condition, such assumptions were less likely to arise naturally, so we made the partner disagreement explicit. This allowed us to isolate the effects of topic partisanship while holding dissent constant.

Reactors read about another ingroup member on Prolific who had changed their mind on the randomly assigned topic, adopting the opposing point of view. To ensure that perceived disagreement was held constant across conditions, we told all reactors that this person initially held the same view as the reactor on the issue—before changing their mind to the opposing view.

Measures

After completing the experimental procedure, participants responded to the dependent measures. We employed the five-item scale measure of social sanctions used in Study 1 as the focal dependent measure in this study. We also included two bi-polar measures of social evaluation: (1) "How do you think another [ingroup member] would feel towards you if they learned about your belief change on this topic" (-3 = *Extremely negatively*; 0 = *Neutral*; 3 = *Extremely positively*); (2) "How much do you think another [ingroup member] would like you if they learned about your belief change on this topic?" (-3 = *Strongly dislike*; 0 = *Neutral*; 3 = *Strongly like*[6]). These items were included to measure—in absolute terms—how much dissenters expect to be and are liked following dissenting belief change. These items showed high internal

[6] These items have been recoded to facilitate interpretation. They were presented to participants on a 1-7 Likert-type scale with the same endpoint labels.

reliability *(r* = .86) and were combined into a composite measure for analyses. Reactors reported their actual reactions and evaluations on the same dimensions using modified items.

Next, predictors were asked two questions about how much the belief change would influence their partner's assessment of their loyalty to the ingroup political party: "This person will think I am…(1) loyal to the [ingroup] Party, (2) likely to vote for [outgroup]s in future elections" (1 = *Definitely Not*; 7 = *Definitely Yes*). Reactors responded to the same two items with modified wording to capture their judgment of the target's ingroup loyalty. The second item was reverse-coded before the two items were combined as a single composite measure (r = .65).

To conclude the study, participants completed two attention checks—one about the focal topic, and one about the political affiliation of the person they read about—along with basic demographic questions[7]. Participants who failed these checks were excluded from analyses.

Results

To examine whether the perception gap was larger in the partisan topic condition compared to the non-partisan topic condition, we first conducted a 2(Role: predict vs. react) × 2(Topic: partisan vs. non-partisan) ANOVA using type III sum of squares to test for an interaction and main effects of role and topic conditions on participants' social sanction composite scores (α = .96). As predicted, this analysis revealed a significant interaction between role and topic conditions, $F(1,389) = 11.21$, $p < .001$, $\eta p^2 = .03$, indicating that the perception gap in social sanctions was significantly larger in the partisan condition. This supports the idea that reputational concerns—beyond general social pessimism—contribute to overestimation.

[7] Study 4a included several exploratory measures based on helpful suggestions from reviewers of this paper. These measures were ultimately not central to the theoretical account presented in this manuscript.

To clarify the nature of the role x topic condition interaction, we conducted pairwise comparisons. In the partisan condition, predictors anticipated significantly more social sanctions (M = 4.68, SD = 1.45) than reactors reported (M = 2.87, SD = 1.74), $t(389) = 8.64$, $p < .001$, d = 1.13. In the non-partisan condition, predictors again anticipated more rejection (M = 2.48, SD = 1.45) than reactors reported (M = 1.67, SD = 1.17), $t(389) = 3.84$, $p < .001$, $d = 0.61$. Although the overestimation effect was present in both topic conditions, it was substantially larger for partisan topics. This pattern suggests that the larger perception gap in partisan topics is unlikely to reflect a general tendency to expect negative social evaluations and may be explained by psychological mechanisms specific to politically polarized contexts.

To test whether the larger perception gap observed in the partisan condition was explained by meta-perceptual loyalty concerns, we conducted a moderated-mediation analysis using structural equation modeling with 5,000 bootstrap samples (N = 393) in R (lavaan package). In this model, role (predictor = 1, reactor = 0) was entered as the independent variable, topic condition (partisan = 1, non-partisan = 0) served as the moderator, loyalty was the mediator, and social sanctions were the outcome. In keeping with the study design, loyalty reflected meta-perceptions of how loyal participants expected to be seen (for predictors) or actual judgments of target loyalty (for reactors), and the outcome variable captured anticipated or reported social sanctions, respectively. This model allowed us to test whether the effect of role on social sanctions was mediated by loyalty and whether this indirect effect was stronger for partisan (versus non-partisan) topics.

We first examined whether the effect of role on loyalty (i.e., the proposed mediator) was moderated by topic condition—a necessary step to establish moderated mediation. We observed a significant role × condition interaction on loyalty, $b = –0.58$, $SE = 0.27$, $z = –2.11$, $p = .035$,

such that predictors in the partisan topic condition expected to be seen as significantly less loyal (M = 3.08, SD = 1.49) than reactors actually perceived them to be (M = 3.76, SD = 1.43)—an overestimation that was attenuated in the non-partisan topic condition where loyalty ratings were closely aligned (M = 4.98 vs. M = 5.08).

This pattern is consistent with a group-specific signal amplification bias in partisan contexts: Predictors in the partisan topic condition expected to be seen as significantly less loyal than reactors actually judged them—an overestimation that did not emerge for non-partisan topics. Consistent with this interpretation, lower loyalty scores were strongly associated with higher anticipated rejection, b = –0.55, SE = 0.05, z = –10.28, p < .001, supporting the hypothesized link between perceived disloyalty and social sanctions.

A moderated mediation model confirmed that this indirect effect was significantly stronger in the partisan topic condition, index of moderated mediation = 0.32, SE = 0.15, z = 2.11, p = .035. Specifically, the indirect effect of role on rejection via loyalty was significant for partisan topics (b = 0.37, SE = 0.12, z = 3.22, p = .002), but not for non-partisan topics (b = 0.06, SE = 0.10, z = 0.58, p = .563).

Together, these findings suggest that dissenting belief change on partisan issues—where attitudes serve as loyalty signals—is perceived as more reputationally risky. This overestimation of social sanctions appears to be explained by predictors' belief that their dissent will signal greater disloyalty than it actually does.

We next examined whether the core results replicated using a bipolar composite measure of social evaluations. This measure was included for two key reasons. First, it served as a robustness check to ensure that these findings were not an artifact of using unipolar rejection

measures. Second, it provided additional interpretive value by capturing both relative (i.e., between-condition) and absolute (i.e., compared to a neutral midpoint) social evaluations. The bipolar scale allowed us to assess whether dissent was expected and received as globally negative, positive, or neutral.

The moderated mediation analysis using this outcome yielded the same pattern of results reported above (full model output reported in the OSM). To further contextualize these effects, Figure 4 presents predictors' expected evaluations and reactors' actual evaluations, centered on zero (i.e., neutral evaluation). In the partisan condition, predictors expected to be evaluated negatively (M = -1.05, SD = 1.13), whereas reactors evaluated dissenters only slightly below neutral (M = -0.25, SD = 1.21), $t(389) = 5.00$, $p < .001$, $d = .71$. In the non-partisan condition, predictors expected (M = 0.17, SD = 1.12) and reactors reported (M = 0.48, SD = 1.00) relatively positive evaluations, $t(389) = 1.94$, $p = .209$, $d = .28$. These findings reinforce the presence of a perception gap: Although dissent on partisan topics does carry interpersonal costs, predictors substantially exaggerate the extent of these costs. The full ANOVA and one-sample t-test results are reported later in the OSM under "Study 4A: Additional Analyses".

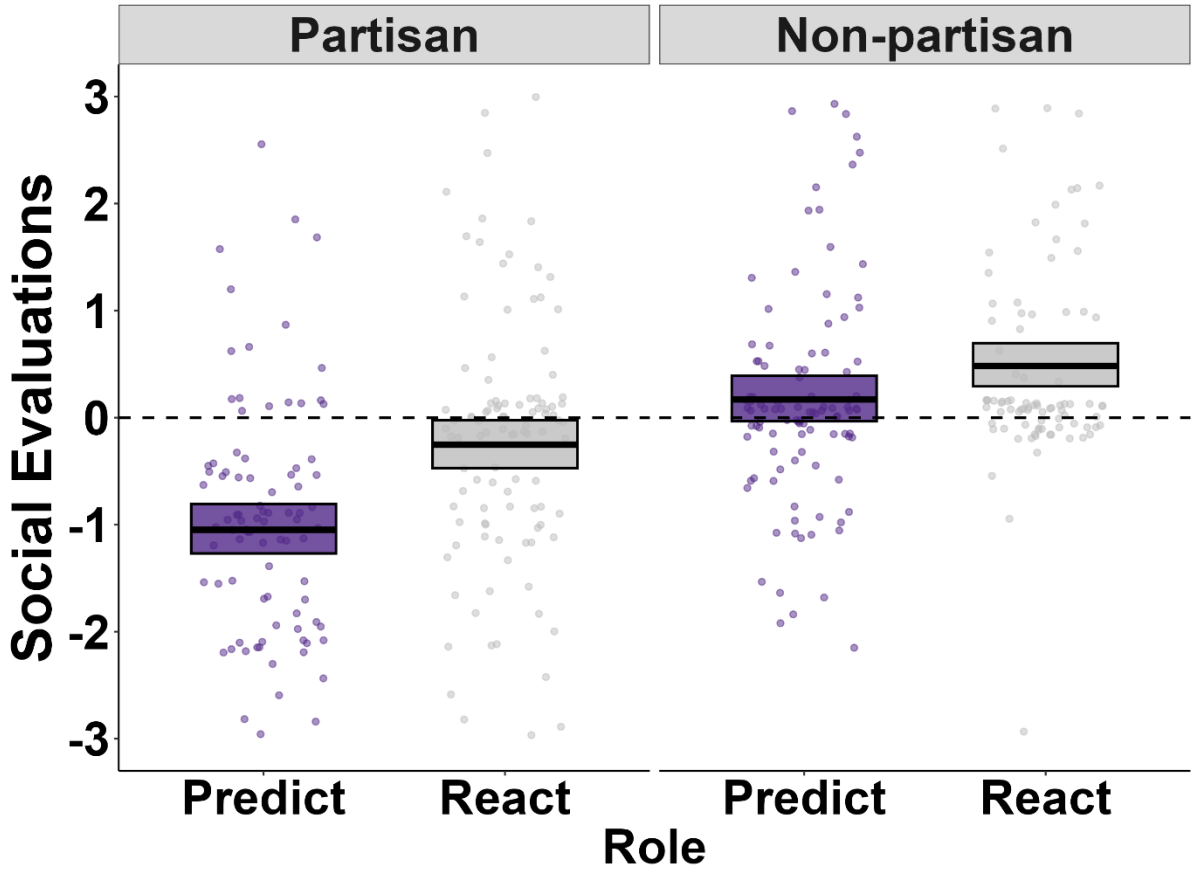


Figure 4. Comparing predictor estimates of social evaluations against reactors' actual evaluations within each topic condition. *Note.* The bipolar composite has been rescaled on the y-axis to center the midpoint on zero, which represents a neutral evaluation. Positive scores represent a positive social evaluation. Negative scores represent a negative social evaluation. The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval. Bipolar social evaluations were measured on seven-point response scales (e.g., -3 = *Strongly dislike*; 0 = *Neutral*; 3 = *Strongly like*).

Study 4B

Study 4a established a boundary condition and tested a core mechanism underlying the perception gap. By showing that the gap was significantly larger for partisan (versus non-partisan) belief change, Study 4a provided evidence that overestimation of social sanctions for dissenting belief change is not merely a reflection of more global bias leading to negative social evaluations, but instead is better explained by a group-specific reputational concern when such dissent might signal disloyalty to the group. It further demonstrated that this gap was explained, in part, by predictors' exaggerated fears of being perceived as disloyal—a pattern consistent with the proposed signal amplification bias account.

One limitation of Study 4a is that the partisan and non-partisan topics differed along more dimensions than partisanship. For example, beliefs about the USPS may not only be less partisan than beliefs about immigration, but also less perceived political relevance or controversy—factors that could themselves affect loyalty judgments and rejection. To address this, Study 4b held the topic content constant across all conditions and experimentally manipulated whether each issue was framed as partisan or non-partisan. This approach allowed us to isolate the effect of perceived partisanship on loyalty and rejection.

Methods

Participants. We recruited 328 U.S.-based participants identifying as either Democrat or Republican from Prolific Academic. After excluding participants who failed comprehension and attention checks, the final sample consisted of $N = 282$ (38% Republicans; 62% Democrats).

We conducted an a priori power analysis to determine the sample size needed to detect the 2 × 2 interaction between role (predictor vs. reactor) and topic framing condition (partisan vs.

non-partisan) on the social-sanctions composite. Using G*Power 3.1 (Faul et al., 2009), we set α = .05, power = .90, and an effect size of $f = 0.17$ (partial $\eta^2 \approx .03$), which was the interaction effect observed in Study 4a. This analysis indicated that 342 participants would be required to achieve 90% power. Although our final sample after exclusions fell slightly below this threshold, a post hoc sensitivity analysis using the observed effect size in Study 4b ($\eta p^2 = .029$) indicated achieved power of 83.7%, suggesting adequate sensitivity to detect the targeted interaction.

Procedure. Participants were randomly assigned to a 2 (Role: predict vs. react) × 2 (Topic Framing: partisan vs. non-partisan) between-subjects design. The study was presented to participants as part of "a large-scale, nationally representative public opinion polling study that seeks to assess American's attitudes towards various issues." Participants began by reporting their views on two issues: digitized gambling and the use of AI in government surveillance. For each issue, they selected the statement (from two opposing options) that best reflected their attitude. The statements were: "The U.S. government should be (a) **allowed** to use AI to support surveillance programs, or (b) **banned** from using AI to support surveillance programs" and, "Digitized gambling should be (a) **banned**, except for in designated areas, or (b) **permitted** nationwide".

These issues were selected based on a pretest (see Pretest Studies 4a & 4b in the OSM) designed to identify issues that (a) were perceived as moralized and (b) did not strongly signal partisan identity. The pretest confirmed that both issues met these criteria and were suitable for the framing intervention: the partisan framing manipulation was perceived as plausible and not especially surprising for either topic.

After reporting their views, participants learned they would see how Democrat and Republican participants from the sample responded to one of the two issues, which was

randomly selected as the focal topic. They again saw the two position statements and were told that, based on the data, the issue was either “deeply split along party lines” (partisan condition) or “NOT split along party lines” (non-partisan condition). This framing manipulation allowed us to vary the perceived partisanship of the issue while holding the topic content constant—enabling a clean test of whether perceived partisanship amplifies the perception gap.

In both conditions, Position A was matched to each participant’s own view on the topic. That is, Position A always reflected the participant’s selected stance, and Position B reflected the opposing stance. Below is an example of the intervention text shown to a Democrat participant assigned to the digitized gambling condition.

> **Partisan Condition:** When we look at Democrats' and Republicans' views on this issue, we see that their opinions are **deeply split along party lines**. We find that the vast majority of **Democrats strongly support Position A:** "Digitized gambling should be [**banned**, except for in designated areas / **permitted** nationwide]. In contrast, the vast majority of **Republicans strongly support Position B**: “Digitized gambling should be [**permitted** nationwide / **banned**, except for in designated areas]”. This suggests that this issue will be **highly polarized** between Democrats and Republicans in future elections.
>
> **Non-Partisan Condition:** When we look at Democrats' and Republicans' views on this issue, we see that their opinions are **NOT split along party lines**. We find that both Democrats and Republicans are equally likely to support either Position A or Position B on this issue. This suggests that this issue will **NOT be polarized** between Democrats and Republicans in future elections.

Participants then completed two comprehension check items asking them to indicate which position both ingroup and outgroup partisans typically supported on the assigned issue (1 = *Strongly support position A*; 4 = *Equally supportive of both positions*; 7 = *Strongly support position B*). Participants in the partisan condition failed the attention check if they selected any response other than "Strongly support position A" for the ingroup and "Strongly support position B" for the outgroup. Participants in the non-partisan condition failed the comprehension check if they selected any response other than "Equally supportive of both positions" for both the ingroup and the outgroup.

Any participant who failed this comprehension check was presented with the experimental manipulation again and asked the comprehension questions a second time. If a participant failed to provide correct answers to both questions the second time, they were removed from the study and compensated $0.53 for their time. Ninety percent of participants passed the comprehension check on the first attempt. Among the sixty-five participants who failed initially, twenty-nine (45%) failed the second attempt and were excluded from the study.

Next, participants completed survey measures based on their assigned role. Reactors read about another ingroup member on Prolific who had initially held the same view on the assigned topic (Position A) but changed their mind to adopt the opposite position (Position B). Reactors were asked to evaluate this person, "knowing that the issue is [highly polarized vs. not very polarized] between Democrats and Republicans." Reactors then rated how loyal the target is to the political ingroup using the same two-item measure of group loyalty as in Study 4a (r = .79); however, the second item was re-worded to measure support for the ingroup (rather than support for the outgroup, as in Study 4a). This modification may explain the relatively stronger

correlation between the two items observed in this study. Next, reactors reported social sanctions using the five-item measure from Study 1 ($\alpha = .98$).

Predictors were told that another participant from their political ingroup would learn that they had changed their view from Position A (their own view) to Position B (the opposing view). Predictors were asked to estimate how this ingroup member would evaluate them in terms of (a) loyalty to the political ingroup and (b) social sanctions, "knowing that the issue is [highly polarized vs. not very polarized] between Democrats and Republicans." They completed modified versions of the same measures as reactors.

In the final section of the study, participants responded to several check measures: (a) a manipulation check item that asked to what extent members of their political ingroup support Position A (1 = *Strongly oppose*; 4 = *Neither strongly oppose nor strongly support*; 7 = *Strongly support*), (b) an attention check item that asked which political party the person they read about in the study supported, (c) a measure of their partisan identity strength, and (d) a believability check rating how credible they found the partisan framing manipulation (1 = *Not at all*; 4 = *Moderately*; 7 = *Very much*). Participants were then debriefed and compensated for their participation.

Results

Manipulation Check. To assess whether the partisan framing manipulation was effective, we conducted a two-way ANOVA to compare manipulation check scores by role and topic conditions. This analysis showed that the manipulation had its intended effect, revealing a main effect of the topic condition, $F(1, 278) = 375.19$, $p < .001$, $\eta p^2 = .57$, demonstrating that participants in the partisan topic condition believed their ingroup to strongly support position A

(M = 6.78, SD = 0.98) significantly more than participants in the non-partisan topic condition (M = 4.09, SD = 0.56). There was no main effect of role, p = .969, nor a significant interaction, p = .530, demonstrating that the partisan framing manipulation had a consistent effect for both predictors and reactors.

Believability Check. To test whether participants in the sample believed the partisan framing manipulation, we analyzed the descriptive statistics for the believability measure. In general, participants found the partisan framing information believable (M = 5.00, SD = 1.70), with over 83% of participants reporting a score of four or higher on the seven-point scale[8]. Next, we conducted a two-way ANOVA to test whether the believability of the manipulation varied between any of the conditions, which revealed that believability ratings did not vary significantly by role condition, $F(1, 278) = 0.004$, $p = .951$, $\eta p^2 < .001$, or by topic condition $F(1, 278) = 0.208$, $p = .648$, $\eta p^2 < .001$. Moreover, the interaction between role and topic conditions was not significant, $F(1, 278) = 0.10$, $p = .752$, $\eta p^2 < .001$. These findings demonstrate that participants generally found the partisan framing manipulation believable, and they did so to a similar extent across conditions.

