

Overestimating the Social Costs of Political Belief Change

[Trevor Spelman](#)

Northwestern University and IPR

[Abdo Elnakouri](#)

University of Houston

[Nour Kteily](#)

Northwestern University

[Eli Finkel](#)

Northwestern University and IPR

Version: June 23, 2025

DRAFT

Please do not quote or distribute without permission.

Abstract

How do U.S. partisans expect members of their political ingroup to react when they diverge from the typical view of their party on a partisan issue (e.g., a Democrat adopting a more conservative stance on private gun ownership)? How accurate are these expectations, and how do they influence whether people choose to speak up or stay silent? Five main studies and five supplemental studies ($N = 4,535$) employing diverse research methods—including surveys, behavioral outcomes, live participant interactions, and coded open-ended responses—revealed that partisans consistently overestimate the social sanctions they will face for changing their minds (average weighted effect size (d) of .87). These inflated expectations, which are associated with a greater likelihood of self-censoring dissenting views, appear to stem from a concern that dissent will signal greater group disloyalty than it actually does. Indeed, a brief intervention prompting individuals to reflect on their past loyalty to the group reduced this concern and produced more accurate expectations about ingroup reactions to their dissenting belief change. By examining the social forces that suppress dissent within political groups, this work offers insight into how to reduce conformity pressures and promote more open political discourse.

Statement of Limitations

Despite the robustness and consistency of our findings across studies, we note several limitations. First, our research was conducted within the specific political and cultural context of the United States during a period of heightened political polarization, which may limit the generalizability of our results to other cultural or political settings. Second, our studies primarily involved exchanges between participants who had limited prior acquaintance or were strangers, which allowed us to isolate certain psychological mechanisms but may not fully capture the complexities of interactions within close relationships. Third, our research focused on dyadic exchanges, which may not reflect the dynamics of larger group interactions, particularly in online environments. Finally, although our findings demonstrate that a loyalty affirmation intervention can reduce concerns about appearing disloyal and lead to significantly more accurate social judgments about rejection, this intervention did not fully eliminate the perception gap. Future research should explore ways to strengthen loyalty affirmation interventions and examine other factors that may contribute to this perception gap.

Introduction

In an era marked by deep political divides, it might seem unlikely that American partisans would update their views on polarizing political issues. Despite the well-documented tendency to perceive reality according to one's political allegiances (Van Bavel & Pereira, 2018), individuals do sometimes update their beliefs—even when this means departing from the prevailing orthodoxy of their political party (e.g., Kossowska et al., 2023; Tappin et al., 2023; Xu & Petty, 2022). Such shifts can carry social costs: A committed Republican who adopts a less conservative stance on abortion, for example, may face harsh social backlash from other Republicans (e.g., Jetten & Hornsey, 2014; Kruglanski & Webster, 1991; Schachter, 1951). This Republican confronts a dilemma: They can disclose their evolving views—risking social sanction—or conceal the shift to preserve their image as a loyal group member at the expense of authenticity.

If the committed Republican truly risks social rejection by disclosing their dissenting belief change¹, concealing this shift may be a wise choice. But what if assessments of this risk are systematically miscalibrated? When the risk of rejection is salient, dissenters—sensitive to the possibility that their dissenting belief change will signal disloyalty to the group—may overestimate how negatively other group members will respond. With the threat of social rejection looming large, group members might think twice before voicing their newfound dissenting views. Such self-censorship may stifle open political discourse, reinforce pluralistic ignorance, and distort the representation of ideas in the public sphere. If U.S. partisans consistently self-censor dissenting opinions, the information environment will falsely reflect

¹ We use the term “dissenting belief change” to characterize the phenomenon in which an individual's beliefs regarding a divisive political issue become less aligned with the typical view of their political party to any extent. We refer to individuals who experience such changes as “dissenters.”

homogenous opinions when in fact a wider plurality exists (Allport, 1924; Miller 2023; Noelle-Neumann, 1993).

Why might dissenters overestimate how much other group members will reject them for their dissent? We theorize that this tendency is grounded in humans' fundamental need to belong and an aversion to social rejection shaped by deep-rooted evolutionary pressures.