Social Sanctions. We first tested whether the perception gap for social sanctions was significantly larger for partisan (vs. non-partisan) topics. A two-way ANOVA using type III sum of squares revealed a significant interaction between role and topic condition, $F(1, 278) = 8.34$, $p = .004$, $\eta p^2 = .029$, replicating the core pattern from Study 4a by showing that the perception gap was significantly larger in the partisan topic condition (vs. the non-partisan condition).

[8] As a robustness check, we replicated all analyses for Study 4b and Study 4c, respectively, with only participants who reported a 4 or higher on the believability measure. For both studies, these analyses revealed identical conclusions for all hypothesis tests reported in the main text. These analyses are reported in the OSM.

To clarify the nature of the interaction, we conducted pairwise comparisons of predictors and reactors within each topic framing condition. In the polarized condition, predictors anticipated significantly more social sanctions ($M = 4.54$, $SD = 1.49$) than reactors reported ($M = 2.56$, $SD = 1.64$), $t(278) = 7.53$, $p < .001$, $d = 1.27$. In the non-polarized condition, predictors again anticipated more rejection ($M = 3.08$, $SD = 1.47$) than reactors reported ($M = 2.18$, $SD = 1.62$), $t(278) = 3.36$, $p = .001$, $d = 0.58$. Although the overestimation effect was present across both framing conditions, it was substantially larger when belief change occurred in the context of politically polarized topics. This pattern is consistent with the hypothesis that dissent on politicized topics is seen as especially reputationally risky, which may be associated with a group-specific signal amplification bias.

Moderated Mediation. We next examined whether this interaction effect could be accounted for by the proposed mechanism of a loyalty-focused signal amplification bias. Using the same model structure as in Study 4a, we conducted a moderated mediation analysis in which role (predictor = 1, reactor = 0) was the independent variable, topic condition (partisan = 1, non-partisan = 0) was the moderator, loyalty was the mediator, and social sanctions served as the dependent variable.

We first tested whether role × condition predicted loyalty judgments. This interaction was significant, $b = –1.26$, $SE = 0.34$, $z = –3.75$, $p < .001$, such that predictors in the partisan condition expected to be seen as significantly less loyal ($M = 2.91$, $SD = 1.37$) than reactors actually rated them ($M = 4.27$, $SD = 1.38$)—an overestimation that was attenuated in the non-partisan condition ($M = 4.62$ vs. $M = 4.72$). Lower loyalty scores significantly predicted higher social sanctions, $b = –0.52$, $SE = 0.07$, $z = –7.76$, $p < .001$, consistent with the proposed mediation pathway.

We then tested whether this indirect pathway was moderated by condition. The index of moderated mediation was significant, $b = 0.65$, $SE = 0.19$, $z = 3.44$, $p = .001$, indicating that the indirect effect of role on rejection via loyalty was significantly stronger in the partisan condition, $b = 0.70$, $SE = 0.15$, $z = 4.83$, $p < .001$, than in the non-partisan condition, $b = 0.05$, $SE = 0.13$, $z = 0.40$, $p = .687$.

Together, these results replicate and extend the findings from Study 4a, providing further evidence that overestimation of rejection in partisan contexts is explained, in part, by a type of group-specific signal amplification bias. That is, predictors appear especially sensitive to the reputational consequences of belief change on partisan issues, potentially because such issues are perceived to carry stronger signals of group disloyalty. This heightened vigilance to social threat is consistent with a self-protective "better safe than sorry" strategy: To avoid costly exclusion, group members are overly sensitive to how their dissent will be interpreted, which may contribute to their tendency to overestimate social rejection.

Study 4C

Study 4c was designed as a preregistered, higher-powered replication of Study 4b, using the same design, measures, and analytic strategy. Although the core hypotheses and procedures remained unchanged, we note a few minor procedural differences below.

Methods

Participants. This preregistered replication study used a more conservative effect size estimate than Study 4b for the targeted 2 × 2 interaction between role (predictor vs. reactor) and topic framing condition (partisan vs. non-partisan) on the social-sanctions composite. Specifically, we powered the study to detect $f = 0.13$ (partial $\eta^2 \approx .017$), based on the lower bound of the 95% confidence interval around the interaction effect observed in Study 4b.

Using G*Power 3.1 (Faul et al., 2009), we set $\alpha = .05$ and power = .90, which indicated that a sample of 610 participants would be required. To allow for exclusions, we recruited 670 participants. After excluding those who failed comprehension or attention checks, the final sample consisted of $N = 596$ participants (272 Republicans, 324 Democrats; 247 males, 343 females, 6 self-identifying participants; $M_{age} = 41.46$, $SD_{age} = 13.34$).

Procedure. Study 4c used the same design, measures, and procedure as Study 4b. The minor differences between the two studies were as follows. First, in Study 4c, participants reported their attitudes on only one randomly assigned question (rather than answering both and then being randomly assigned to one after the fact, as in Study 4b). Second, the manipulation check in Study 4c was modified to, “How DIFFERENT is this new belief from the view that most [ingroup members] have on this issue?” (1 = *Not very different;* 4 = *Somewhat;* 7 = *Very different*).

Results

Manipulation Check. To examine whether the experimental manipulation of partisan framing had its intended effect, we conducted a two-way ANOVA to compare manipulation check scores by role and topic conditions. This analysis showed that the manipulation had its intended effect, revealing a main effect of the topic condition, $F(1, 592) = 340.33$, $p < .001$, $\eta p^2 = .37$, demonstrating that participants in the partisan topic condition believed that the dissenting belief represented a view that was different from the majority view of their ingroup ($M = 6.43$, $SD = 1.04$) to a significantly greater extent than participants in the non-partisan condition ($M = 3.37$, $SD = 1.60$). The main effect of role ($p = .34$) and the interaction of role × topic condition ($p = .174$) were both not significant, demonstrating that the partisan framing manipulation had a consistent effect for both predictors and reactors.

Believability Check. To test whether participants in the sample found the partisan framing manipulation credible, we analyzed the results for the believability measure. In general, participants found the partisan framing information believable ($M = 4.77$, $SD = 1.71$), with 80% of participants reporting a score of four or higher. Next, we conducted a two-way ANOVA to test whether the believability of the manipulation varied between any of the conditions, which revealed that believability ratings did not vary by role condition, $F(1, 592) = 2.76$, $p = .097$, $\eta p^2 = .005$; however, the main effect of the topic condition was statistically significant, $F(1, 592) = 14.03$, $p < .001$, $\eta p^2 = .023$. Moreover, the interaction between role and topic conditions was significant, $F(1, 592) = 4.30$, $p = .039$, $\eta p^2 = .007$. This pattern diverges from Study 4b, where believability ratings did not vary by condition—an unexpected finding we address in the discussion section.

Tukey's HSD post-hoc tests were conducted to account for multiple comparisons and the lack of a priori hypothesis about the direction of significant differences in believability between conditions. These tests revealed that predictors in the non-partisan topic condition found the framing manipulation significantly less believable ($M = 4.37$, $SD = 1.74$) than predictors in the partisan topic condition ($M = 5.11$, $SD = 1.61$), $t(592) = 3.75$, $p = .001$, $d = .44$. Among reactors, believability did not significantly differ between the non-partisan topic condition ($M = 4.70$, $SD = 1.66$) and the partisan topic condition ($M = 4.86$, $SD = 1.75$), $t(592) = 0.85$, $p = .830$, $d = .01$.

Social Sanctions. Study 4c replicated the central findings from Studies 4a and 4b, again revealing a significant perception gap between predictors and reactors that was amplified in polarized contexts. A two-way ANOVA using Type III sum of squares revealed a significant interaction between role and topic condition, $F(1, 592) = 7.14$, $p = .008$, $\eta p^2 = .012$. This interaction mirrored the pattern observed in Studies 4a–b, such that predictors in the polarized condition anticipated significantly greater rejection than reactors reported, whereas the gap was attenuated in the non-polarized condition.

To clarify the nature of the interaction, we conducted post-hoc pairwise comparisons of predictors and reactors within each topic framing condition. In the polarized condition, predictors anticipated significantly more social sanctions ($M = 4.42$, $SD = 1.58$) than reactors reported ($M = 2.74$, $SD = 1.64$), $t(592) = 9.59$, $p < .001$, $d = 1.05$. In the non-polarized condition, predictors also anticipated more rejection ($M = 3.13$, $SD = 1.48$) than reactors reported ($M = 2.10$, $SD = 1.37$), $t(592) = 5.81$, $p < .001$, $d = 0.72$. Although the overestimation effect was present across both framing conditions, it was substantially larger in the polarized condition, providing further support of the hypothesis that dissent on politicized topics is perceived as more reputationally risky—consistent with a group-specific signal amplification bias.

Moderated Mediation. To test whether the significant interaction between the topic condition (partisan vs. non-partisan) and role (predictor vs. reactor) on anticipated rejection was mediated by perceived disloyalty, we conducted a moderated mediation analysis using the same model structure and analytic approach as in Studies 4a and 4b. We first examined whether the effect of role on anticipated rejection was moderated by the topic condition. Although both role ($b = –0.94$, $SE = 0.16$, $z = –6.06$, $p < .001$) and topic condition ($b = –1.33$, $SE = 0.15$, $z = –8.85$, $p < .001$) significantly predicted loyalty perceptions, their interaction was not significant ($b = 0.23$, $SE = 0.23$, $z = 1.00$, $p = .317$). indicating that the conditions for moderated mediation were not met.

Nevertheless, consistent with the prior findings, loyalty remained a significant mediator of the relationship between role and anticipated rejection across both conditions. Lower loyalty scores significantly predicted greater anticipated rejection, $b = –0.47$, $SE = 0.05$, $z = –10.11$, $p < .001$. The indirect effect of role on rejection via loyalty was significant in both the partisan condition, $b = 0.33$, $SE = 0.09$, $z = 3.80$, $p < .001$, and the non-partisan condition, $b = 0.44$, $SE = 0.08$, $z = 5.21$, $p < .001$. Although the interaction effect from Studies 4a and 4b was not replicated, these findings provide further support for the central role of loyalty concerns in shaping predictors' expectations of social sanctions following belief change.

Study 1: Additional Analyses

Study 1: Comparing Democrats and Republicans in the sample.

Our sample (N = 500) consisted of 248 participants who identified as a Democrat and 252 participants who identified as a Republican. To explore whether the main effect of role varied between Democrats and Republicans in our sample, we conducted a 2 (role: predict vs. react) x 2 (party ID: Republican vs. Democrat) ANOVA. The results revealed a significant main effect of role, $F(1, 496) = 93.93$, $p < .001$, $\eta^2 = .16$, indicating that participants in the predict role expected significantly more rejection than those in the react role. There was also a significant main effect of party identification, $F(1, 496) = 6.27$, $p = .013$, $\eta^2 = .01$, suggesting that overall, rejection expectations differed between Republicans and Democrats. The interaction between role and party identification was not significant, $F(1, 496) = 0.45$, $p = .505$, $\eta^2 = .00$, indicating that the effect of role on rejection—in other words, the perception gap—did not differ significantly between Republicans and Democrats.

Given the significant main effect of party identification, we conducted follow-up independent samples t-tests to further explore whether predicted or actual rejection differed between Democrats and Republicans within each role condition. In the predict condition, the t-test revealed a significant difference in rejection expectations between Republicans (M = 3.73; SD = 1.56) and Democrats (M = 4.18; SD = 1.57), $t(247) = -2.25$, $p = .025$, $d = -0.29$. These results suggest that Democrats in the predict role expected marginally significantly more rejection than Republicans, but the effect size is small.

In the react condition, the t-test showed no significant difference in rejection expectations between Republicans (M = 2.46; SD = 1.57) and Democrats (M = 2.72; SD = 1.61), $t(249) = -$

1.29, $p = .197$, $d = -0.16$, suggesting that the difference in actual rejection between Republicans and Democrats was small and not statistically significant.

Study 1: Exploring the effect of role for each topic.

Participants were randomly assigned to consider any of the following topics upon which they reported they held the typical view of their party: abortion ($n = 177$), immigration ($n = 160$), or gun control ($n = 163$). We refer to the assigned topic as the "topic condition". To explore whether any of our effects varied across topic conditions, we first ran a 2 (role: predict vs. react) x 3 (topic: abortion, immigration, gun control) ANOVA which revealed a significant main effect of role, $F(1, 494) = 93.12$, $p < .001$, $\eta^2 = .16$; however, the main effect of issue was not significant, $F(2, 494) = 2.04$, $p = .131$, $\eta^2 = .01$, nor was the interaction between role and issue, $F(2, 494) = 0.15$, $p = .864$, $\eta^2 = .00$, suggesting that the effect of role on rejection did not significantly differ by topic condition.

Next, we ran a series of independent samples t-tests to compare predict versus react condition scores within the subset of participants who were assigned to each topic. These tests revealed that the effect of role condition was significant for each topic. For the abortion topic, there was a significant difference in rejection expectations between participants in the predict ($M = 4.17$, $SD = 1.54$) and react ($M = 2.75$, $SD = 1.61$) roles, $t(175) = 5.99$, $p < .001$, $d = .90$. For the immigration topic, participants in the predict role ($M = 3.83$, $SD = 1.56$) also expected significantly more rejection than those in the react role ($M = 2.59$, $SD = 1.66$), $t(158) = 4.89$, $p < .001$, $d = .77$. For the gun control topic, those in the predict role ($M = 3.82$, $SD = 1.64$) expected more rejection compared to participants in the react role ($M = 2.43$, $SD = 1.51$), $t(161) = 5.63$, $p < .001$, $d = 0.89$.

Study 1: Comparing participants who held at least one atypical view to participants who held only typical views.

As mentioned above, participants reported their attitudes on abortion, gun control, and immigration and were then randomly assigned to consider one topic upon which they reported the typical view of their party. A reasonable question is whether our effects varied between participants who reported the typical view on each of the three topics and participants who reported at least one atypical (i.e., "dissenting") view. On average, both Republican (78.44%) and Democrat (86.56%) participants reported the typical view on these topics the majority of the time. However, a greater proportion of participants reported at least one dissenting view on at least one of the topics (Republicans: 48.41%; Democrats: 34.27%). Participant responses, sorted by party ID, to each of the binary-choice attitude measures for all three topics are reported in Table S2 below.

Table S2: Percentages of Views by Party ID

Issue	Party ID	Liberal View (%)	Conservative View (%)
Abortion	Republican	29.76	70.24
	Democrat	97.98	2.02
Immigration	Republican	13.10	86.90
	Democrat	75.81	24.19
Gun Access	Republican	21.83	78.17
	Democrat	85.89	14.11

To investigate whether participants who reported typical views on all three topics (referred to as "orthodox" participants) responded differently to our rejection items compared to those who reported at least one atypical view (referred to as "unorthodox" participants), we conducted a 2x2 ANOVA with orthodoxy (orthodox vs. unorthodox) and role (predict vs. react) as factors. The results did not reveal a significant main effect of orthodoxy, $F(1, 496) = 0.45$, $p = .505$, $\eta^2 = .00$, indicating that there was no significant difference in rejection scores between participants who held orthodox views ($M = 3.31$, $SD = 1.68$) and those who held at least one unorthodox views ($M = 3.21$, $SD = 1.79$). The main effect of role was again significant, $F(1, 496) = 92.52$, $p < .001$, $\eta^2 = .16$; however, the interaction between orthodoxy and role was not significant, $F(1, 496) = 0.81$, $p = .367$, $\eta^2 = .002$, indicating that the effect of role on rejection did not significantly differ between orthodox and unorthodox participants.

"Acceptance" scale items. Study 1 also included a four-item Likert-type scale to measure *social acceptance* ($\alpha = .98$), which was developed for exploratory purposes to complement the social rejection measure. Whereas the rejection items captured expectations of exclusion and criticism, the acceptance items assessed perceived inclusion and recognition within the ingroup. The items in the predict condition were: "If you were to change your mind on the issue of [topic] to believe [the opposite], to what extent do you think that another [Democrat / Republican] would… (1) accept you as a member of the group? (2) treat you as a member of the group? (3) view you as a member of the group? (4) include you as a member of the group?" (1 = Not very much; 7 = Very much). The items in the react condition were identical in structure, but the wording was modified such that reactors reported how they would respond to another member of their political ingroup who changed their mind.

To examine whether predictors accurately estimated how much ingroup members would *accept* them for dissenting belief change, we conducted a 2 (Role: predict vs. react) × 3 (Topic: abortion vs. immigration vs. gun control) ANOVA on the acceptance composite. This analysis revealed a significant main effect of role, $F(1, 494) = 24.61$, $p < .001$, $\eta p^2 = .05$, such that predictors ($M = 3.25$, $SD = 1.49$) *underestimated* how much reactors would accept them ($M = 3.96$, $SD = 1.71$; $d = 0.44$). The main effect of topic was not significant, $F(2, 494) = 1.21$, $p = .300$, $\eta p^2 = .005$, nor was the interaction between role and topic, $F(2, 494) = 0.09$, $p = .918$, $\eta p^2 < .001$, indicating that the effect of role on acceptance did not differ significantly across topics.

These findings closely mirrored the pattern observed for the rejection items, demonstrating a consistent perception gap between predictors and reactors. Just as predictors overestimated rejection, they also underestimated acceptance—suggesting that partisans generally miscalibrate the social consequences of dissenting belief change in a pessimistic direction. However, the magnitude of the gap for acceptance ($d = 0.44$) was smaller than that for rejection ($d = 0.87$ in the rejection composite), indicating that people's expectations about inclusion may be somewhat less distorted than their expectations about exclusion. One possible explanation is that individuals are more attuned to cues of potential rejection than to cues of acceptance, reflecting a negativity bias in social perception (Baumeister et al., 2001; Rozin & Royzman, 2001). Because the costs of social exclusion are typically more salient and consequential than the benefits of inclusion, predictors may have erred more strongly on the side of anticipating rejection than underestimating acceptance.

Study 2: Additional Analyses

Study 2: Key Analysis (social sanctions, behavioral measures, qualitative measures) for the Entire Sample.

Due to a technical difficulty, a portion of participants who completed Study 2 were not successfully paired with another participant based on our criterion. Specifically, a large number of participants were either (a) paired with another participant who was assigned to the same role condition (e.g., two predictors in a dyad), or (b) not paired with anyone. These participants are not included in the analyses reported in the main text, which only includes participants who were paired with another participant who was (a) assigned to the opposite role condition, (b) assigned to the same topic condition, and (c) identified with the same political party as the participant. Here, we report auxiliary analyses including all participants who passed our study attention checks—even those for whom the randomization system failed. Overall, the pattern of results is consistent with what we observed among the subset of participants who met our criterion.

Social Sanctions. Predictors expected that their partner would socially reject them (M = 2.46; SD = 1.29) for dissenting belief change significantly more than reactors reported (M = 1.46, SD = 1.46), $t(560) = 10.71$, $p < .001$, $d = .89$.

Qualitative Free Response Data. Next, we analyzed predictor and reactor responses to the free response items, which asked them to describe how they think their partner felt (predictors) or how they felt (reactors) after learning about the dissenting belief change. We assembled a team of four hypothesis-blind research assistants who independently coded participant responses (ICC = .86) based on the expected and actual effects on interpersonal evaluations ranging from -3 (very negative) to +3 (very positive). These findings were again consistent with what we observed on survey measures: Participants in the predict condition

anticipated significantly worse interpersonal outcomes (M = -0.79, SD = 1.04) than participants in the react condition reported (M = 0.01, SD = 1.11), $t(436) = -7.73$, $p < .001$, $d = .74$.

Behavioral Measures. Lastly, we observed a similar pattern of results on the behavioral measures. On average, predictors overestimated how often their partner would choose to work with someone else on the cooperative task (Predict: 20.4%; React: 8.8%), $X^2(1, 561) = 14.16$, $p < .001$, $V = .16$, and expected their partner to keep significantly more of the bonus ($M_{predict}$ = 5.3 cents) than their reactor partners kept (M_{react} = 4.8 cents), $t(560) = -4.49$, $p < .001$, $d = .38$.

Study 2: Testing for Attrition Effects Between the pre-screen and the full study.

Almost all participants from the pre-screen (N = 817; 92%) met our recruitment criterion and were invited to participate in Study 2. In total, 613 participants who were invited to participate in Study 2 accepted the study HIT (208 participants did not). To examine whether there were significant differences between participants who completed Study 2 and those who did not, several attrition tests were conducted. First, an independent samples t-test was conducted to compare political identity centrality scores between participants who completed Study 2 and those who did not. The results indicated that there was no significant difference in political identity centrality between participants who did not complete Study 2 (M = 4.46; SD = 0.80) and those who did complete Study 2 (M = 4.60, SD = 0.75), $t(841) = -1.03$, $p = .302$.

Next, we conducted chi-square test of independence to see if the attrition rate was different among Democrats and Republicans in our sample. The analysis revealed a significant association between these variables, χ^2 (1, N = 843) = 8.96, p = .003. Examination of the contingency table showed that a higher proportion of Republicans (102 out of 337, 30.3%) failed to complete Study 2 compared to Democrats (106 out of 506, 20.9%). Therefore, in the next

section of the supplement we explore for potential differences between Democrats and Republicans in our final sample for the key dependent measures.

Study 2: Testing for effects of partisan identity, topic, attitude strength, and partisan identity strength on social sanctions.

Partisan Identity. First, we conducted two-way ANOVA to examine the effect of role (predictor vs. reactor) and party identification on social sanction composite scores. The analysis revealed a significant main effect of role, $F(1, 274) = 86.90$, $p < .001$; however, the main effect of party identification was not significant, $F(1, 274) = 0.47$, $p = .496$, and the interaction effect between role and party identification was also not significant, $F(1, 274) = 1.82$, $p = .179$, indicating that the effect of role on social sanction composite scores did not differ depending on party identification.

Topic. We conducted a 2 (role: predictor vs. reactor) × 3 (topic: abortion, guns, immigration) ANOVA to examine whether social sanction scores varied by role, topic, or their interaction. This model revealed a significant main effect of role, $F(1, 272) = 87.86$, $p < .001$, $\eta p^2 = .244$, such that predictors anticipated more rejection than reactors reported. There was also a small but significant main effect of topic, $F(2, 272) = 3.55$, $p = .030$, $\eta p^2 = .025$, indicating that overall rejection levels varied modestly across topics. The interaction between role and topic was not significant, $F(2, 272) = 0.12$, $p = .89$, $\eta p^2 = .001$, suggesting that the perception gap between predictors and reactors was consistent across issues.

Pairwise comparisons using Tukey-adjusted post hoc tests to account for multiple comparisons revealed that abortion elicited significantly higher rejection scores than immigration ($p = .029$), while differences between abortion and guns ($p = .186$) and between guns and immigration ($p = .796$) were not statistically significant. Means and standard deviations for each

topic (collapsed across role) were as follows: abortion (M = 2.78, SD = 1.23, n = 108), guns (M = 2.51, SD = 1.09, n = 76), and immigration (M = 2.42, SD = 1.05, n = 94).

Although the overall topic effect was statistically significant, its effect size was small and not qualified by an interaction with role. Thus, while there may be minor topic-level differences in rejection, these do not appear to meaningfully alter the strength of the perception gap observed across the studies.

Attitude Strength. To test whether attitude strength moderated the effect of role on social sanctions, we regressed social sanction scores on role, attitude strength (measured as agreement with the issue position on the randomly assigned topic; 1 = *Strongly disagree*; 7 = *Strongly agree*), and their interaction. The overall model was significant, $F(3, 274) = 30.10$, $p < .001$, $R^2 = .25$. There was a significant main effect of role, $b = –2.67$, $SE = 1.03$, $t = –2.60$, $p = .009$, such that predictors anticipated more rejection than reactors reported. However, there was no significant main effect of attitude strength, $b = –0.05$, $SE = 0.12$, $t = –0.46$, $p = .64$, nor a significant interaction between role and attitude strength, $b = 0.24$, $SE = 0.16$, $t = 1.52$, $p = .13$. These results suggest that the strength of participants' agreement with the issue position on the randomly assigned topic did not meaningfully moderate anticipated rejection, nor the size of the perception gap.

Partisan Identity Strength. A linear regression was conducted to examine whether political identity strength(1 = *Strongly Democrat*; 4 = *Somewhere in the middle*, 7 = *Strongly Republican*) moderated the effect of role on social sanction. The overall model was significant, $F(3, 274) = 30.00$, $p < .001$, $R^2 = .25$. There was a significant main effect of role, $b = –1.37$, $SE = 0.22$, $t = –6.23$, $p < .001$, such that predictors anticipated greater social rejection than reactors reported. The main effect of political identity was not statistically significant, $b = –0.07$, $SE = 0.04$, $t = –1.68$, p

= .095, suggesting that rejection did not vary by partisan identity strength. Additionally, the interaction between role and political identity was not significant, $b = 0.08$, $SE = 0.06$, $t = 1.41$, $p = .161$, suggesting that the perception gap between predictors and reactors did not vary significantly by political identity.

Study 3: Additional Analyses

Study 3: Individual Sample Information.

Due to challenges in collecting a large sample of participants who reported a dissenting shift in their political beliefs following the counter-attitudinal writing task, participants in Phase 1 of Study 3 were collected from three separate studies with a similar experimental procedure. Here, we describe the characteristics of each of these samples. Later in the Supplement (Study 3: Additional Information), we report the minor differences between the studies.

Sample A consists of only predict condition participants who completed the counter-attitudinal writing task. Sample B consists of only react condition participants who reported their reactions to an ingroup member who completed a counter-attitudinal writing task. Sample C consisted of both predictors who completed the counter-attitudinal writing task and reactors who evaluated an ingroup member who completed a counter-attitudinal writing task. Dissenting belief change was calculated by subtracting T1 (pre-task) agreement with the ingroup's view on the randomly assigned political topic (abortion, gun control, immigration if participants reported that they agreed with the typical view of their party).

We originally planned to treat Sample A (predict condition) and Sample B (react condition) as a pilot test of this experimental paradigm and then to run Sample C as a fully-powered pre-registered replication including both predict and react conditions. However, only a small minority of predict condition participants in Sample C reporting dissenting belief change after completing the writing task (17.82%), which left us significantly under-powered. Given the similarity between these resource-intensive study procedures, we combined them for analysis in Study 3. Below, we report the belief change outcomes observed in each sample.

Sample A. In sample A, 85 (35.42%) predictors reported dissenting belief change, 111 (46.25%) reported no change, and 44 (18.33%) reported stronger agreement after completing the counter-attitudinal writing task.

Sample B. Sample B consisted of reactors only and did not employ the writing task.

Sample C. In sample C, 62 (17.82%) predictors reported dissenting belief change, 251 (72.12%) reported no change, and 35 (10.05%) reported stronger agreement after completing the counter-attitudinal writing task.

Key Analysis in Each Sample.

Samples A-B. Predictors in Sample A (n = 85) who reported dissenting belief change (M = 3.90; SD = 1.60) anticipated significantly harsher social sanctions compared to the actual social sanctions reported by reactors in Sample B (n = 204) who evaluated an ingroup member who reported dissenting belief change after completing the writing task (M = 2.57; SD = 1.64), t(326) = 6.87, p < .001, d = .82.

Sample C. Predictors in Sample C (n = 62) who reported dissenting belief change (M = 3.36.; SD = 1.66) anticipated harsher social sanctions compared to the actual social sanctions reported by reactors in Sample C (n = 143) who evaluated an ingroup member who reported dissenting belief change after completing the writing task (M = 3.01; SD = 1.67); however, this difference was not statistically significant, t(203) = 1.37, p = .174, d = .21. The results of this analysis should be interpreted with caution due to concerns about the statistical power of the test.

Additional Analysis: Sample C - Dictator Game. Sample C also included a dictator game measure. For this item, we found that participants in the predict condition whose beliefs decreased significantly underestimated how much their partner would keep for themselves

($M_{predict}$ = \$0.69) compared to what participants in the react condition decided to keep for themselves (M_{react} = \$0.82), $t(203) = -3.74$, $p < .001$, $d = .54$.

In Study 2, we observed that predictors overestimated how much their partner would take in the dictator game, and in Study 2 we replicated the same pattern with paired predictor-reactor dyads. In Sample C we found the opposite pattern. We believe that this discrepant pattern of results was due to several key differences in the way that the question was presented to participants in Sample C.

First, reactors in Studies 2 and 3 distributed the bonus between themselves and another participant in the study whom they were told they were paired with (in Study 2 this was someone whom they had met and with whom they believed they would interact again in the future). In Sample C, however, reactors were asked to split the bonus with another participant from a previous study. Thus, punishment was relatively less "costly" (Henrich et al., 2006) for participants in Sample C compared to Studies 2 and 3 given the temporal disconnect between the reactor and the recipient in Sample C, which likely led to more self-interested behaviors in the dictator game. Moreover, this measure may have been relatively less realistic and believable for participants in Sample C, who may have questioned whether we would actually retroactively grant a bonus to a participant from a previous study.

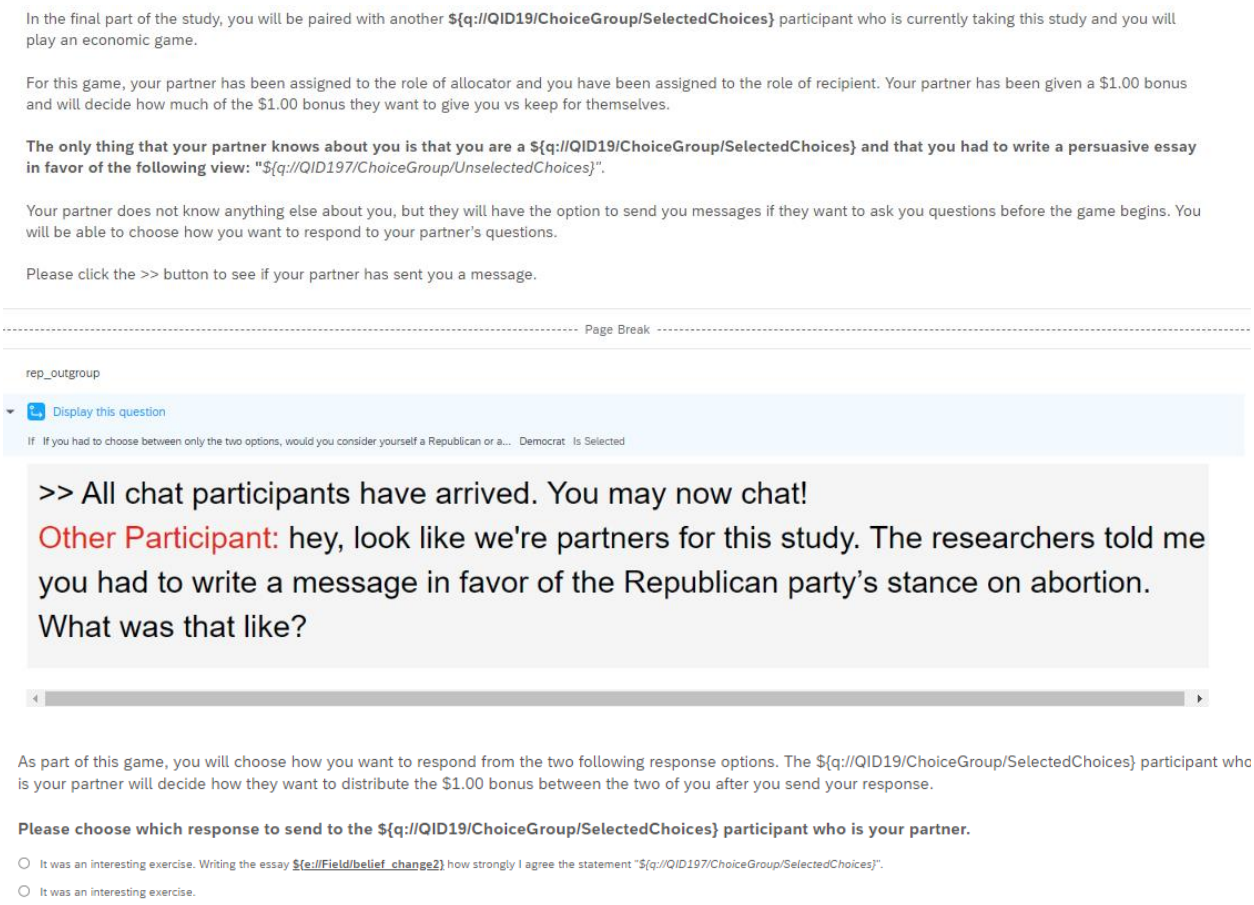
Furthermore, participants in Studies 2 and 3 were told that one participant would be randomly assigned to the role of "bonus allocator" (dictator) and asked to split up the bonus money between themselves and their partner (in reality, reactors were always assigned to the dictator role). In Sample C, participants were *given* a \$1.00 bonus and asked how much, if any, they would give to the participant from the previous study. This form of endowment likely led

reactors in Sample C to behave more selfishly than reactors in Studies 2 and 3 (i.e., "The Endowment Effect; Morewedge & Giblin, 2015).

Findings from previous research support the conclusion the presentation of the dictator game item in Study C biased how participants responded. Studies that have examined typical distribution patterns in dictator games show that participants in the dictator role—absent any experimental manipulation—tend to distribute the amount about 50/50 on average (Branas Garza et al., 2017). The same study found that participants in the recipient role, on average, expect that they will be distributed approximately 50% of the bonus. In our work, predictors at every level of belief change (including those whose beliefs did not change and whose beliefs strengthened) underestimated how much their react condition counter-part would keep in the dictator game, suggesting that some artifact of our experimental design biased results in a way that is inconsistent with previous research and our findings in Studies 2 and 3. Taken together, these findings suggest that our measure of behavioral punishment was not well calibrated to our study design in Sample C.

Study C: Behavioral Measure of Self-Censorship.

Participants in the predict condition in Sample C were shown the following text and asked the following question.



We initially intended to compare self-censorship frequency between participants in the predict condition who reported that their agreement with their party's view (a) weakened, (b) did not change, and (c) strengthened after completing the counter-attitudinal writing task. We hypothesized that participants whose agreement weakened (i.e., dissenting belief change) would self-censor more frequently than participants whose agreement did not change and strengthened; however, we later realized that this approach is methodologically and conceptually flawed in a way that compromises these comparisons. We originally intended to treat the different belief change outcomes as an independent variable; however, given the lack of random assignment among predictors to belief change conditions (i.e., one cannot experimentally assign people to experience a certain type of belief change), it would not be appropriate to compare predict condition participant scores across levels of belief change (weaken, no change, strengthen). This

approach is especially flawed because it is possible that an unaccounted-for person-level variable simultaneously influenced predictors' a) belief change outcomes, and b) sensitivity to rejection.

For example, there was a significant difference across belief change outcomes for predictors' baseline (T1) attitudes, $F(2, 345) = 56.48$, $p < .001$, $\eta^2 = 0.25$, predictors' strength of political identity, $F(2, 345) = 5.04$, $p = .007$, $\eta^2 = .003$, and the number of Democrats and Republicans for each belief change outcome level, $\chi^2(2, 347) = 7.56$, $p = .023$. These are a few examples that highlight the differences in stable characteristics between predictors whose beliefs weakened, did not change, and strengthened following the counter-attitudinal writing task, which led us to conclude that it would be invalid to treat belief change outcomes as an independent variable in our analyses. For transparency, we report the results from this pre-registered analysis below.

Self-Censorship Results. We found that the proportion of participants in the predict condition whose agreement weakened who self-censored (51.39%) was significantly higher than the proportion of participants whose agreement did not change (8.86%), $\chi^2(1, N = 389) = 70.38$, $p < .001$, 95% CI [30.50%, 54.56%]. However, a comparison of the proportion of participants whose agreement weakened (51.39%) and participants whose agreement strengthened (34.78%) did not yield a statistically significant difference, $\chi^2(1, N = 389) = 3.13$, $p = .077$, 95% CI [-1.36%, 34.57%]. While there was a trend suggesting a higher likelihood of self-censorship among participants in the decrease condition, this difference did not reach the conventional levels of statistical significance (again, we note the small sample sizes in these cells).

Study 3: Robustness Check Controlling for Partisan Identity Covariates. To assess whether the perception gap between predictors and reactors remained significant after accounting for individual differences in political identity, we conducted a linear model that included partisan

identity centrality (centered continuous) and partisan identity (Democrat vs. Republican) as covariates. This model tested whether predictors continued to anticipate more social rejection than reactors, above and beyond individual-level ideological orientation.

The model revealed a significant effect of role, $b = –0.73$, $SE = 0.16$, $t(490) = –4.66$, $p < .001$, such that predictors anticipated more social rejection than reactors actually reported. Political identity centrality was also a significant positive predictor of rejection, $b = 0.23$, $SE = 0.04$, $t(490) = 5.63$, $p < .001$. Party affiliation was not a significant predictor, $b = –0.09$, $SE = 0.14$, $t(490) = –0.63$, $p = .53$. Together, these predictors explained a modest but meaningful portion of the variance in rejection scores, $R^2 = .114$.

These results demonstrate that partisan identity centrality is associated with heightened expectations of social rejection: individuals who view their political identity as more central to their self-concept tend to anticipate more rejection. However, the effect of role remained significant even after adjusting for political identity centrality and party affiliation, indicating that the perception gap between predictors and reactors reflects a robust pattern not reducible to individual differences in political orientation or identity strength.

Study 3 Phase 1: Testing for moderation effects of partisan identity and partisan identity centrality. We note that no significant moderation effects were observed. For transparency, we report these findings below.

Partisan Identity. A 2 (role: predict vs. react) × 2 (party affiliation: Democrat vs. Republican) ANOVA on rejection scores revealed a significant main effect of role, $F(1, 490) = 29.15$, $p < .001$, $\eta^2 = .056$, such that predictors anticipated more rejection than reactors reported. There was no main effect of party affiliation, $F(1, 490) = 0.08$, $p = .78$, and no interaction between role and party affiliation, $F(1, 490) = 0.34$, $p = .56$. These results indicate that the perception gap in anticipated rejection did not vary by participants' political affiliation.