The Need to Belong and Fear of Social Rejection

The need to belong is a fundamental feature of human psychology (Baumeister & Leary, 1995). People largely define themselves by the groups they affiliate with, and group membership provides many benefits—including a sense of self-worth and access to various forms of material and social capital (Correll & Park, 2005; Gaertner et al., 2015; Hogg, 2016; Kruglanski et al., 2006; Swann et al., 2012; Van Bavel & Packer, 2021). Given the central role of group membership in human social life, people are highly averse to the threat of social rejection. To preempt exclusion, people often internalize and adhere to group norms—ranging from table manners to religious beliefs to political ideology.

This heightened sensitivity to social rejection is likely to be especially active in contemporary political contexts, where beliefs often serve as identity-relevant markers of group membership (Bakker et al., 2020; Connors, 2020; Dovidio et al., 2017; Funkhouser, 2022; Gaertner et al., 2015; Haidt, 2012; Swann et al., 2012; Van Bavel & Pereira, 2018). In such contexts, ideological conformity is often rewarded, whereas dissent may be interpreted as a signal of disloyalty—a reputational threat that can provoke social sanction from fellow group members.

Indeed, a growing body of research suggests that these intragroup dynamics present partisans with strong social pressures to conform to their group's ideological norms. American partisans reward group members who selectively attend to information that supports their own party's views (Moore et al., 2023) and penalize group members who are receptive to opposing views (Hussein & Wheeler, 2024; c.f., Heltzel & Laurin, 2021)—especially when such receptiveness signals dissent (Heltzel & Laurin, 2021, Study 3). Taken together, these findings suggest that group members who dissent may be prudent to fear harsh social reactions to dissenting belief change. But what if these expectations are miscalibrated?

Given the high social costs of rejection, individuals may not merely fear negative evaluations for dissenting belief change—they may systematically overestimate how negatively fellow group members will react. This tendency is likely rooted in our evolutionary past. Throughout human history, the consequences of social rejection posed serious threats to survival (see Williams, 2007, for review). In early hunter-gatherer societies, the excommunicated were likely to die if they could not repair damaged social bonds or form new ones with another group (Baumeister & Leary, 1995; Henrich, 2015). Given the severe costs of social exclusion, many scholars have argued that humans have evolved a “better safe than sorry” orientation to risks of social rejection (Baumeister & Tice, 1990; Haselton & Buss, 2000; MacDonald & Leary, 2005). Consequently, we suggest, in contexts where dissent could plausibly signal disloyalty to the group, individuals may adopt a conservative estimation strategy—erring on the side of caution by overestimating the severity of social sanctions in order to minimize the risk of exclusion. In polarized political environments, where belief change may be perceived as a signal of disloyalty, such hypervigilance to reputational threat may be especially acute.

Loyalty as a Core Reputational Concern

Loyalty—a psychologically and morally motivated commitment to act on behalf of a person, group, or organization (Berry et al., 2021; Hildreth et al., 2016; Zdaniuk & Levine, 2001)—is a foundational virtue in coalitional settings (Goodwin et al., 2014) and has been described as the “social glue” that holds groups together (Van Vugt & Hart, 2004). In competitive intergroup contexts, where coalitions depend on mutual trust and coordinated action, loyalty serves as a critical cue for determining who is dependable and who might defect (Tooby & Cosmides, 2010). People care deeply about being seen as loyal because loyalty promotes access to social support, status, and opportunities within the group (Berry et al., 2021; Goodwin et al., 2014; Kunst et al., 2019; Shaw et al., 2017; Tajfel & Turner, 1979). As a result, group members are attuned not only to the loyalty of others but to how their own actions may signal loyalty—or disloyalty—to others. This reputational sensitivity may be especially pronounced in political groups, where ideological alignment often functions as a shorthand for group commitment and loyalty (Van Bavel & Pereira, 2018).

Given group members’ strong aversion to social rejection, this better-safe-than-sorry orientation may result in a heightened sensitivity to the signals their behavior sends about group commitment. In the case of dissenting belief change, individuals may not only worry about backlash—they may specifically fear that their dissent will be seen as a signal of disloyalty or betrayal. Because loyalty is a reputationally fragile quality—easily questioned, difficult to affirm (Everett et al., 2018; McManus et al., 2020)—dissenters may assume that even minor deviations from group orthodoxy will cast doubt on their allegiance. As a result, they may overestimate how much their belief change will be interpreted as disloyal and, consequently, how harshly others will respond. This pattern reflects a broader psychological distortion in socially evaluative

contexts: the belief that one's actions will be seen as more revealing than observers actually perceive.