Partisan Centrality Strength. A linear model tested whether political identity centrality moderated the perception gap between predictors and reactors. The interaction between role and political identity centrality was not significant ($b = 0.00$, $SE = 0.10$, $p = .997$), suggesting no evidence of moderation. Neither the main effect of role ($b = –0.73$, $SE = 0.49$, $p = .136$) nor the main effect of political identity centrality ($b = 0.23$, $SE = 0.08$, $p = .006$) were robust in this model. Although the direction of the role effect remained consistent with prior findings, it was no longer statistically significant in this model, likely due to inflated standard error from including a non-informative interaction term. These findings suggest that political identity centrality does not meaningfully moderate or account for the perception gap.

Study 4A: Additional Analyses

ANOVA: Bipolar Social Evaluations. To help contextualize the expected and actual interpersonal penalties associated with various types of belief change in both absolute and relative terms, we report the ANOVA results and a series of one-sample t-tests in the following section.

We conducted a 2 (Role: predict vs. react) × 2 (Topic: partisan vs. non-partisan) ANOVA using type III sum of squares to test for an interaction and main effects of role and topic conditions on participants' bi-polar social evaluation scores, which detected a significant interaction of role and topic condition on bi-polar social evaluation composite scores, $F(1, 389) = 4.55$, $p = .033$, $\eta p^2 = .01$. Tukey's HSD post-hoc tests were conducted to account for multiple comparisons and revealed that predictors in the partisan topic condition anticipated significantly less favorable social evaluations ($M = -1.05$; $SD = 1.13$) compared to predictors in the non-partisan topic condition ($M = 0.17$; $SD = 1.12$), $t(389) = 7.67$, $p < .001$, $d = 1.10$. Relative to reactor reports, predictors in the partisan topic condition significantly overestimated negative evaluations compared to reactor ratings in the partisan topic condition ($M = -0.25$, $SD = 1.21$), $t(389) = 5.00$, $p < .001$, $d = .71$. In contrast, predictors in the non-partisan topic condition did not significantly overestimate negative social evaluations relative to reactors' scores in the non-partisan topic condition ($M = 0.48$; $SD = 1.00$), $t(389) = 1.94$, $p = .209$, $d = .28$. Finally, post-hoc tests showed that reactors rated dissenters in the partisan topic condition ($M = -0.25$; $SD = 1.21$) as significantly more negative than reactors rated dissenters in the non-partisan topic condition ($M = 0.48$; $SD = 1.00$), $t(389) = 4.57$, $p < .001$, $d = .66$.

To further contextualize these effects, we tested whether social evaluation scores in each condition differed significantly from zero (i.e., the midpoint of the scale representing a "neutral"

evaluation). On this scale, negative scores reflect negative social evaluations, while positive scores reflect positive evaluations. Predictors in the partisan topic condition anticipated being evaluated significantly more negatively than positively (M = -1.05, SD = 1.13), $t(92) = -8.96$, $p < .001$, d = -0.93. Reactors in the partisan condition also evaluated dissenters significantly below neutral (M = -0.25, SD = 1.21), $t(104) = -2.12$, $p = .037$, d = -0.21. In contrast, predictors in the nonpartisan condition expected to be evaluated more positively than negatively, but these ratings were not significantly different from zero (M = 0.17, SD = 1.12), $t(105) = 1.56$, $p = .121$, d = 0.15, Reactors in the nonpartisan condition evaluated dissenters positively (M = 0.48, SD = 1.00), $t(88) = 4.53$, $p < .001$, d = 0.48.

These findings suggest that group members who dissent on partisan topics do incur a mild social penalty: they are evaluated more negatively than those who dissent on nonpartisan topics, and these evaluations fall slightly below the neutral midpoint. In this sense, predictors are not wrong to anticipate some interpersonal cost for partisan dissent. However, they substantially overestimate the severity of that cost—expecting to be strongly disliked when, in fact, reactors respond with only mild negativity. By contrast, dissenting belief change on nonpartisan topics is expected to elicit—and does elicit—positive evaluations. Thus, while the perception gap reflects an exaggeration, it is grounded in a kernel of truth: partisan dissent carries social risk, but not nearly to the extent people imagine.

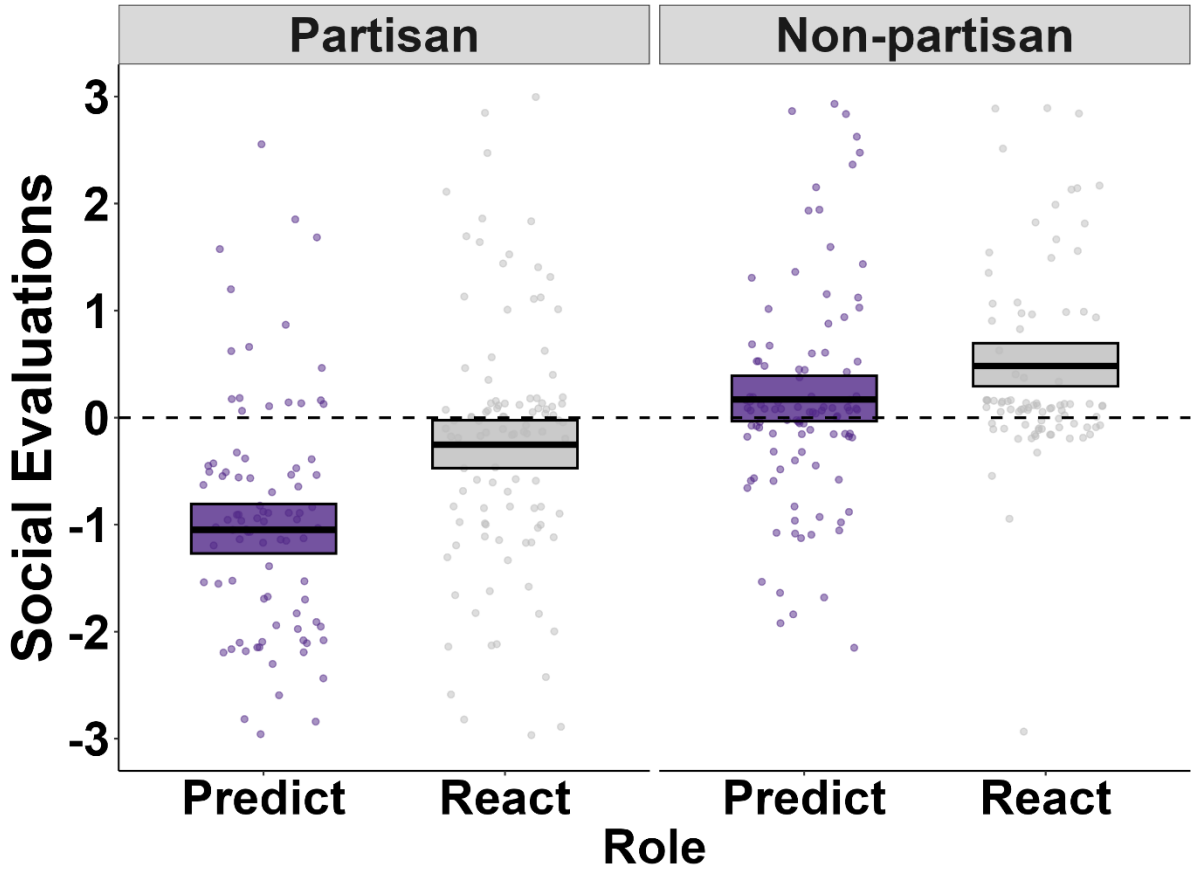


Figure. Comparing predictor estimates of social evaluations against reactors' actual evaluations within each topic condition. *Note.* The bipolar composite has been rescaled on the y-axis to center the midpoint on zero, which represents a neutral evaluation. Positive scores represent a positive social evaluation. Negative scores represent a negative social evaluation. The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval. Bipolar social evaluations were measured on seven-point response scales (e.g., -3 = *Strongly dislike*; 0 = *Neutral*; 3 = *Strongly like*).

The significant interaction of role and topic conditions on both the social sanctions composite measure and the bi-polar social evaluation measure suggests that the overestimation of social sanctions for dissenting belief change on partisan issues cannot be explained by a more general undersociality bias. The moderated mediation model results support the hypothesis that predictors overestimate the social sanctions they will face for belief change in partisan contexts, in part because they expect their dissent to send a stronger signal of partisan disloyalty than

reactors actually perceive. This finding aligns with the proposed mechanism of signal amplification bias, whereby dissenters expect that their belief change will be more diagnostic of their group loyalty than ingroup observers perceive. Importantly, this pathway was significantly stronger in the partisan condition—where loyalty concerns are heightened—suggesting that signal amplification bias is most likely to distort social expectations in politically charged contexts.

Moderated Mediation Analysis: Bipolar Social Evaluations. To test whether the interaction between topic condition (partisan vs. non-partisan) and role (predictor vs. reactor) on social evaluations was mediated by perceived disloyalty, we conducted a moderated mediation analysis using structural equation modeling (N = 393, 5,000 bootstrap samples). The bipolar composite measure of social evaluation served as the dependent variable (centered on zero, with higher scores reflecting more favorable evaluations).

We observed a significant role × topic condition interaction on loyalty, $b = –0.58$, $SE = 0.28$, $z = –2.10$, $p = .036$, such that predictors in the partisan condition expected to be seen as more disloyal compared to reactors' actual judgments. Loyalty scores, in turn, significantly predicted bipolar evaluations, $b = 0.44$, $SE = 0.04$, $z = 11.14$, $p < .001$, indicating that lower perceived loyalty was associated with more negative evaluations.

A significant moderated indirect effect emerged, $b = –0.26$, $SE = 0.13$, $z = –2.03$, $p = .042$, indicating that the indirect path from role to social evaluation via loyalty was stronger in the partisan condition. Specifically, the indirect effect was significant in the partisan condition, $b = –0.30$, $SE = 0.10$, $z = –3.02$, $p = .003$, but not in the non-partisan condition, $b = –0.05$, $SE = 0.08$, $z = –0.59$, $p = .557$.

The total moderated effect of topic condition on bipolar evaluations was also significant, $b = –0.48$, $SE = 0.23$, $z = –2.14$, $p = .033$, providing additional evidence that overestimation of negative social evaluation in polarized contexts is partly driven by exaggerated fears of appearing disloyal to the group.

Study 4a: Robustness Check Controlling for Political Identity Covariates. To examine the robustness of the central interaction between role and condition on social sanctions, we conducted a two-way ANOVA that included three covariates: partisan identity, partisan identity centrality, and partisan identity strength. This model tested whether the main effects and interaction remained significant when accounting for individual differences in political identification strength and centrality.

The 2 (Role: Predict vs. React) × 2 (Condition: Partisan vs. Non-partisan) ANOVA with covariates revealed a significant main effect of role, $F(1, 386) = 72.29$, $p < .001$, $\eta p^2 = .158$ and a significant main effect of condition, $F(1, 386) = 108.48$, $p < .001$, $\eta p^2 = .220$. Additionally, the role × condition interaction remained significant, $F(1, 386) = 11.15$, $p = .001$, $\eta p^2 = .028$, indicating that the perception gap was significantly larger in the partisan condition than the non-partisan condition, even after controlling for the covariates.

None of the covariates were significant predictors of social sanctions: party affiliation ($F(1, 386) = 0.14$, $p = .705$), political identity centrality ($F(1, 386) = 0.79$, $p = .375$), or partisan identity strength ($F(1, 386) = 0.47$, $p = .494$). These results demonstrate that the perception gap between predictors and reactors—and its amplification in the partisan condition—cannot be explained by differences in political identity strength or centrality.

Next, to test whether the effect of role (predict vs. react) on social sanctions remained significant after accounting for variation across topics, we conducted two linear models—one for

the partisan condition and one for the non-partisan condition—controlling for issue as a fixed effect. In the partisan condition, a Type III ANOVA revealed a significant main effect of role, $F(1, 194) = 62.50$, $p < .001$, $\eta p^2 = .116$, indicating that predictors anticipated significantly more social sanctions ($M = 4.63$, $SE = 0.16$) than reactors reported ($M = 2.86$, $SE = 0.15$). The effect of issue was also significant, $F(2, 194) = 6.20$, $p = .002$, $\eta p^2 = .025$, suggesting some variability in baseline sanction levels across topics. Estimated marginal means indicated that the difference between roles was robust across issue types, $t(194) = 7.91$, $p < .001$, $d = 1.13$.

In the non-partisan condition, the same pattern emerged. The role effect remained significant, $F(1, 191) = 17.27$, $p < .001$, $\eta p^2 = .083$, with predictors again anticipating greater social sanctions ($M = 2.47$, $SE = 0.13$) than reactors reported ($M = 1.68$, $SE = 0.14$). The issue effect was also significant, $F(2, 191) = 4.12$, $p = .018$, $\eta p^2 = .041$. The estimated difference between predictors and reactors was statistically significant, $t(191) = 4.16$, $p < .001$, $d = 0.60$.

These findings indicate that although social sanction scores vary somewhat across topics, the role-based perception gap is robust: predictors consistently anticipated more social sanctions than reactors reported across all issue types. The perception gap persists even after statistically adjusting for topic-level differences.

Study 4a: Self-censorship SEM Analyses

To assess predictors' willingness to openly share their political belief change, we administered a single-item self-censorship measure to participants in the predictor role only. After reporting their revised opinion, predictors were asked: *"*If this topic came up in conversation with another [ingroup member], how likely would you be to share that your beliefs on this topic changed?*"* Responses were recorded on a 7-point Likert-type scale (1 = *Not at all*

likely, 7 = *Very likely*). For interpretability, scores were reverse-coded such that higher values reflect greater likelihood of self-censoring.

We examined whether predictors' distorted expectations were associated with self-censorship. That is, we tested whether heightened concerns about loyalty and anticipated rejection in partisan contexts would be associated with individuals' willingness to disclose their belief change. An independent samples t-test revealed that predictors in the partisan topic condition reported significantly greater likelihood of self-censorship ($M = 4.68$, $SD = 1.90$) than those in the non-partisan condition (M = 3.*75*, $SD = 2.06$), $t(197) = –3.00$, $p = .001$, $d = –0.47$.

To test whether this effect was sequentially mediated by loyalty meta-perceptions and anticipated social sanctions, we conducted a mediation analysis among predictors ($N = 199$) using structural equation modeling with 5,000 bootstrap samples. In this model, topic condition (partisan = 1, non-partisan = 0) was the independent variable; loyalty meta-perceptions were the first mediator; anticipated social rejection was the second mediator; and self-censorship was the dependent variable (reverse-scored so that higher values reflect greater likelihood of censoring).

The results revealed a significant indirect effect of topic condition on self-censorship through both mediators—i.e., via loyalty meta-perceptions and then anticipated rejection ($\text{indirect}_1 = 0.503$, $SE = 0.132$, $z = 3.83$, $p < .001$, 95% CI [0.243, 0.768]). A second indirect pathway also emerged through anticipated rejection alone, bypassing loyalty ($\text{indirect}_2 = 0.448$, $SE = 0.133$, $z = 3.38$, $p = .001$, 95% CI [0.186, 0.708]). The total indirect effect was significant (total = 0.932, $SE = 0.276$, $z = 3.37$, $p = .001$), whereas the direct effect of topic condition on self-censorship was not ($c' = –0.019$, $SE = 0.314$, $z = –0.06$, $p = .951$), consistent with the possibility of full mediation.

These findings are consistent with the hypothesis that participants were more likely to self-censor in the partisan condition because they expected their belief change to signal disloyalty and provoke harsher social sanctions. The absence of a significant direct effect alongside a robust total indirect effect further indicates that this behavioral reluctance to disclose dissent was mediated by loyalty meta-perceptions and anticipated rejection—underscoring signal amplification as a key psychological mechanism to help explain this relationship.

Study 4B: Additional Analyses

Study 4b: Robustness Check Controlling for Political and Topical Covariates. To test the robustness of the key role × condition interaction on social sanctions observed in Study 4b, we conducted a linear model that included all relevant participant and topic-level covariates. Specifically, we included partisan identity, partisan identity strength (centered), perceived believability of the manipulation (centered), and issue topic as covariates. This model allowed us to assess whether the effect of role and condition on social sanctions remained significant after statistically adjusting for these variables.

A Type III ANOVA revealed that the main effect of role was significant, $F(1, 274) = 10.82$, $p = .001$, $\eta p^2 = .038$, and the main effect of condition was highly significant, $F(1, 274) = 30.65$, $p < .001$, $\eta p^2 = .101$. Additionally, the role × condition interaction remained significant, $F(1, 274) = 7.51$, $p = .006$, $\eta p^2 = .027$, indicating that the perception gap was robust even after controlling for partisan identity, partisan identity strength, perceived believability, and issue topic.

None of the covariates explained significant variance in social sanctions: political party ($F(1, 274) = 1.26$, $p = .264$), perceived believability ($F(1, 274) = 0.10$, $p = .750$), or issue topic ($F(1, 274) = 2.66$, $p = .104$). Political identity strength approached significance, $F(1, 274) = 3.69$, $p = .056$, suggesting a possible weak association. Together, these results demonstrate that the role × condition interaction observed in Study 4b is not an artifact of participant ideology, identity strength, topic believability, or issue content.

Study 4b: Exploratory Moderation Effects. To assess whether the core perception gap effect observed in Study 4b varied by participant characteristics or topic content, we conducted a series of exploratory moderation analyses. Specifically, we tested whether the role × condition

interaction was moderated by participants' political party (Democrat vs. Republican), political identity strength, believability of the manipulation, or issue topic (gambling vs. surveillance). Across all models, the key role × condition interaction remained statistically significant (*ps* < .025), and no three-way interaction reached significance. This suggests that the role × condition interaction on social rejection was robust across partisan identities, levels of political identity strength, perceived credibility of the manipulation, and issue topic.

Study 4b: Robustness Check - Results Replicate Among Participants Who Believed the Manipulation. To test the robustness of our findings, we conducted a moderated mediation analysis restricted to participants who rated the experimental framing as moderately to highly believable (believability ≥ 4 on a 7-point scale; $n = 235$). This analysis replicated the same findings for all hypothesis tests reported in the main text. These findings are reported below.

Using the same structural equation model employed for the main analyses of Study 4b in the main text, we found that the role × topic condition interaction significantly predicted loyalty judgments, $b = –1.16$, $SE = 0.35$, $z = –3.32$, $p = .001$, such that predictors in the polarized condition expected to be seen as less loyal (compared to reactors), with no such difference in the non-polarized condition. Lower loyalty scores, in turn, significantly predicted higher anticipated social sanctions, $b = –0.50$, $SE = 0.08$, $z = –6.33$, $p < .001$. A moderated mediation analysis revealed a significant index of moderated mediation, $b = 0.57$, $SE = 0.19$, $z = 2.96$, $p = .003$, indicating that the indirect effect of role on anticipated rejection via loyalty was significantly stronger in the polarized condition, $b = 0.62$, $SE = 0.14$, $z = 4.35$, $p < .001$, than in the non-polarized condition, $b = 0.05$, $SE = 0.13$, $z = 0.39$, $p = .695$.

Study 4C: Additional Analyses

Study 4c: Robustness Check Controlling for Political and Topical Covariates. To test whether the role × condition interaction on social sanctions observed in Study 4c remained robust when accounting for relevant covariates, we ran a linear model including partisan identity, partisan identity strength (centered), perceived believability of the manipulation (centered), and issue topic as covariates.

The results confirmed a significant main effect of role, $F(1, 588) = 35.07$, $p < .001$, $\eta p^2 = .056$, and a significant main effect of condition, $F(1, 588) = 46.04$, $p < .001$, $\eta p^2 = .073$. Additionally, the role × condition interaction remained significant, $F(1, 588) = 6.06$, $p = .014$, $\eta p^2 = .010$, indicating that the perception gap for social sanctions persisted even after adjusting for political affiliation, identity strength, believability, and issue topic.

Among the covariates, only perceived believability emerged as a significant predictor, $F(1, 588) = 7.07$, $p = .008$, $\eta p^2 = .012$. Neither party affiliation ($F = 0.02$, $p = .900$), political identity strength ($F = 0.98$, $p = .323$), nor issue topic ($F = 0.70$, $p = .402$) significantly predicted anticipated rejection.

Study 4C: Moderation Analyses. To test the robustness of the role × condition interaction, we conducted a series of exploratory moderation analyses including issue type, party affiliation, party ID strength, and believability of the manipulation. In each model, the role × condition interaction remained statistically significant or marginal, but none of the three-way interactions were significant. However, we did detect a marginally significant three-way interaction with believability, $F(1, 588) = 3.78$, $p = .052$. The believability manipulation also interacted significantly with condition, $F(1, 588) = 19.90$, $p < .001$, suggesting that participants' perceptions of the credibility of the framing influenced their rejection expectations. This

pattern—absent in Study 4b—may help explain why the role × condition interaction failed to translate into a moderated mediation effect in Study 4c.

Robustness Check: Moderated Mediation Among Participants Who Believed the Manipulation. To assess the robustness of our findings, we re-ran the moderated mediation analysis using only the subset of participants ($N = 480$) who rated the framing manipulation as at least moderately believable (i.e., ≥ 4 on the 7-point believability scale). The model structure and analytic approach were identical to those used in the main analysis. In summary, we replicate the same pattern of results reported in the main text with the full sample of participants. The model results for this subset is reported below.

We first examined whether the effect of role on anticipated rejection was moderated by topic condition. As in the full sample, the interaction between role and condition did not significantly predict loyalty, $b = 0.006$, $SE = 0.27$, $z = 0.02$, $p = .983$. However, both role and condition significantly predicted loyalty perceptions, $b = –0.75$, $SE = 0.18$, $z = –4.20$, $p < .001$ and $b = –1.39$, $SE = 0.18$, $z = –7.90$, $p < .001$, respectively.

Loyalty judgments in turn significantly predicted anticipated rejection, $b = –0.46$, $SE = 0.05$, $z = –8.76$, $p < .001$. Consistent with this, the indirect effect of role on rejection via loyalty was significant in both the partisan condition, $b = 0.34$, $SE = 0.10$, $z = 3.54$, $p < .001$, and the non-partisan condition, $b = 0.34$, $SE = 0.09$, $z = 3.77$, $p < .001$. However, the difference in indirect effects (i.e., moderated mediation) remained non-significant, $b = –0.003$, $SE = 0.12$, $z = –0.02$, $p = .983$.

The total moderated effect of role × condition on rejection remained significant, $b = 0.88$, $SE = 0.28$, $z = 3.13$, $p = .002$. These findings suggest that the perception gap in anticipated rejection remains larger in polarized versus non-polarized topics, even among participants who

found the framing manipulation believable. However, the absence of a significant interaction on loyalty continues to indicate that the psychological mechanism driving this effect does not vary meaningfully by topic condition.

Study 5: Additional Analyses

Testing the Specificity of the Loyalty Manipulation

To examine whether the manipulation had unintended effects on related constructs, we conducted robustness checks testing for potential spillover to variables that might plausibly serve as alternative mediators. Specifically, we analyzed whether condition affected predictors' meta-perceptions of how aligned they would be seen by others, or their own self-reported alignment with the ingroup.

A one-way ANOVA comparing predictors' expectations of how much the ingroup member would perceive their views as misaligned with typical party beliefs (relative to reactors' actual perceptions) was not statistically significant, $F(2, 617) = 2.41$, $p = .091$, $\eta p^2 = .008$. This suggests that predictors in both conditions were similar and reasonably accurate in estimating how aligned others would perceive them to be. Similarly, an independent samples t-test revealed no significant difference in self-reported alignment between the high and low loyalty conditions ($p = .680$). Taken together, these results suggest that the loyalty intervention specifically influenced perceptions of loyalty without altering perceptions or meta-perceptions of issue alignment—supporting the theoretical specificity of the manipulation.

Supplementary Results: Covariate and Topic-Robustness Models

We conducted a series of robustness checks to assess whether the effects observed in our planned contrasts were robust to topical and demographic variation. Specifically, we re-ran both preregistered contrasts controlling for the topic of belief change, participant gender, age, political identification strength, party affiliation, self-reported alignment with political ingroup, and perceived disagreement with the ingroup's views (i.e., meta-perceptions of disagreement).

For Contrast 1, the overestimation effect remained robust after including all covariates. Predictors anticipated significantly more social rejection than reactors reported, $b = 0.83$, $SE = 0.08$, $t(598) = 10.01$, $p < .001$. The model accounted for 17.7% of the variance in rejection expectations, $R^2 = .177$, $F(13, 598) = 9.88$, $p < .001$. None of the covariates emerged as statistically significant predictors (ps > .10), including perceived disagreement ($b = –0.09$, $p = .18$) and self-alignment ($b = –0.05$, $p = .48$). Topic and demographic covariates also had no meaningful influence on the outcome. These findings support the robustness of the overestimation effect observed in Contrast 1.

For Contrast 2, the effect of loyalty affirmation also remained significant in the fully adjusted model. Participants in the high loyalty condition anticipated significantly less social rejection than those in the low loyalty condition, $b = –0.28$, $SE = 0.08$, $t(598) = –3.64$, $p < .001$. The model accounted for 6.0% of the variance in rejection expectations, $R^2 = .060$, $F(13, 598) = 2.92$, $p < .001$. Among covariates, only self-alignment emerged as a significant predictor, $b = –0.18$, $p = .011$. Perceived disagreement was nonsignificant, $b = 0.02$, $p = .77$. This reinforces the idea that affirming loyalty reduces anticipated rejection independently of broader perceived belief divergence from the group.

***Mediation Analyses Including Covariates*.**

To ensure the robustness of our mediation findings, we conducted re-ran the mediation analyses for both Contrast 1 (Predictors vs. Reactors) and Contrast 2 (High vs. Low Loyalty within Predictors) reported in the main text while controlling for key covariates: participant gender, age, political identity strength, party affiliation, self-reported ingroup alignment, topic label, and level of perceived disagreement.

For Contrast 1, we examined whether the effect of role (predictor vs. reactor) on anticipated social rejection was mediated by perceived loyalty (meta-perceptions for predictors; perceptions for reactors). The indirect effect remained significant when covariates were included, ACME = 0.15, 95% CI [0.07, 0.23], $p < .001$. The direct effect of role on rejection (ADE) also remained significant, ADE = 0.68, 95% CI [0.53, 0.83], $p < .001$, indicating partial mediation. The total effect was consistent with previous analyses, Total Effect = 0.83, 95% CI [0.67, 0.99], $p < .001$. These results confirm that loyalty judgments partially mediated the perception gap in anticipated rejection, even after adjusting for theoretically relevant covariates.

For Contrast 2, we tested whether the effect of loyalty affirmation (high vs. low) on anticipated rejection was mediated by loyalty meta-perceptions. Again, the indirect effect was significant with covariates included, ACME = –0.10, 95% CI [–0.18, –0.03], $p = .003$. The direct effect (ADE) also remained significant, ADE = –0.18, 95% CI [–0.31, –0.04], $p = .010$, suggesting partial mediation. The total effect was nearly identical to the unadjusted model, Total Effect = –0.28, 95% CI [–0.43, –0.13], $p < .001$. These results support the interpretation that affirming one's loyalty attenuated anticipated rejection in part by boosting loyalty meta-perceptions.

Parallel Mediation Analyses

We conducted parallel mediation models to test whether loyalty meta-perceptions remained a significant mechanism when adjusting for three competing mediators: self-reported loyalty, meta-perceptions of ingroup alignment (disagreement), and self-reported ingroup alignment. We report each contrast below.

Contrast 1 (Predictors vs. Reactors).

The full parallel mediation model revealed a significant indirect effect of role condition (predict vs. react) on anticipated rejection through loyalty meta-perceptions, $b = 0.15$, 95% CI [0.07, 0.23], $p < .001$. None of the three competing mediators showed significant indirect effects: self-reported loyalty ($b = –0.007$, 95% CI [–0.025, 0.005], $p = .35$), disagreement ($b = 0.018$, 95% CI [–0.001, 0.045], $p = .14$), and self-reported ingroup alignment ($b = 0.007$, 95% CI [–0.014, 0.033], $p = .55$). The total indirect effect was significant, $b = 0.17$, 95% CI [0.09, 0.25], and the direct effect remained significant after accounting for all mediators, $b = 0.67$, 95% CI [0.52, 0.82], $p < .001$. These findings suggest that loyalty meta-perceptions uniquely accounted for the overestimation of social rejection in this context.

Contrast 2 (High vs. Low Loyalty).

We next tested whether the effect of the loyalty affirmation manipulation on anticipated rejection was uniquely mediated by loyalty meta-perceptions, adjusting for self-reported loyalty and disagreement. The indirect effect through loyalty meta-perceptions was significant, $b = –0.14$, 95% CI [–0.23, –0.07], $p < .001$. Neither self-reported loyalty ($b = 0.04$, 95% CI [–0.01, 0.08], $p = .11$) nor disagreement ($b = 0.01$, 95% CI [–0.01, 0.04], $p = .32$) showed significant indirect effects. The total indirect effect was significant, $b = –0.10$, 95% CI [–0.18, –0.02], and the direct effect of the manipulation remained significant, $b = –0.18$, 95% CI [–0.32, –0.05], $p = .008$. These results confirm that the intervention reduced anticipated rejection primarily by shifting participants' perceptions of how loyal they would be seen by other group members.

Study 5: Topics of Belief Change by Party ID

At the beginning of Study 5, participants were asked if they had changed their views on a way that put them at odds with their political ingroup (i.e., dissenting belief change) on any of

the following topics. The table below depicts the frequency by which each topic was selected by Democrat and Republican participants, respectively. Note: These topics were selected based on pilot data which showed these topics to be the most frequently listed topics of belief change for Democrats and Republicans, respectively. As a result, two topics were only shown to Republicans (Climate Change, Economic Policy), and two topics were shown only to Democrats (Israel/Palestine Conflict, Gender Issues).

Table. Topics of belief change in Study 5 by participant partisan identity.

Topic	**Republican**		**Democrat**	
Immigration	103	28%	69	27%
Abortion Access	79	22%	43	17%
Gun Ownership	55	15%	47	19%
Climate Change	67	18%		
Economic Policy	63	17%		
Israel/Palestine Conflict			58	23%
Gender Issues			36	14%

Study 5: Topic Distribution Check.

For our primary pre-registered analyses, we planned to pool variance across belief change topics within each of our three conditions; therefore, we sought to ensure that topic distributions were approximately balanced across conditions. As described in the study procedures, we used a weighted random assignment procedure for reactors based on pilot base rates in order to approximate balance. We tested this assumption using chi-square tests of independence for Democrat and Republican participants separately. This served as a check for the success of our

weighted randomization procedure, which was designed to ensure that reactors were assigned to evaluate belief change topics at proportions that reflected the actual distribution of predictor-reported topics.

The chi-square test of independence for Democrats indicated no significant difference in topic distributions between predictors and reactors, $\chi^2(4) = 9.00$, $p = .06$, suggesting that topics were evenly distributed between predict and react conditions. For Republicans, the chi-square test of independence was significant, $\chi^2(4) = 15.00$, $p = .005$, suggesting that the distribution of topics between Republican predictors and reactors was unbalanced. Follow-up chi-square tests for each topic revealed that immigration was the only topic where there was a significant imbalance, where a significantly higher percentage of Republican predictors (26.6%) reported belief change than the percentage of Republican reactors assigned to evaluate belief change on immigration (12.2%), $\chi^2(1) = 9.00$, $p = .002$. Other topics did not show statistically significant mismatches, suggesting that the overall chi-square effect was driven by the topic of immigration.

To ensure this imbalance did not bias our primary analyses, we implemented the topic-matched analytic strategy detailed in our preregistration. This involved computing topic-specific prediction error scores (anticipated rejection minus average reactor response for the same topic) and comparing these scores between predictor conditions. We found that all key effects using a pooled variance approach replicated in this topic-matched framework. Therefore, because topic distributions were successfully balanced for Democrats, and the topic-matched analysis for Republicans produced convergent results, we report our primary planned analysis using pooled condition-level variance in the main text. Full details and results of the topic-matched robustness checks are reported in the next section.

Study 5: Topic-Matched Analyses

As a pre-registered contingency plan, we conducted a set of topic-matched analyses to account for potential imbalances in the distribution of topics across conditions. Given the significant chi-square test among Republican participants, we executed this topic-matched analysis plan. These analyses replicated the primary results using *topic-matched prediction error*, calculated for each participant as the difference between their anticipated rejection score and the average rejection score reported by reactors assigned to the same topic.

Republican Sample. Chi-square analyses indicated significant topic imbalance among Republican participants across predictor and reactor conditions. We therefore implemented the pre-registered contingency plan to compute topic-matched prediction error and tested whether this differed by condition. As expected, participants in the high loyalty condition (M = 1.04) exhibited significantly smaller prediction error than those in the low loyalty condition (M = 1.49), $t(242) = -2.00$, $p = .02$, 95% CI [–0.84, –0.06], $d = 0.29$.

Next, we tested whether this effect was mediated by loyalty meta-perceptions. A causal mediation analysis revealed a significant indirect effect of condition on prediction error through loyalty meta-perceptions (ACME = 0.22, 95% CI [0.02, 0.44], $p = .030$). The direct effect was nonsignificant (ADE = 0.23, 95% CI [–0.11, 0.58], $p = .193$), and the total effect was significant (Total Effect = 0.45, 95% CI [0.06, 0.84], $p = .023$). Approximately 49.5% of the total effect was mediated, 95% CI [3%, 161%], $p = .043$. These findings replicate the primary results and confirm that the effect of loyalty affirmation on prediction error is significantly explained by changes in loyalty meta-perceptions.

Democrat Sample. Although topic imbalance was not observed for Democrat participants, we conducted a parallel topic-matched analysis for transparency and comparison. The pattern of results closely mirrored the findings observed among Republicans. Participants in the high

loyalty condition (M = 0.94) again exhibited significantly smaller prediction error than those in the low loyalty condition (M = 1.60), $t(154) = -3.00$, $p = .005$, 95% CI [–1.11, –0.20], $d = 0.46$.

A mediation analysis confirmed that loyalty meta-perceptions significantly mediated this effect (ACME = –0.39, 95% CI [–0.69, –0.12], $p = .002$). The direct effect was not statistically significant (ADE = –0.27, 95% CI [–0.65, 0.12], $p = .178$), while the total effect remained significant (Total Effect = –0.65, 95% CI [–1.10, –0.20], $p = .005$). The proportion mediated was 58.9%, 95% CI [25%, 144%], $p = .006$. Together, these findings provide convergent support for the role of loyalty meta-perceptions in explaining variation in perceived social sanctions for dissenting belief change across both partisan groups.

Study 5: Contrast and Mediation Analyses Using Bi-Polar Social Evaluation Measure.

We re-ran the two key contrast analyses using a separate, bi-polar measure of social evaluation: **liking**. This item asked participants how much they thought their group would like (predictors) or actually liked (reactors) someone who changed their mind on the issue, using a 7-point scale (1 = *strongly dislike*; 4 = *neutral*; 7 = *strongly like*).

Contrast 1: Predictors vs. Reactors. We first tested whether predictors underestimated how much they would be liked, relative to how much reactors actually reported liking someone who changed their mind. A contrast-coded regression revealed a significant effect, b = –0.42, SE = 0.07, t(618) = –6.14, p < .001, d = 0.49, 95% CI [–0.56, –0.29]. As expected, predictors anticipated being liked less (M = 4.11; *SD* = 1.25) than reactors actually reported liking them (M = 4.74; *SD* = 1.16), replicating the perception gap with a distinct evaluative measure.

Contrast 2: High vs. Low Loyalty (within predictors). We next examined whether affirming one's loyalty improved anticipated evaluations among predictors. The analysis revealed that

participants in the high loyalty condition anticipated significantly more liking (M = 4.33; *SD* = 1.27) than those in the low loyalty condition (M = 3.89; *SD* = 1.22) , b = 0.22, SE = 0.06, t(618) = 3.58, p < .001, d = 0.29, 95% CI [0.10, 0.47]. These results conceptually replicate the main findings using an alternative measure of social evaluation, suggesting that signal amplification bias extends beyond anticipated rejection to broader social judgments like likability.

Next, we re-ran our primary mediation models using liking as the outcome variable.

Contrast 1. We first tested whether the effect of role (predictor vs. reactor) on liking was mediated by loyalty meta-perceptions. The indirect effect was significant, ACME = –0.15, 95% CI [–0.23, –0.07], p < .001, indicating that predictors' lower anticipated liking was partially explained by lower loyalty meta-perceptions. The direct effect remained significant, ADE = –0.28, 95% CI [–0.39, –0.16], p < .001, and the total effect was b = –0.42, 95% CI [–0.56, –0.29], p < .001. Approximately 35% of the total effect was mediated by loyalty meta-perceptions, Prop. Mediated = 0.35, 95% CI [0.18, 0.54], p < .001.

Contrast 2. We then tested whether the effect of loyalty affirmation on anticipated liking was mediated by loyalty meta-perceptions. Again, the indirect effect was significant, ACME = 0.13, 95% CI [0.06, 0.20], p < .001, suggesting that affirming loyalty increased perceived standing, which in turn improved anticipated social evaluations. The direct effect was marginal, ADE = 0.09, 95% CI [–0.01, 0.20], p = .069, and the total effect remained significant, b = 0.22, 95% CI [0.10, 0.35], p < .001. Over half of the total effect was mediated by loyalty meta-perceptions, Prop. Mediated = 0.58, 95% CI [0.32, 1.05], p < .001.

These findings provide further evidence that signal amplification bias extends beyond anticipated rejection to more general evaluations of social standing, such as likability, and that these perceptions can be shifted through targeted affirmation of loyalty to one's group.