Research on signal amplification bias has documented the tendency for individuals who are concerned about social rejection to overestimate how strongly their behavior will shape others' impressions. This bias reflects a metaperceptual error in which individuals believe their actions will be seen as more revealing or diagnostic than observers actually perceive them to be (Vorauer et al., 2003; Savitsky, Epley, & Gilovich, 2001; see also research on the spotlight effect, Gilovich et al., 2000). Research has shown that in situations where individuals feel interpersonally vulnerable—such as initiating relationships (Vorauer et al., 2003) or navigating intergroup boundaries (Vorauer & Sakamoto, 2006)—they assume their behavior will convey significantly more diagnostic information than observers actually perceive.

We argue that dissenting political belief change constitutes a similarly fraught context. In political groups, where beliefs often function as signals of loyalty (Van Bavel & Pereira, 2018), partisans may worry that changing their view on a politicized issue will be interpreted as disloyal. Drawing on research on signal amplification bias (Vorauer et al., 2003), we hypothesize that individuals—sensitive to the steep costs of social rejection—will overestimate how negatively others will react to their dissent, in part because they believe their belief change will send a stronger signal of disloyalty than it actually does.

Hypotheses and Research Overview

Based on the preceding analysis, we advance four central hypotheses in the present research:

H1: Partisans will systematically overestimate how much other ingroup members will reject them for dissenting belief change.

H2: Anticipated social rejection for dissenting belief change will be positively associated with self-censorship: The more rejection individuals expect, the less likely they are to disclose their change in beliefs.

H3: The overestimation of rejection for dissenting belief change is partially explained by a form of signal amplification bias wherein dissenters expect that they will be seen as more disloyal for their dissent than ingroup observers actually perceive.

H4: Interventions that mitigate the signal amplification bias of loyalty judgments will, in turn, reduce the overestimation of social rejection for dissenting belief change.

We tested these hypotheses across five main studies ($N = 3,063$; four pre-registered) and five supplemental studies ($N = 1,472$; two pre-registered), using behavioral outcomes, paired dyads, and live interactions. First, we examined how expectations about ingroup member reactions to dissent influence decisions to disclose versus conceal dissenting political belief change (Pilot Study 1, Supplemental Study 1, Study 4a). Next, we tested whether these expectations align with how ingroup members actually react (Studies 1-5). Finally, we examined how signal amplification bias—specifically, the tendency to expect that dissenting belief change will send a stronger signal of disloyalty than observers actually perceive—contributes to the overestimation of social rejection for dissenting belief change (Studies 4-5). We also tested whether affirming dissenters' sense of group loyalty can attenuate this bias in loyalty meta-judgments and, in turn, reduce the overestimation of social sanctions (Study 5).

Transparency and Openness

The design, hypotheses, and analysis plan for studies 2, 3, 4c, 5, and supplementary studies 1 and 2 were pre-registered². Any deviations from the pre-registered document are reported in Table S1 in the Online Supplementary Materials (OSM), following guidelines by Willroth & Atherton, (2024). To facilitate accessibility, we provide access to all experimental measures, procedures, statistical analyses, data, and pre-registrations at the same link. Our studies used varied samples that included participants from the online participant recruitment platforms Amazon Mechanical Turk (via CloudResearch) and Prolific. For all studies, we report how we determined our target sample size, all data exclusions, all manipulations, and all measures in the methods sections. This research was approved by the [blinded for peer-review] University Ethics Board. ChatGPT was used to edit code for data analysis, revise select sentences in the manuscript, and conduct basic literature reviews in the preparation of this manuscript (OpenAI, 2024).

Pilot Study 1

Although the relationship between fear of social rejection and self-censorship of minority opinion has been well established in previous work (Glynn et al., 1997; Matthes et al., 2017; Moy et al., 2001; Noelle-Neumann, 1993; Scheufele et al., 2001), this relationship has not been examined for dissenting belief change. To examine whether this relationship is present in this context, we conducted a pilot study in which we surveyed U.S. partisans who reported dissenting belief change on a political topic in the past 12 months ($N = 131$) and asked them (a) how they expect ingroup members to react when learning about their dissenting belief change, and (b) whether they have become more or less likely to self-censor their views on this topic in the time

² Preregistrations can be accessed here: https://researchbox.org/3336&PEER_REVIEW_passcode=NBUYVQ
Password: NBUYVQ

since their belief change occurred. This design represents an ecologically valid context wherein partisans who have experienced dissenting belief change reported their real-world behaviors and expectations when interacting with other group members in their daily lives³.