Supplemental Study 1: Additional Analyses

Supplemental Study 1*: Examining Selection Bias When Selecting the Sample of Participants for Supplemental Study 1.*

We first conducted a pre-screen survey to identify participants to recruit for Supplemental Study 1 based on their responses to an attitude strength measure in the pre-screen. During the pre-screen, participants were asked to report the following: (1) Their partisan identity, (2) which of the following statements best represents their views on abortion access: "Abortion access in this country should be [**protected/restricted**]", and (3) how strongly they agree with the statement that they selected on a 0-100 scale (0 = *Completely Disagree;* 100 = *Completely Agree*). Only participants who reported that they agreed with the typical view of their party were eligible for Supplemental Study 1. The goal of this pre-screen survey was to identify at least 240 (120 Republicans; 120 Democrats) "extreme" participants who indicated that they agreed with the statement they selected on abortion between 95-100 on the agreement scale and at least 240 (120 Republicans; 120 Democrats) "moderate" participants who indicated they agree with the statement between 50-85 to re-recruit for Supplemental Study 1. In total, $N = 2{,}069$ participants completed the pre-screen. Among these, we randomly selected an equal number of Democrats and Republicans who met our recruitment criteria in each category ($n = 515$) and invited them to participate in Supplemental Study 1 twenty-four hours later.

Among participants who reported that they supported the typical view of their party ($n = 1{,}406$), the vast majority reported that they agreed with the statement between 95-100 on the 0-100 scale of agreement (82%), whereas a small minority (18%) reported agreement between 50-85 on the agreement scale. Thus, we re-recruited all eligible "moderate" participants for Supplemental Study 1 and we had to randomly select a subset of "extreme" participants to

participate in Supplemental Study 1. In Table S3 below, we report the characteristics of the full sample of "extreme" participants from the pre-screen who met our recruitment criterion and compare them to the subset who were randomly selected to be invited for Supplemental Study 1 (n = 515). The results from these analyses suggest that there is a very low probability of selection bias in our Supplemental Study 1 sample.

Table S3: Comparing the randomly selected subset of "Extreme" participants to the full sample of "Extreme" participants in the pre-screen.

Variable	Full Sample Extremists	Selected Subset Extremists	Mean Difference	t-value	df	p-value
Agreement (0-100)	99.76	99.70	0.06	1.20	1409	.231
Poli ID Centrality	4.56	4.64	-0.08	-0.69	1409	.493

Supplemental Study 1: Exploring Attrition Effects Between the pre-screen and Supplemental Study 1.

We invited 515 participants to participate in Supplemental Study 1. Among these, 435 accepted the invitation to participate. An attrition analysis was conducted to examine whether there were significant differences in the characteristics of participants who accepted the invitation to participate in Supplemental Study 1versus those who did not. First, we explored whether any of our four recruitment criterion subsets were more or less likely to accept the Supplemental Study 1 invitation: (1) "Moderate" Republicans, (2) "Moderate" Democrats, (3) "Extreme" Republicans, and (4) "Extreme" Democrats.

We performed a chi-square test of independence to examine the association between subset criterion and participation in Supplemental Study 1. The chi-square test was not

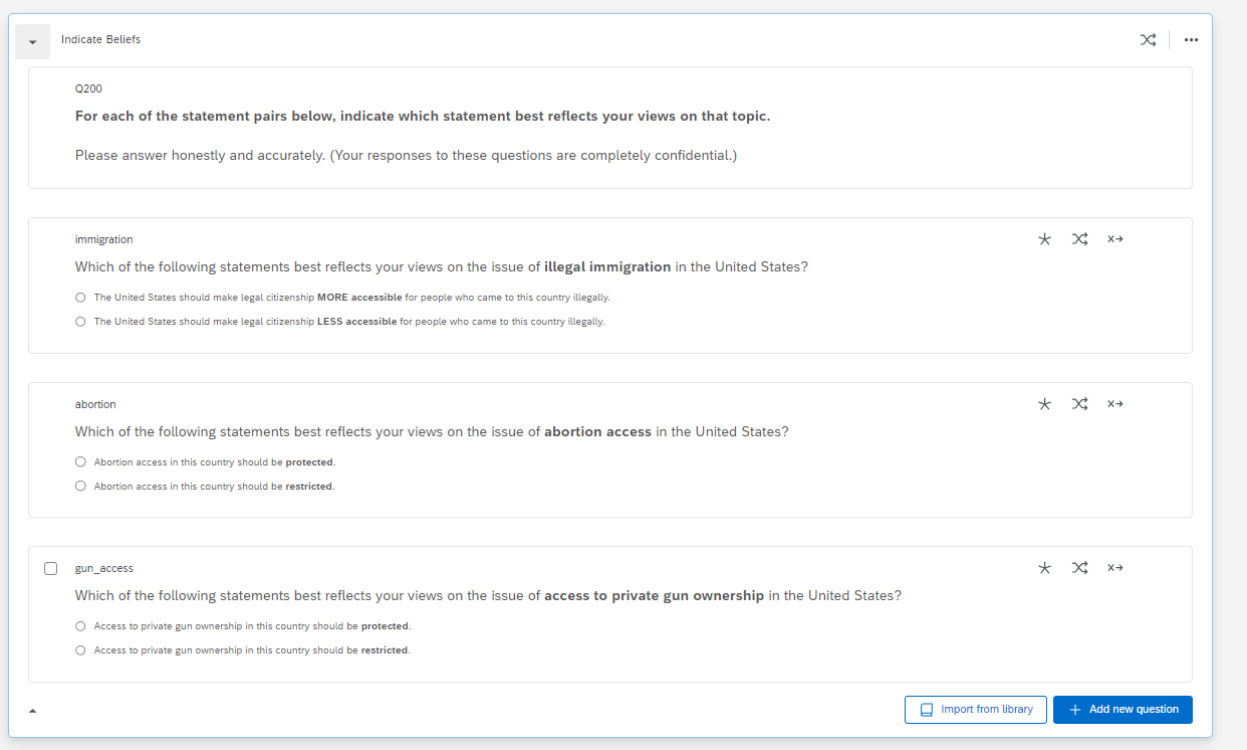
significant, $\chi^2(3, N = 515) = 6.99$, $p = .072$, suggesting that a difference in participation rates across the groups is unlikely. Moreover, we find that "agreement" (from 0-100) among participants who did not accept the Supplemental Study 1 invitation ($M = 86.89$) was not significantly different from participants who did accept the Supplemental Study 1 invitation ($M = 88.27$), $t(517) = 0.82$, $p = .414$. Taken together, these results suggest that attrition effects are unlikely to impact results observed in the final sample of the study.

Meta-Analytic Tests

A fixed-effects meta-analysis including all eight experimental tests of the overestimation effect (including Study 3 Phase 2 and using only the partisan condition subsamples in Studies 4a–4c) yielded a weighted mean Cohen's *d* of 0.87, 95% CI [0.77, 0.94]. This estimate reflects consistent evidence across multiple designs and topics that individuals overestimate the social rejection they will face for changing their mind on a polarized political issue.

Study 1: Additional Study Information

Study 1 - Political Attitudes. Participants reported whether a liberal position on an issue or a conservative position on an issue best represents their own views on the issue using the following items.



Study 2: Additional Study Information

Note: Study 2 employed six individual Qualtrics surveys in order to ensure correct pairing of participants based on (a) role, (b) ingroup affiliation, and (c) topic condition. The supporting materials added to the Research Box page for this project includes only one example of such a survey ('Democrats – Guns'), which is representative of the other surveys that recruited Democrats and Republicans for each topic, respectively. These materials will be provided upon request.

Study 2: Participant instructions at the beginning of Study 2.

Welcome back!

Thank you for completing Part 1 of the study and for returning for Part 2. Part 2 will now begin.

Please read all questions and instructions carefully throughout this study.

The purpose of this study is to examine how well people perform on cooperative tasks when they are working with another person who holds similar or different political beliefs.

In this part of the study, you will be paired with another participant who either has similar or different political beliefs than you. We will use your responses from Part 1 of the study (that you completed yesterday) to randomly assign you to one of these two conditions.

The participant you are paired with will be your partner in this study. You and your partner will try to complete a brief (2 minute) cooperative teamwork task together. If you and your partner are able to complete this task, you will be awarded a $0.10 bonus that you will distribute amongst yourselves.

In this study, you will (in order):

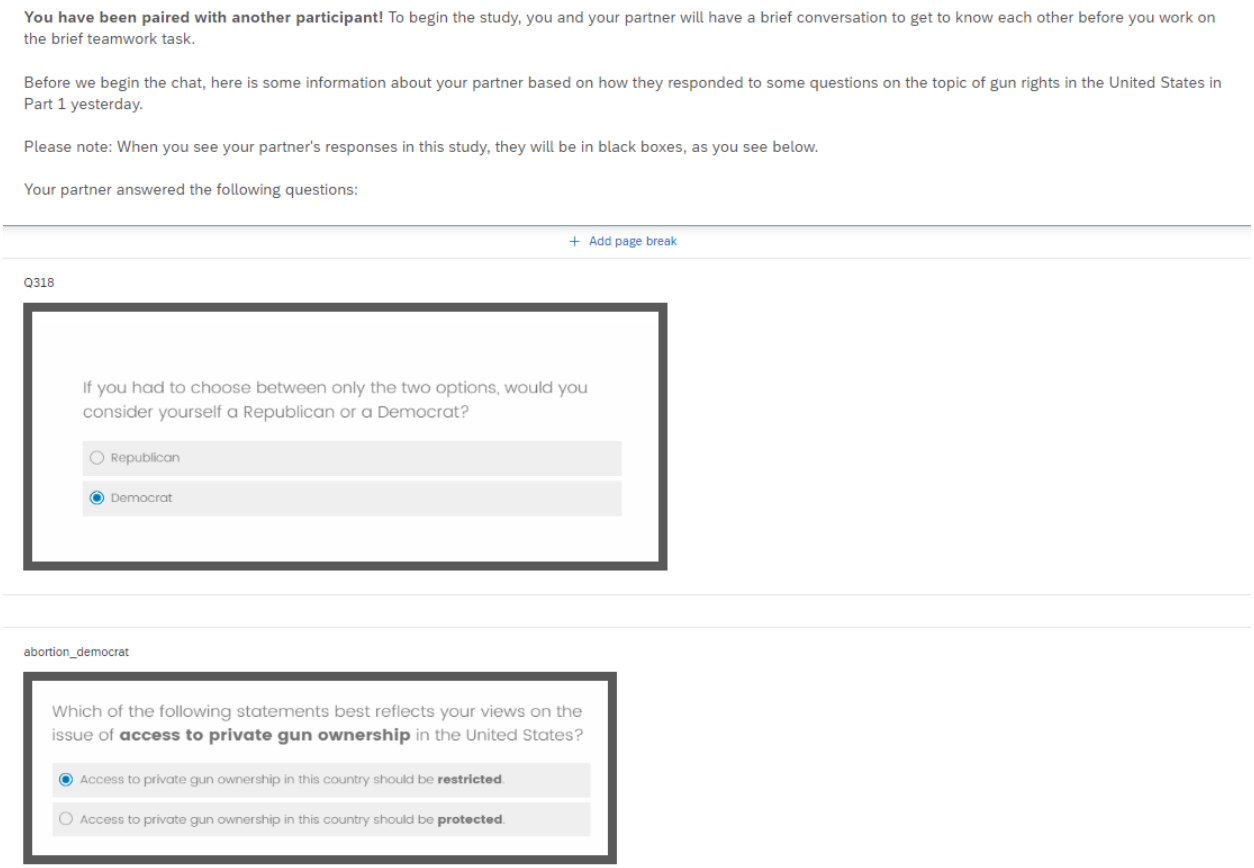
1. **Have a conversation with your partner.**
2. **Complete a brief survey (on your own).**
3. **Work together with your partner to complete a brief cooperative teamwork task.**

This part of the study should take anywhere between 10-20 minutes to complete, depending on how long it takes you to be paired with someone in the study.

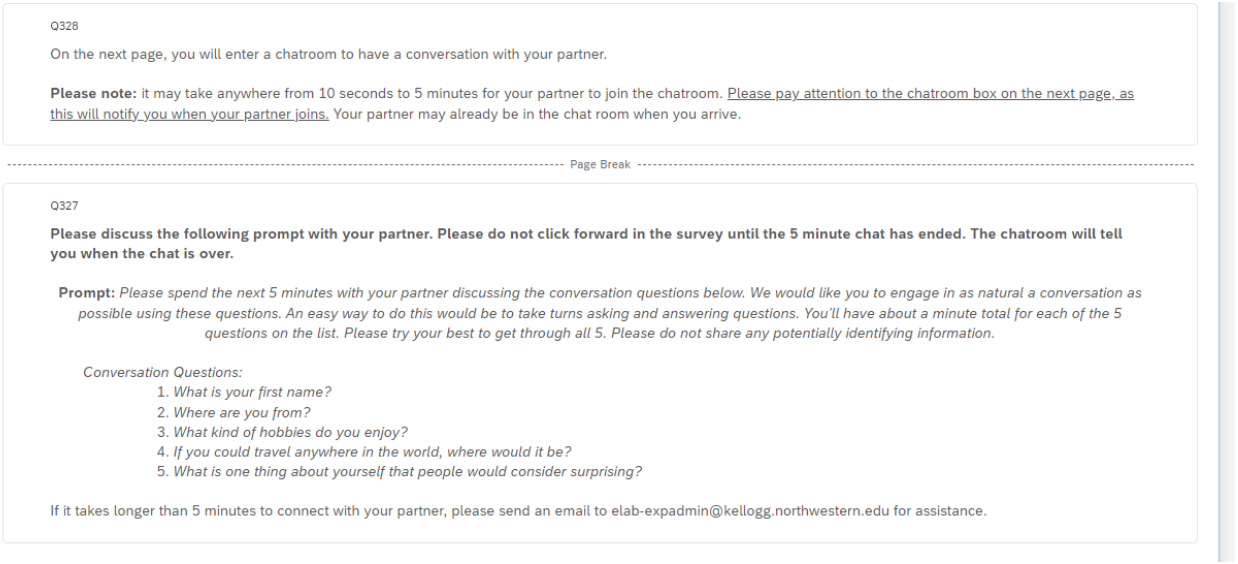
When you are ready to begin the study, please check the box below. You will be paired with your partner on the next page.

☐ I have read and understand the instructions. I will continue to read the instructions carefully throughout this study.

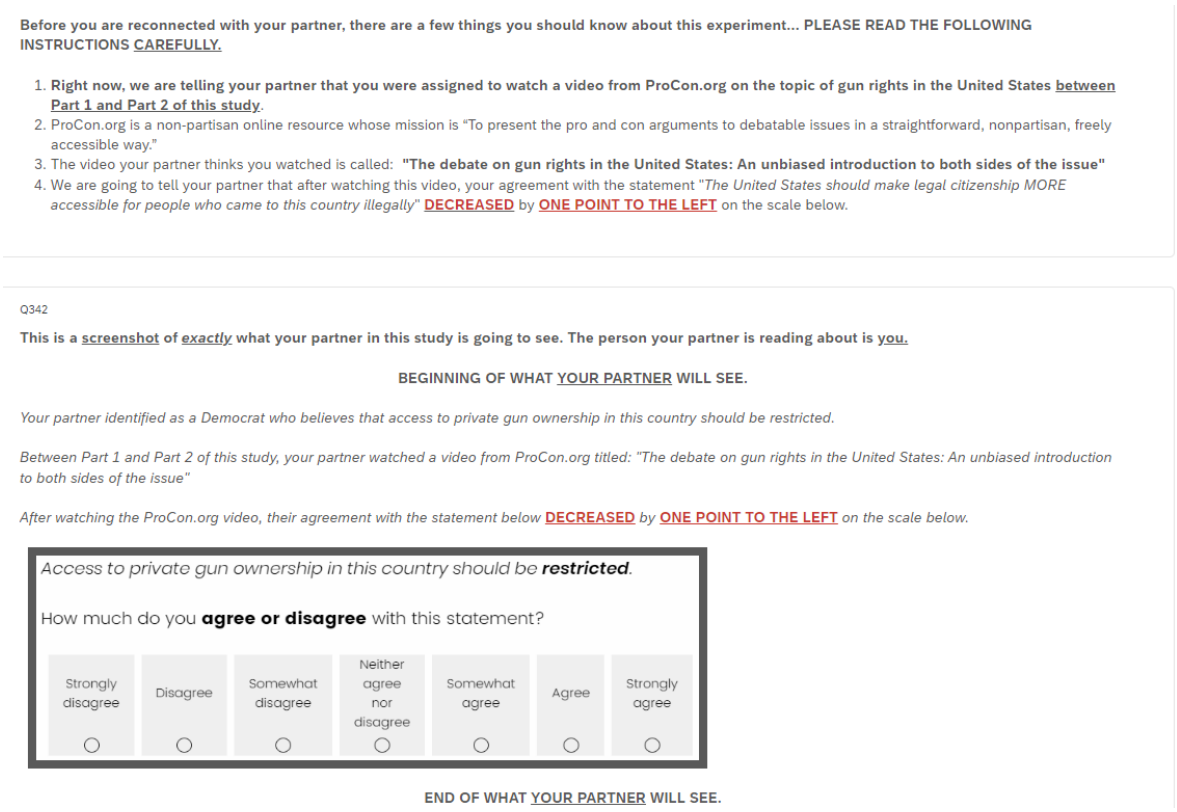
Study 2: Partner Information. Here is how all participants learned, veridically, about their partners' political views based on their Pre-screen survey responses.



Study 2: Chat Instructions. Here are the instructions that all participants received for the chat.



Study 2: Predict condition Post-Chat Manipulation. Predictors were shown the following text immediately following the chat with their partner.



Before you are reconnected with your partner, there are a few things you should know about this experiment... PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY.

1. **Right now, we are telling your partner that you were assigned to watch a video from ProCon.org on the topic of gun rights in the United States between Part 1 and Part 2 of this study**.
2. ProCon.org is a non-partisan online resource whose mission is "To present the pro and con arguments to debatable issues in a straightforward, nonpartisan, freely accessible way."
3. The video your partner thinks you watched is called: **"The debate on gun rights in the United States: An unbiased introduction to both sides of the issue"**
4. We are going to tell your partner that after watching this video, your agreement with the statement "*The United States should make legal citizenship MORE accessible for people who came to this country illegally*" **DECREASED** by **ONE POINT TO THE LEFT** on the scale below.

Q342

This is a screenshot of *exactly* what your partner in this study is going to see. The person your partner is reading about is you.

BEGINNING OF WHAT YOUR PARTNER WILL SEE.

Your partner identified as a Democrat who believes that access to private gun ownership in this country should be restricted.

Between Part 1 and Part 2 of this study, your partner watched a video from ProCon.org titled: "The debate on gun rights in the United States: An unbiased introduction to both sides of the issue"

After watching the ProCon.org video, their agreement with the statement below **DECREASED** *by* **ONE POINT TO THE LEFT** *on the scale below.*

*Access to private gun ownership in this country should be **restricted**.*

How much do you **agree or disagree** with this statement?

Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
○	○	○	○	○	○	○

END OF WHAT YOUR PARTNER WILL SEE.

Study 2: React Condition Post-Chat Manipulation. React condition participants were shown the following text immediately following the chat with their partner. Note that this is the exact information that we told predictors we would tell them.

Before you are reconnected with your partner, there are a few things you should know about this experiment... PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY.

1. **After Part 1 of this study, your partner was assigned to watch a video from ProCon.org on the topic of gun rights in the United States.**
2. ProCon.org is a non-partisan online resource whose mission is "To present the pro and con arguments to debatable issues in a straightforward, nonpartisan, freely accessible way."
3. The video from ProCon.org that your partner watched is titled: "**The debate on gun rights in the United States: An unbiased introduction to both sides of the issue"**
4. On the next page, we are going to show you how your partner responded to questions about gun rights in the United States **both before AND after watching the ProCon video.**

When you have read and understand these instructions, please check the box below to see how your partner responded to these questions before and after watching the ProCon video.

Study 2: Cooperative Task Description.

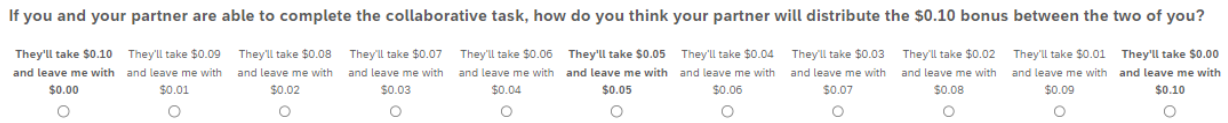
Great, we are now going to begin the portion of the study in which you and your partner will complete a collaborative task.

Please read the description of the task and respond to the questions below before the task begins.

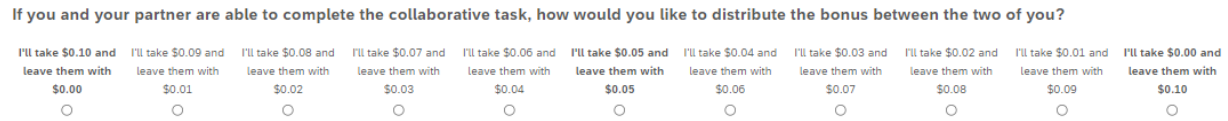
Task Description: You and your partner have been assigned to complete a collaborative maze navigation task. During this task, you and your partner will work together to navigate a virtual maze. This task requires coordination, communication, and the ability to work together to reach a common goal. You will receive specific instructions for your role on the page with the task.

Study 2: Behavioral Measures.

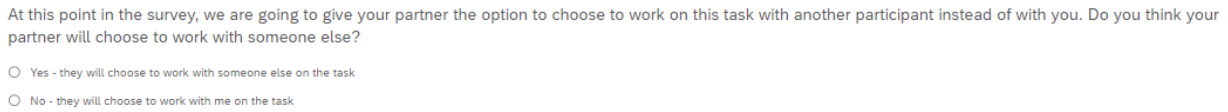
Predict: Dictator Game.



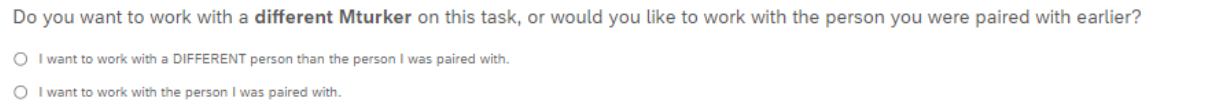
React: Dictator Game.



Predict: Partner Choice.



React: Partner Choice.



Study 2: Research Assistant Coding Instructions.

Coding for Effects on Interpersonal Consequences

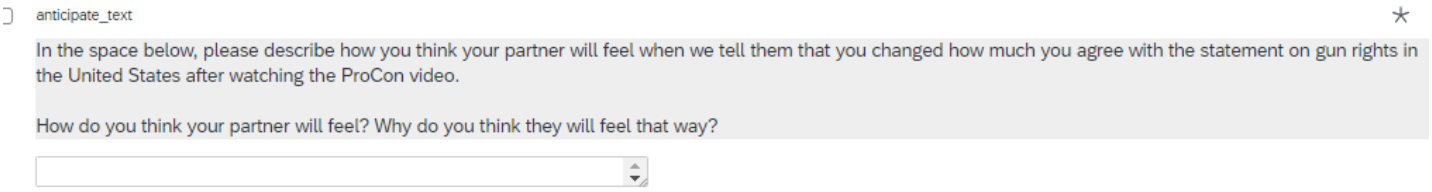
In this study, participants were asked to provide written responses to two types of questions:

1. **Predict Questions:** How they think someone else would judge and feel towards themselves.
2. **React Questions:** Their own judgement of and feelings towards someone else.

PREDICT QUESTIONS

Participants were asked to **predict** how someone else would evaluate them if this person found out that their beliefs on a political topic deviated away from the majority view of the ingroup (e.g., a Democrat who shifted to a more conservative stance on the issue of immigration). The topics were related to gun control, immigration, and abortion.

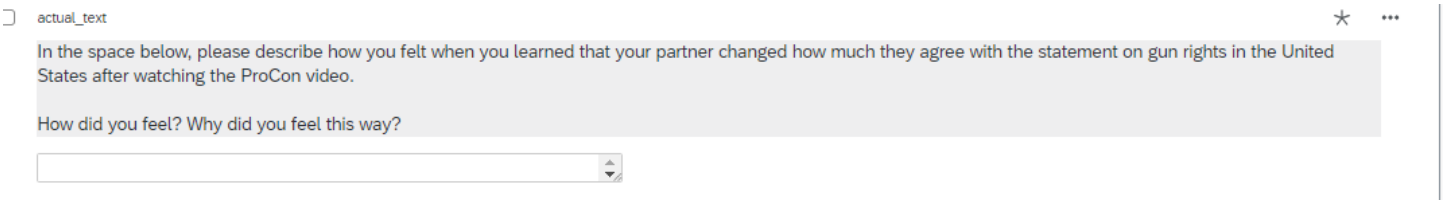
Here is the **Predict Question** they were asked. In this example the question is about gun control.



REACT QUESTION

Participants were told that someone else's beliefs on a political topic deviated from the normative view of the group (e.g., a Democrat who shifted to a more conservative stance on the issue of immigration) and were asked to evaluate them.

Here is the **React Question** they were asked. In this example the question is about gun control.



Your task is to code the responses that participants provided for both the Predict Question and React Question in response to these two questions.

We will apply the same coding scheme to the question in both Predict and React questions.

PLEASE NOTE that for react questions participants wrote about their evaluation of someone else and for predict questions participants wrote about how they think someone else will evaluate them.

For example, responses to each question might look something like…

- predict: "I think this person will feel ____ towards me."
- react: "I feel ____ towards this person."

We are going to code these responses on one dimension: **INTERPERSONAL CONSEQUENCES.**

More detail is provided on the following page.

CODING FOR INTERPERSONAL CONSEQUENCES

For responses to the **Predict Question,** we will code for the expected effect that learning that their political beliefs change will have on their partner's judgement, feelings, and evaluation of them.

For responses to the **React Question,** we will code for the actual effect that learning that their partner's political beliefs changed has on their judgement, feelings, and evaluation of their partner.

We will apply the same coding scheme to responses to both Predict and React questions, but keep in mind that predictors are making predictions and reactors are reporting their actual reactions.

For each response, you will code it as follows:

Coding Guidelines

-3 Very Negative Effect: This [has (react) / will have (predict)] a **very negative effect** on interpersonal outcomes.

-2: Negative Effect: This [has (react) / will have (predict)] a **negative effect** on interpersonal outcomes.

-1: Moderately Negative Effect: This [has (react) / will have (predict)] a **moderately negative effect** on interpersonal outcomes.

0: Neutral / No Effect: This [has (react) / will have (predict)] **NO effect** on interpersonal outcomes.

1: Moderately Positive Effect: This [has (react) / will have (predict)] a **moderately positive effect** on interpersonal outcomes.

2: Positive Effect: This [has (react) / will have (predict)] a **positive effect** on interpersonal outcomes.

3: Very Positive Effect: This [has (react) / will have (predict)] a **very positive effect** on interpersonal outcomes.

5: Ambiguous / Unclear / Did Not Answer the Question: You will mark this for any response that is ambiguous, unclear, or otherwise does not apply (e.g., if the participant did not understand the question, or wrote something that doesn't apply such as "good").

For all cases: only consider information that is about INTERPERSONAL CONSEQUENCES specifically.

For example, if someone says "*It is important to be able to learn and change your mind.* ***That being said, I know that there is not much more to learn about the abortion debate for people who are on the right side and want to preserve choice***" you would ignore the negative moral value judgement involved with the second half of the sentence and focus on the first part that discusses Interpersonal Outcomes.

Examples for each category are provided on the following page.

-3 Very Negative: Outcomes that are defined by intensely negative interpersonal outcomes. Look for qualifiers such as "very" and "extremely" (e.g., I feel **very** upset). You will know a -3 when you see one.

Predict

- *I think my partner would be disappointed in me. I think **anyone who truly believes in women's rights would know that there is a disinformation campaign going around trying to slowly take away all of our rights**.*
- *I think this person will feel **very betrayed. They will think of me as unreasonable, and disloyal to my party.***

React

- *They're falling for **lefty propaganda**. That is bad.*
- *I felt like this **is typical reaction for a woman**. I feel this way because **women generally don't have their own opinions and are easily manipulated** by emotion.*

-2 Negative: Outcomes that are unambiguously negative. Look for unqualified judgements/feelings, such as upset, sad, angry, disappointed. This is when people unambiguously state a negative interpersonal outcome.

Predict

- *Access to abortion is a very hot topic in the country right now and people have very strong opinions about it. If they think abortion rights should be protected and are told that I think they should be denied or restricted in some way**, I would expect them to be annoyed or upset with me.** I think people are very set in their ways on abortion so any change of opinion is seen as a defeat especially if its against your position.*
- *I think this person will be curious why I changed my mind**, but ultimately will be sad** that I no longer agree with this position on immigration.*

React

- *I felt surprised by how my partner changed their response that I was curious about what the video was about**. I felt unhappy for the partner to easily changed their response** after watching the clip since I thought the topic needed no debate.*
- ***I feel upset with this person**. They should not be so easily swayed on such an important issue.*

-1 Moderately Negative: Outcomes that are in between negative and neutral, but carry somewhat of a negative connotation. Look for hedge words to qualify statements, such as "kind of", "sort of", "a little", "I guess", "maybe" (e.g., "I feel **a little** upset).

Predict

- ***Maybe I am too quick with my judgments/decisions*** *if I am willing to change my mind based on one video. I think they will feel that way because the video decreased my thoughts on restricting guns.*

React

- ***I am a little disappointed*** *but would like to hear their reasons for the change. They still are on the agree with protections side of the scale, they just a little less. I'm interested in what the video was like but I know I'll never change my position on this issue.*

0 Neutral / No Effect: These are responses where the person reports no change in how this affects interpersonal outcomes. Be careful not to confuse these with examples of responses coded as 5 (unambiguous), which could seem neutral (e.g., "I don't want to make a judgement of this person") but is actually the participant not correctly answering question.

Look for words and phrases, like: "I feel the same", "I don't judge them for changing their opinion", "I trust their judgement", "I feel neutral", "I don't feel any differently".

Predict

- *I think my partner will be a little surprised* ***but doesn't get upset*** *because I'm only 1 notch to the left (somewhat agree). My partner may want to ask me questions as to why I changed my vote a little to the left instead of agree or extremely agree. I would then explain my thoughts to her, (possibly affected by what I saw on the video) but let her know I'm still voting ""agree""*
- *I think that my partner will* ***not consider this a big deal*** *- after all it is just a one point decrease. It would be a logical thing to do.*
- *I believe* ***my partner will not think much about my decisioin to change, except to wonder why****. After all, a movement of one point is not that much.*

React

- ***I feel okay with this information****. I don't really know what was in the video so it doesn't change how I feel about my partner. Also, it was mentioned that she only changed one point to the left which could be fairly insignificant of a change anyway for all I know.*
- ***I felt neutral*** *because maybe they learned something that they didn't consider before about gun rights. It show that they do use critical thinking and is willing to take in new information about things.*
- ***I think they are entitled to their own opinion****. I would want them to respect mine as well.*

1 Moderately Positive: Between neutral and positive. This will include mild feelings of positivity. Look for hedge language ("kind of", "sort of", "a little", "I guess", "maybe")

Predict

- *My partner seemed nice.* ***Majority of people can have civil conversations and still be friends even if their opinions differ****. There are too many other things we have in common.*
- *I'm not sure what they'll think of me, but I suppose they will think I'm* ***slightly open-minded***

React

- *I don't know, this doesn't really say all that much about a person. I guess if anything* ***I like that they are not too set in their ways to consider other opinions****.*
- *I* ***suppose it's a good thing*** *that they can think about things different.*
- *I think anyone can be influenced by learning things they didn't know before. It's one of the reasons why continually learning and educating yourself is a good idea. I didn't feel upset or betrayed by my partner. I just assumed his positional change was just based on him learning something new he didn't know before.*

2 Positive: Outcomes that are unambiguously positive. Look for unqualified judgements/feelings, such as happy, pleased, respectful. This is when people unambiguously state a negative outcome.

Predict

- *They will think I am* ***open-minded and willing to consider all perspectives****.*
- *I think they will appreciate and respect the fact that I am able to think beyond my previously held beliefs.*

React

- *I felt like she / he seemed open to new perspectives and ideas. it is good to change your opinion with new facts.*
- *It makes me feel that they are open minded to the issues and not an ideologue.. I respect rational thinkers no matter what side of the political spectrum they are on.*

3 Very Positive:

Predict

- *They will be respectful of my change, and I think they will actually* ***think highly of me for being open-minded.*** *They will be glad to know that members of their party are so open-minded*
- *I think my partner will* ***respect me a lot****. It shows that I have a very clear way of evaluating information.*

React

- ***I feel very supportive of my partner's decision****. It shows that they are able to think past their biases and carefully consider new information****. I respect that a lot.***
- ***I am very pleased to learn*** *that this person is so open-minded.*

5: Ambiguous / Unclear / Did Not Answer the Question: You will mark this for any response that is ambiguous, unclear, or otherwise does not answer the question (e.g., if the participant did not understand the question, refused to answer it, or wrote something short such as "good"). A frequent occurrence of a 5 involves a participant saying that they don't have enough information to make a judgement. As mentioned previously, be careful not to confuse some 5 (ambiguous/unclear) responses with 0 (neutral) responses.

- *Sometimes I feel rejected by someone*
- *It's hard to say because the person did not engage in the chat so there is no way of gauging their reactions to the questions.*
- *i'm not sure. i dont know them well fenough to make a judgement on that type of things after a short conversation*
- *Abortion is a tricky situation since sometimes it is needed for medical reasons.*
- *I am not sure they will believe that I would lower my stance on this.*

SPECIAL CASES

1. **Coding for "Respect"**

Often you will see people talk about respect. To determine whether or not these responses should be coded as 'neutral' instead of 'positive' or 'moderately positive' It is important to identify whether the respect is geared towards the person, or towards the general idea of changing one's mind. It is also important to consider the other information included in the response.

If the respect is geared towards the right to change one's mind **and there is no other information about interpersonal outcomes in the response**, then the response would be coded as neutral.

For example, these would be coded as neutral:

- "I respect anyone's right to change their mind".
- "This person has a right to change their mind. I respect that"

If the respect is geared towards the right to change one's mind **and there IS other information about interpersonal outcomes in the response**, then the response would be coded based on the other information that is present.

For example:

- **-1 Moderately Negative:** "I feel a bit upset with this person, but I respect anyone's right to change their mind"

- **2 Positive:** “I am happy for this person. They have a right to change their mind, and it shows they thought carefully about the other side”.

If the respect is geared towards the **person** (i.e., an interpersonal outcome), it will likely be coded as very positive, positive, or moderately positive. Be sure to evaluate other information that might be present as well.

For example:

- **2 Positive:** “I think they will respect me”
- **1 Moderately Positive:** “I am surprised, but I kind of respect this person for being open-minded.”

2. **Coding for curiosity and surprise**

You will often see people list that they are curious or surprised. We will not ascribe value to either of those reactions in isolation. Instead, we’ll make our judgement based on the other information present.

For example:

- **2 Negative:** “I think they will be surprised. They will be upset that I changed my mind”
- **0 Neutral:** “I think this person will be surprised” AND “They will be curious why I changed my mind”
- **2 Positive:** “They will be curious why my mind changed, but I think they will respect me for being open-minded”

3. **When PREDICT questions evaluate the other person, or the REACT questions assume how the person will feel about THEM, ignore this information.**

For example, you would **ignore the bolded text** and use the unbolded text to score the following **PREDICT** questions:

- My partner will understand that watching the video could have an impact on my beliefs about gun control. **She was a mother, so I think she is pretty open minded when it comes to people having their own opinions. In the chat, she seemed pretty friendly, open. I didn't get the impression that she is going to reject somebody that has differences of opinions.**
- I think my partner will hate me. **They seemed very friendly and kind.**

For example, you would **ignore the bolded text** and use the unbolded text to score the following **REACT** questions:

- I don’t trust this person. **I think they will assume that I am a nice guy**, but this is not okay with me.
- **This person may think that I would be judgmental of them**, but I actually respect their decision.

Study 3: Additional Study Information

Study 3 Phase 1: Counter-attitudinal writing task instructions. These are the instructions that participants in the Predict condition received. **Note:** Participants in Sample A were randomly assigned to consider either abortion, gun control, or immigration as the focal topic (following the same procedure employed in Studies 1, 3, 5); however, participants in Sample C only considered the topic of abortion access. This decision was made because predictors in Sample A reported dissenting belief change more frequently for abortion access than for the other two topics and we were trying to recruit the greatest number of participants who reported dissenting belief change as possible in Sample C.

Sample A Predict Condition Writing-Task Instructions.

You indicated that the following statement best represents your views on the issue of ${e://Field/issue}.

${e://Field/topic}

In this part of the study, regardless of your personal beliefs, please craft a persuasive argument in support of the **opposite** view on this topic.

Specifically, please write a persuasive essay in favor of the following stance: *${e://Field/opposite}*.

Your aim is to write a convincing essay on the benefits of having these laws for the benefit and well-being of the country. Your goal is to present this stance as persuasively as possible, even if you personally disagree.

You will be able to reference a list of arguments generated by ChatGPT in favor of this view to help you craft this message, which will appear on the next survey page.

Q332

Chance to Win a Bonus: Your essay will be shown to a future sample of ${q://QID19/ChoiceGroup/UnselectedChoices} participants on Mturk who will evaluate how persuasive your essay is. If your essay is rated in the top 10% of persuasive essays out of all essays from participants taking this study, you will be awarded a $5.00 bonus.

All responses will be kept completely anonymous and will be stripped of any identifying information prior to use in the future study.

NOTE: Please do not leave this webpage for the duration of the task. You must craft this message on your own without copy/pasting from the internet or the use generative AI to create your text for you. We will be able to tell if you leave this webpage, and you will be excluded from the bonus pool if you leave this webpage during this task.

Q333

Before you begin, please answer the following attention checks to make sure that you understand these instructions. Your ten minutes to write this message will begin once you've answered these two questions.

Sample C Predict Condition Writing-Task Instructions.

You indicated that the following statement best represents your views on the issue of abortion.

"${q://QID197/ChoiceGroup/SelectedChoices}"

In this part of the study, regardless of your personal beliefs, please craft a persuasive argument in support of the **opposite** view on this topic.