Extending prior research on self-censorship to the context of belief change, we found that participants who anticipated harsher social sanctions for their dissenting belief change ($M = 3.61$, $SD = 1.63$) also reported greater likelihood of self-censoring their dissenting views ($M = 3.88$, $SD = 1.47$), Spearman's $r = -.22$, $p = .014$. This link between expected social sanction and self-censorship highlights the stakes of potentially miscalibrated expectations about social rejection for dissent and, by implication, the accurate representation of perspectives in the public sphere. Of note, we believe this study represents a conservative test of the relationship between anticipated rejection and self-censorship because people may not fully recognize or report the extent to which fear of social sanctions drives their decision to self-censor. As such, any observed relationship between these variables may underestimate the true influence of anticipated rejection on self-silencing behavior.

Study 1

Given the relationship between anticipated social sanctions and self-censorship of dissenting belief change, an important question is whether individual's expectations of social rejection from ingroup members are well calibrated. In Study 1, we investigated this question by comparing predicted and actual social rejection for dissenting belief change on a political topic.

Methods

³ See Pilot Study 1 in the OSM for a full description of the procedure, measures, analyses, and results from this study.

Participants

We conducted an a priori power analysis using G*Power 3.1 (Faul et al., 2009), which suggested a sample size of at least 420 participants to detect a small-to-moderate effect ($f = 0.2$) in a 2×3 between-subjects ANOVA with 90% power at $\alpha = .05$. Given the novelty of this research area, and our aim to examine findings across multiple political topics, we decided to over-recruit to enhance the precision and robustness of the findings. We recruited 519 participants from Amazon's Mechanical Turk using custom recruitment filters on CloudResearch to achieve a sample with equal numbers of Democrat and Republican participants. We excluded thirteen participants who did not complete the study, one participant who failed an attention check, and five participants who reported that the study procedure was confusing. The final sample consisted of 500 participants (248 Democrats, 252 Republicans; 270 females, 225 males, 5 self-identifying participants; $M_{age} = 44.95$, $SD_{age} = 12.68$). This study was not pre-registered.

Materials and Procedure

Study 1 used a 2 (Role: predict vs. react) $\times 3$ (Topic: abortion vs. immigration vs. gun control) between-participants design. At the outset of the study, participants were randomly assigned either to a *predict* (referred to as “predictors”) or a *react* (referred to as “reactors”) condition. Participants were then randomly assigned to one of the three topic conditions according to the following procedure. First, participants reported their political orientation and their attitudes on three key political issues in the United States: immigration policy, abortion access, and private gun ownership. We chose these topics based on evidence that opinions on these topics are strongly polarized between political parties and largely homogenous within political parties (Hawkins et al., 2018).

Participants indicated their stance on each of these issues by selecting from binary response options representing a liberal view and a conservative view on each issue. For example, for the abortion access question, participants were asked, “Which of the following statements best reflects your views on the issue of abortion access in the United States?” Participants then selected either, “Abortion access in this country should be **protected**” (liberal view), or “Abortion access in this country should be **restricted**” (conservative view).

Among the topics regarding which participants reported a view that aligned with the typical view of their political party⁴, we randomly assigned participants to consider one of these topics as the focal topic for the next part of the study. For example, if a Democrat participant reported a conservative view on gun control, that topic would not have been eligible for random selection as the focal topic. This process ensured that all participants were being asked about a topic regarding which they held a view that aligned with the typical view of their party⁵.

Next, participants responded to dependent measures. Predictors were asked to predict how a political ingroup member would react if they adopted the opposing party’s view on the topic (i.e., dissenting belief change). Reactors reported their reactions to an ingroup member who adopted the opposing party’s view on the topic. For example, a Republican participant in the predict condition who reported a conservative view on gun ownership would be asked to predict how another Republican would react if they were to change their mind on the issue of gun ownership from “Access to private gun ownership in this country should be **protected**” to “Access to private gun ownership in this country should be **restricted**.” If this participant were

⁴ Across all topics, participants in this sample reported that they held the typical view of their party 82.4% of the time. We do not see significant differences in the results reported when controlling for whether participants hold counter-normative views on the non-focal topic. These analyses are reported in the OSM.

⁵ We used these three political topics and this process of random assignment to political topics in Studies 2 and 4a.

in the react condition, they would be asked to report their reactions toward another Republican whom they were told adopted the opposing party's view.

We used a five-item Likert-type scale measure of social rejection (adapted from Cavazza et al., 2014), which served as the key dependent measure in this study ($\alpha = .94$). We used this measure because social rejection is an experience that people have a strong desire to avoid (Williams, 2007) and is especially relevant to theorizing about intragroup dynamics, loyalty, and dissent (Jetten & Hornsey, 2014). 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) exclude you? (2) ignore your input? (3) reject you? (4) disrespect you? (5) criticize you?" (1 = *Not at all*; 7 = *Very much*). The items in the react condition were similar, but the wording was modified such that reactors reported how they would react to another member of their political ingroup who changed their mind.

Finally, participants responded to attention checks and demographic questions, including age, gender, ethnicity, and U.S. zip code.⁶

Results

Were predictors' expectations about ingroup member reactions to dissenting belief change accurate? To examine the effects of role and topic conditions on the social sanction composite measure, we conducted a 2 (Role: predict vs. react) \times 3 (Topic: abortion vs. immigration vs. gun control) ANOVA. This test revealed a significant main effect of role, wherein predictors ($M = 3.95$, $SD = 1.58$) overestimated how much reactors would reject them for dissenting belief change ($M = 2.59$, $SD = 1.59$), $F(1, 494) = 93.12$, $p < .001$, $\eta p^2 = .16$ (Figure

⁶ Following our primary dependent measures, participants responded to additional survey items that were included for a separate research project.

1). The main effect of topic was not significant, $F(2, 494) = 2.04, p = .131, \eta p^2 = .01$, nor was the interaction between role and topic, $F(2, 494) = 0.15, p = .864, \eta p^2 = .00$, suggesting that the effect of role on rejection did not differ significantly by topic condition.

Next, we conducted a 2 (Role: predict vs. react) \times 2 (Party: Democrat vs. Republican) to examine whether the perception gap varied in magnitude between Democrat and Republican participants. The interaction was not significant, $F(1, 496) = 0.45, p = .505, \eta p^2 = .00$, suggesting that the perception gap for social sanctions was similarly present for both Democrat and Republican participants. The full results of the analyses for each topic and by participant party identity are reported in the OSM.