Specifically, please write a persuasive essay in favor of the following stance: *"${q://QID197/ChoiceGroup/UnselectedChoices}"*.

Your goal is to present this stance as persuasively as possible, even if you personally disagree. If your essay is rated in the top 10% of most persuasive essays, you will win a $5.00 bonus.

You will be able to reference a list of arguments generated by ChatGPT in favor of this view to help you craft this message, which will appear on the next survey page.

Q332

NOTE: Do not leave this webpage for the duration of the task. You must craft this message on your own without copy/pasting from the internet or the use generative AI to create your text for you. If you leave the webpage or copy/paste into the text box, you will NOT be paid for your participation in this study.

Example ChatGPT Output. This prompt was given to ChatGPT and the actual response was pasted into a Qualtrics survey question.

Q365

Here is the output from ChatGPT that you may reference as you craft your persuasive essay. This answer was in response to the prompt below. You do not need to reference this, but it is here in case it is helpful.

Prompt: I am currently participating in a research study and I have been asked to write a persuasive essay in favor of the following stance: ${q://QID197/ChoiceGroup/UnselectedChoices}. Please come up with a list of five persuasive arguments in favor of this position.

cons_abortion

Display this question

If Which of the following statements best reflects your views on the issue of abortion access in the... Abortion access in this country should be protected Is Selected

ChatGPT Response:

Absolutely, here are five points you might use to build a persuasive argument in favor of restricting abortion access:

1. Moral and Ethical Considerations. Many people, on the basis of their moral, philosophical, or religious beliefs, find abortion ethically objectionable and argue for restrictions to uphold societal moral standards.

2. Health and Psychological Concerns. Critics argue that abortion carries health risks for the individual undergoing the procedure, including complications such as infection and hemorrhage, as well as mental health outcomes such as depression.

3. Alternative Solutions. Advocates for restricting abortion access often promote adoption as a viable alternative, providing a way for individuals to relinquish parental responsibilities while still bringing the pregnancy to term.

4. Legal and Policy Considerations - State Rights. Some advocate for individual states to have greater autonomy in determining their own abortion policies, reflecting the diverse viewpoints of their respective populations.

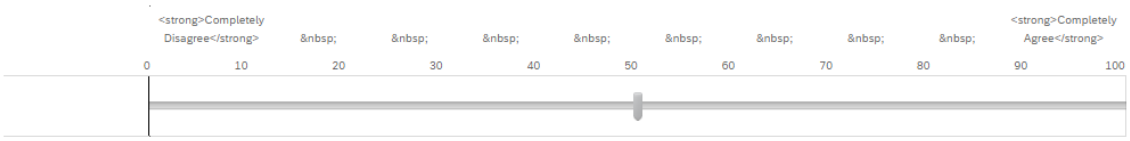
5. Population Dynamics. From a demographic perspective, some argue that restricting abortion could help stabilize population dynamics by potentially increasing birth rates.

Study 3 Phase 1: Time 2 Attitude Measure.

In our research, we find that it is common for people to update their views after completing this type of exercise. In case your views changed at all, we would like to ask you again:

Please indicate the extent to which you agree with the statement below. A score of 100 means that you completely agree with this statement. A score of 0 means that you completely disagree.

Statement: *${q://QID197/ChoiceGroup/SelectedChoices}*



Study 3 Phase 1: Sample C - Information Given to React Condition Participants.

In this part of the study, you will be asked to evaluate another ${q://QID19/ChoiceGroup/SelectedChoices} participant from a previous study. We will tell you some information about this person before you are asked to evaluate them.

Please read the instructions on the next page carefully. There will be attention checks about these instructions.

Page Break

INFO DRAFT 2

The participant you are going to evaluate is a ${q://QID19/ChoiceGroup/SelectedChoices} who participated in one of our recent studies.

This person reported that they agree with the following statement about abortion, "*${q://QID197/ChoiceGroup/SelectedChoices}*".

During the study, this person was instructed to write a persuasive message in favor of the **opposing** political view.

Here are the exact instructions that this person was given, shown in blue:

<beginning of the instructions this person saw>

You indicated that the following statement best represents your views on the issue of abortion.

"${q://QID197/ChoiceGroup/SelectedChoices}"

In this part of the study, regardless of your personal beliefs, please craft a persuasive argument in support of the **opposite** view on this topic.

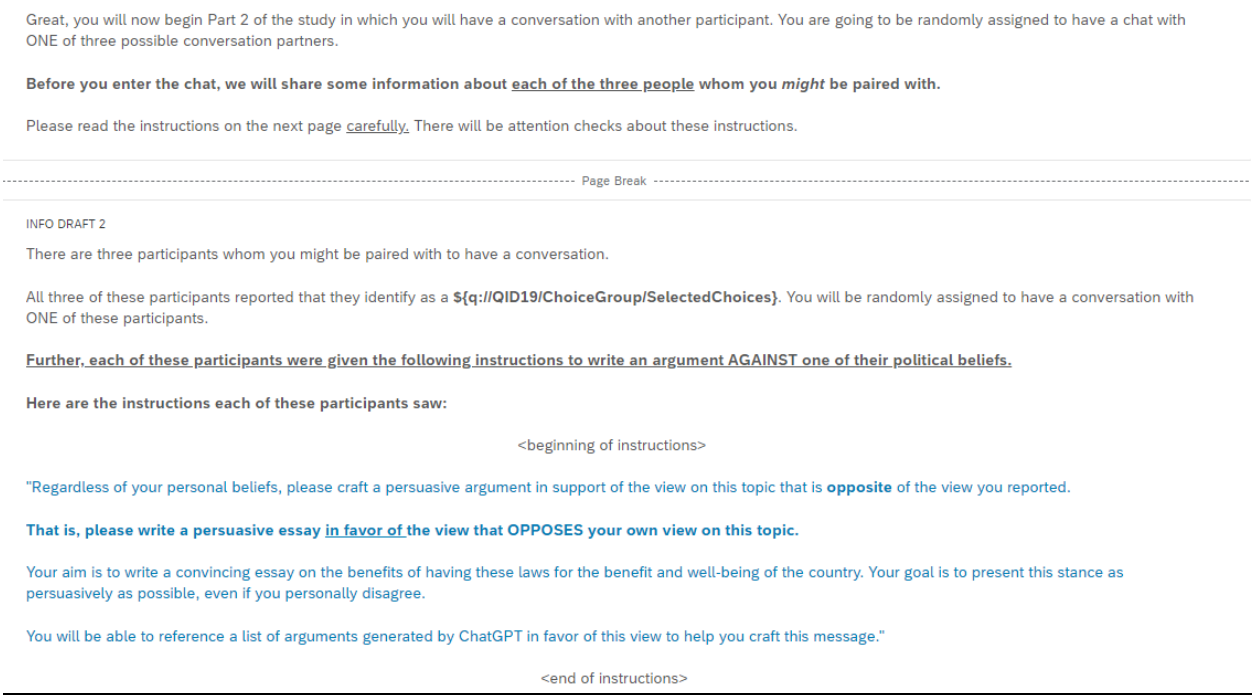
Specifically, please write a persuasive essay in favor of the following stance: "*${q://QID197/ChoiceGroup/UnselectedChoices}*".

Your goal is to present this stance as persuasively as possible, even if you personally disagree. If your essay is rated in the top 10% of most persuasive essays, you will win a $5.00 bonus.

You will be able to reference a list of arguments generated by ChatGPT in favor of this view to help you craft this message, which will appear on the next survey page.

<end of instructions this person saw>

Study 3 Phase 1: Sample B - Information Given to React Condition Participants.



Great, you will now begin Part 2 of the study in which you will have a conversation with another participant. You are going to be randomly assigned to have a chat with ONE of three possible conversation partners.

Before you enter the chat, we will share some information about each of the three people whom you *might* be paired with.

Please read the instructions on the next page carefully. There will be attention checks about these instructions.

Page Break

INFO DRAFT 2

There are three participants whom you might be paired with to have a conversation.

All three of these participants reported that they identify as a **${q://QID19/ChoiceGroup/SelectedChoices}**. You will be randomly assigned to have a conversation with ONE of these participants.

Further, each of these participants were given the following instructions to write an argument AGAINST one of their political beliefs.

Here are the instructions each of these participants saw:

<beginning of instructions>

"Regardless of your personal beliefs, please craft a persuasive argument in support of the view on this topic that is **opposite** of the view you reported.

That is, please write a persuasive essay in favor of the view that OPPOSES your own view on this topic.

Your aim is to write a convincing essay on the benefits of having these laws for the benefit and well-being of the country. Your goal is to present this stance as persuasively as possible, even if you personally disagree.

You will be able to reference a list of arguments generated by ChatGPT in favor of this view to help you craft this message."

<end of instructions>

Study 3: Key Differences Between Samples.

Sample A (predictors) versus Sample C (predictors):

1. Sample A was recruited from Mturk and Sample C was recruited from Prolific.
2. Sample A participants were randomly assigned to write an essay in favor of an opposing view on a randomly assigned political topic (abortion, gun control, immigration) upon which they reported a view that was consistent with the majority view of their party.
3. Sample C included additional exploratory items (economic punishment, self-censorship). In the main text, we report only the key dependent measure that was consistent across samples (the social sanctions composite item). We report the results for the additional items in the "Additional Analyses" section for Study 3, though we note again here that we realized after data collection was complete that both of these variables were flawed in a way that compromised their interpretation.

Sample B (reactors) versus Sample C (reactors):

1. Sample B was recruited from Mturk whereas Sample C was recruited from Prolific.
2. Sample B reactors were asked to evaluate three dissenting ingroup members one at a time in succession, whereas Sample C reactors were asked to evaluate only one dissenting ingroup member. We did not observe order effects in Sample B.

Study 3 Phase 2: Salient vs. Non-Salient condition manipulation text.

Salient.

You participated in one of our research studies on November 30th. At the beginning of that study, you indicated that you support the following view on the topic of abortion access in the United States: "*${e://Field/topic}*".

Later on in that study, you were asked to write a persuasive message on the topic of abortion access in the United States in **favor** of the following **opposing** political stance: "***${e://Field/opposite}***".

After you finished writing a message in favor of the opposite view on this topic, you reported that your agreement with the statement, "${e://Field/topic}" **DECREASED**. That is, your agreement with the majority position of the **${e://Field/party_id}** party on the topic of abortion access in the United States **decreased** after writing a persuasive message in favor of the majority position of the ${e://Field/party_id_opposite} party.

Non-Salient.

Thank you for participating in this study. You have been invited to participate in this study because earlier this fall you participated in one of our research studies and we would like to ask you some follow up questions.

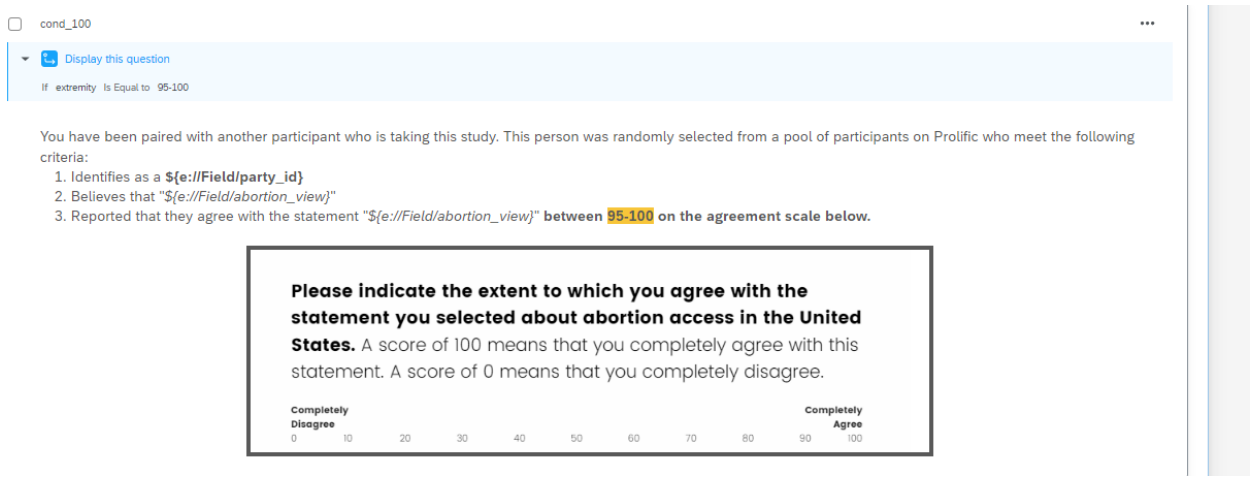
Timeline for Study 3 Phase 2.

Phase 1 and Phase 2 of Sample A (Sept 12, 2023) and Sample C were collected approximately two months apart (Nov 20, 2023) due to bandwidth and resource constraints. We thus conducted Phase 2 for Sample A (Nov 17, 2023) and Sample C (Jan 20, 2024) approximately two months following each of the respective Phase 1 studies in order to ensure that the delay period for Phase 2 was similar for both samples of participants.

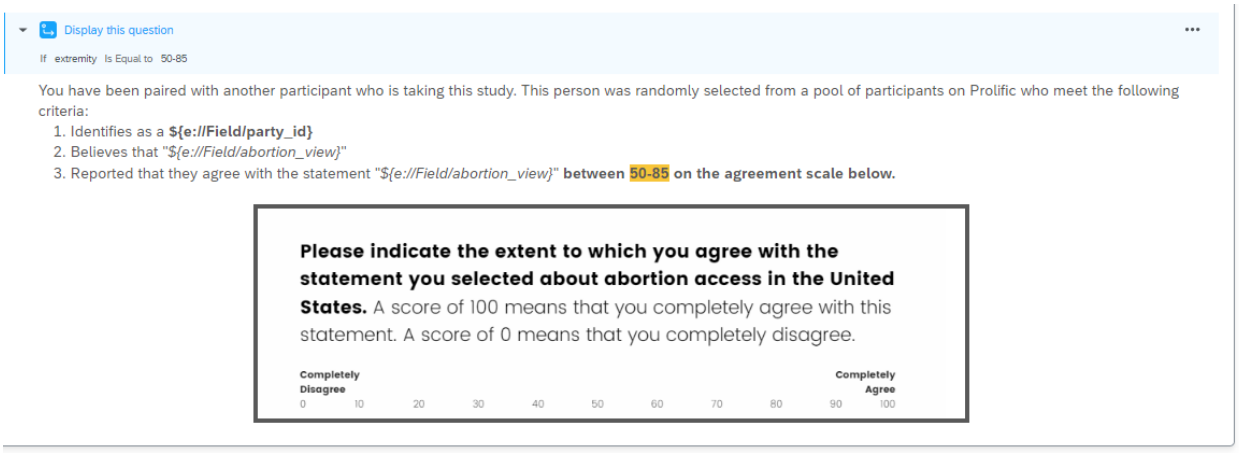
Supplemental Study 1: Additional Study Information

Predict Condition Manipulation. Here is what participants in the predict condition saw when they were informed that their partner was either an "extreme" partisan or a "moderate" partisan.

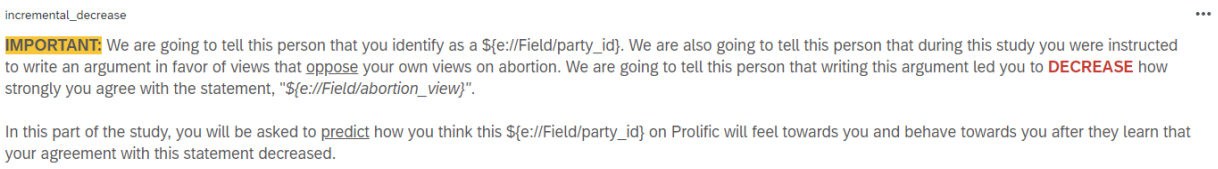
Extreme. Piped-text showed (a) the same party identity, and (b) the same party-aligned attitude on abortion as the participant.



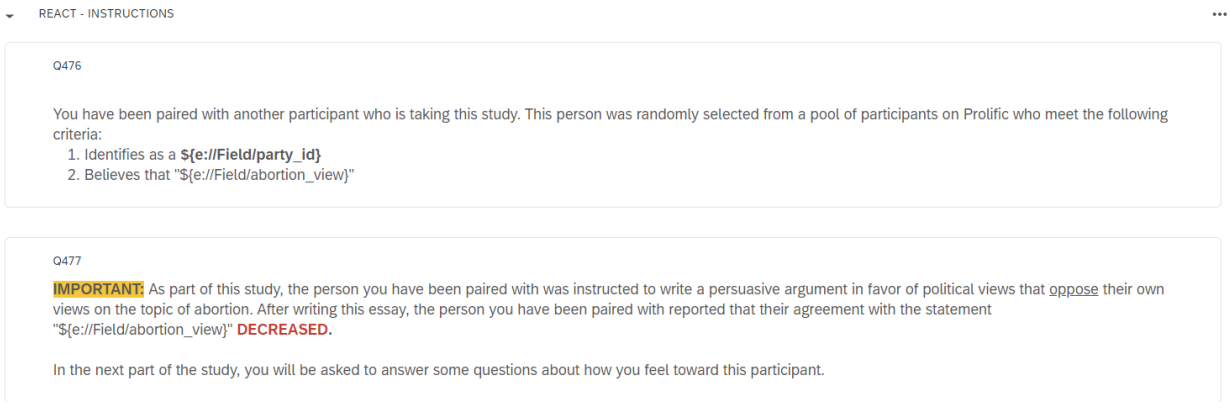
Moderate. Piped-text showed (a) the same party identity, and (b) the same party-aligned attitude on abortion as the participant.



Here is how predict condition participants were informed that their ingroup partner would learn that their agreement with the typical view of the party decreased.



React Condition Instructions. Here is what react condition participants saw.



Supplemental Study 2: Additional Study Information

Third-Party Condition Instructions and Measures.

Please read the following text carefully.

Another **${q://QID19/ChoiceGroup/SelectedChoices}** participant taking this survey named Sam indicated that they believe that ${e://Field/topic}.

In the next part of the survey, we are going to ask you questions about how you think **another ${q://QID19/ChoiceGroup/SelectedChoices} on Prolific** would react if they found out that Sam changed their mind on this issue and now believes that ${e://Field/opposite}.

+ Add page break

Q238

This question lets you record and manage how long a participant spends on this page. This question will not be displayed to the participant.

Page Break

distant_other

If another ${q://QID19/ChoiceGroup/SelectedChoices} on Prolific found out that Sam changed their mind on the issue of ${e://Field/issue} and now believes that ${e://Field/opposite}, to what extent do you think this ${q://QID19/ChoiceGroup/SelectedChoices} on Prolific would...

	1 Not very much	2	3	4 a moderate amount	5	6	7 very much
exclude Sam?	○	○	○	○	○	○	○
ignore Sam's input?	○	○	○	○	○	○	○
reject Sam?	○	○	○	○	○	○	○
disrespect Sam?	○	○	○	○	○	○	○
criticize Sam?	○	○	○	○	○	○	○

References

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Morewedge, C. K., & Giblin, C. E. (2015). Explanations of the endowment effect: an integrative review. *Trends in Cognitive Sciences*, *19*(6), 339–348.

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213800.