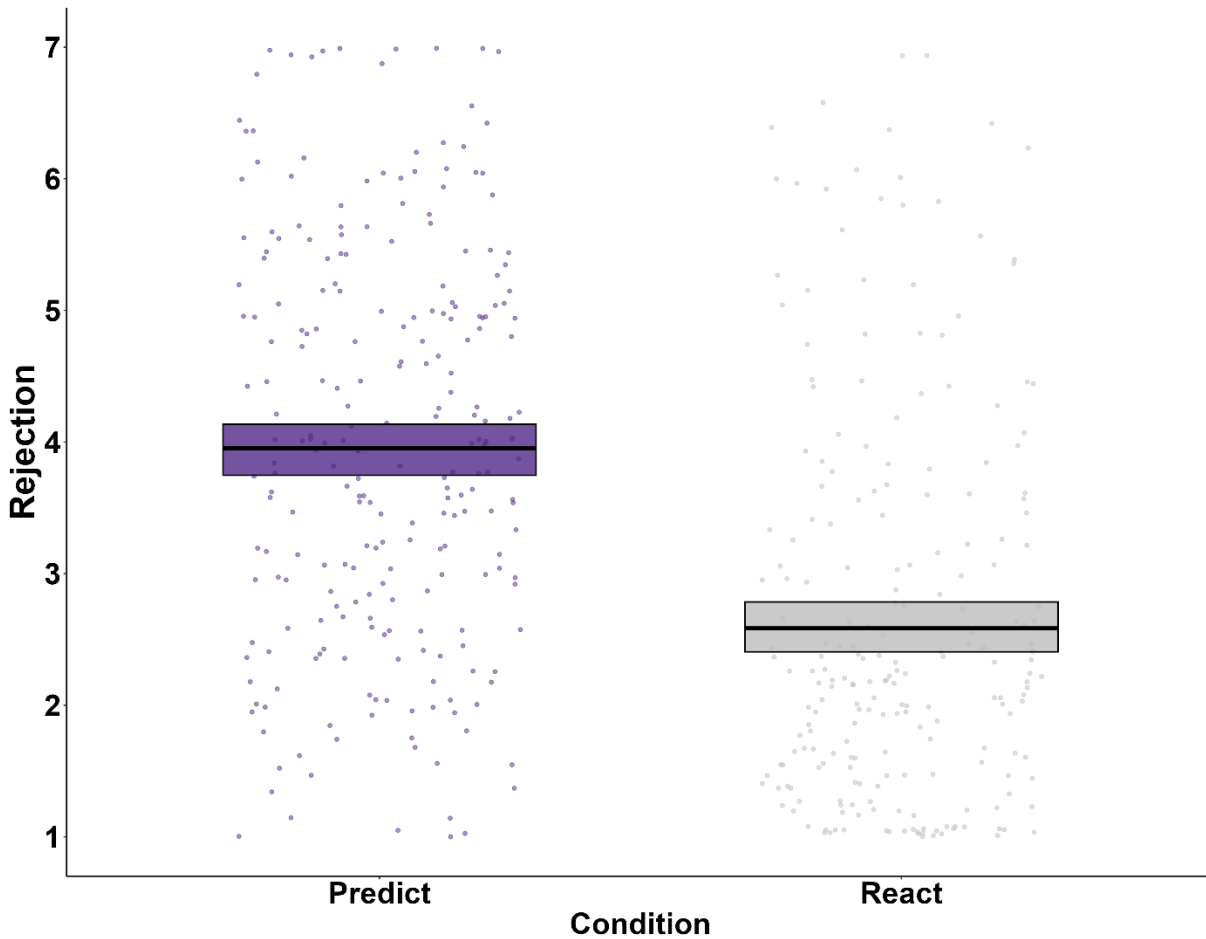


Figure 1. *Predictors overestimate rejection relative to reactors' reports.* Note. The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval.

Discussion

Study 1 found that predictors overestimated how much a political ingroup member would reject them for dissenting belief change. This perception gap emerged across all three political topics and was similar for both Democrats and Republicans. One potential limitation is that participants considered a relatively extreme form of belief change—shifting from one side to the opposite side of a polarizing issue (e.g., from pro-life to pro-choice). To assess generalizability, we tested this effect across varying degrees of dissenting belief change on several different topics

in Studies 2-5.⁷ A second limitation is that Study 1 relied solely on Likert-type items, which may have introduced social desirability concerns among reactors (who may have wanted to appear open-minded). In Study 2, we addressed this by including financially incentivized behavioral measures of rejection.

Study 2: Paired Dyads & Behavioral Measures

Study 1 showed that partisans overestimate how much they would be rejected by an ingroup member for dissenting belief change. Study 2 tested whether this perception gap persists when participants engage in a live interaction with another ingroup member who shares their view on a political topic. This pre-registered dyadic design allowed us to assess predictor accuracy against a specific reactor's response, providing a ground-truth benchmark for misperception. It also enabled the use of behavioral and open-ended measures: After their conversation, participants predicted (or reported) how they would evaluate a partner who changed their mind on a political topic, including how much of a monetary bonus they would allocate and whether they would choose to continue working with that person. In addition, participants provided an open-ended description of how they expected their partner to feel (predict condition) or how they themselves felt (react condition) in response to the belief change; responses were coded to assess perceived versus actual interpersonal effects.

Methods

Participants

We recruited participants from a university participant pool of Amazon Mechanical Turk workers. A priori power analysis indicated that a sample of at least 300 participants was required

⁷ See also Pilot Study 2 in the OSM, which replicated Study 1 for a minor change in beliefs.

to detect a moderate effect size (Cohen's $d = .50$) with 90% power ($\alpha = .05$) in an independent samples t-test comparing social sanction scores across conditions. Given that Study 2 occurred in two distinct phases and involved live participant pairing, we over-recruited heavily to ensure that the final sample was sufficient after accounting for attrition and excluding participants who were not paired with a partner in the survey due to technical issues. Anticipating a 50% eligibility rate for the full study, we aimed to recruit 1,000 participants (500 Democrats, 500 Republicans) to complete the initial screener which was administered twenty-four hours prior to the main study.

In total, 885 participants completed the pre-screen survey. The following day, eligible participants were invited via email to participate in Study 2 if they reported at least one political view that was aligned with the typical view of their political party on the topics of abortion, gun control, and immigration in the United States. Almost all participants ($N = 817$; 92%) met this criterion and were invited to complete the study. Among these, 278 participants (190 Democrats, 88 Republicans; 154 females, 124 males; $M_{age} = 45.90$, $SD_{age} = 12.04$)⁸ completed all study procedures, passed all attention checks, and were successfully paired with another participant according to our pre-registered pairing and exclusionary criteria.

Procedure

Pre-screen. Participants completed a two-minute survey in which responded to political demographic and opinion questions, indicating their political orientation (Republican vs. Democrat); attitudes on the issues of abortion, immigration, and gun control (binary choice: liberal vs conservative stance—similar to Study 1); and how strongly they agreed with the

⁸ Due to a technical difficulty, a portion of participants who completed Study 2 were not successfully paired with another participant based on our pre-registered pairing criteria. These participants are not included in the analyses reported in the main text. Auxiliary analyses including all participants—even those for whom the randomized pairing system failed—yielded identical conclusions for all hypothesis tests and are reported in the OSM.

selected stance on each of the topics (1 = *Strongly Disagree*; 7 = *Strongly Agree*). Participants also responded to basic demographic questions (age, gender, ethnicity).

Study 2. Once participants accepted the invitation to participate, they were randomly assigned to either a predict condition or a react condition. Following a similar procedure as Study 1, we randomly assigned participants to consider one of the political topics from the pre-screen survey on which they held a party-consistent view was randomly selected as the focal topic for Study 2.

Upon entering the survey, we told participants that the purpose of this study was to examine how well people perform on cooperative tasks when working with another person who holds either similar or different political beliefs. To that end, we told participants that they could be paired with an ingroup member or an outgroup member; however, in reality, all participants were paired with ingroup members who shared the same view on the randomly selected political topic.

We then informed participants that the study would proceed in the following order: First, they would have a brief conversation with their partner. Next, they would complete a brief survey on their own. Then, they would rejoin their partner to complete a cooperative task where they would have a chance to win a financial bonus based on their joint performance. In reality, the cooperative task did not exist. We included a fictitious cooperative task to create the pretense of future interaction between participant pairs, thereby enhancing the ecological validity of our findings. We expected that this would increase the stakes for cooperation and coordination between participants, thereby making the measures of anticipated and actual social sanctions both more consequential and realistic. The cover story also allowed us to include behavioral measures of rejection related to the task.

After reading about the study procedure, participants entered a chatroom using Chatplat—an in-survey platform where participants can have a live chat conversation with another participant in the study (www.chatplat.com; Brooks & Schweitzer, 2011)—and were paired with another participant. Using survey logic, participants were randomly paired with another participant who met the following pre-registered criteria: (1) both participants reported that they support the same political party, (2) both participants were randomly assigned to the same political topic condition, regarding which they both held the typical view of the party, and (3) participants were randomly assigned to opposite role conditions.

This procedure allowed us to create dyads of predictor-reactor ingroup members who shared their party's typical view on the assigned topic. Pairing occurred as soon as another eligible participant entered the chatroom. Participants were informed, veridically, that their partner was another participant taking the survey who identified with the same political party and selected the same binary-choice statement representing their view on the topic (all study materials are reported in the OSM under Study 2 Additional Information). Participants completed two comprehension checks to ensure this information was understood before the chat began.

Once dyads were formed, participants were given five minutes to chat with their partner and were instructed to follow a simplified version of the “fast-friends” procedure (Aron et al., 1997) in which participants took turns asking each other get-to-know-you style questions⁹. When the conversation concluded, we told participants that they would soon reconnect with their partner to complete the cooperative task. On the next page, we gave participants the following

⁹ We chose this prompt because we wanted participants to get to know each other, but we did not want them to drift into discussing political topics, which may have reduced the realism of the ensuing experimental procedure.

message, “Before you are reconnected with your partner, there are a few things you should know about this experiment... PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY” (emphasis original). Participants were then given specific instructions based on which role condition they were assigned to.

We told predictors that at this point in the study, the research team was presently informing their partner that between pre-screen and the present phase of the study, the predictor was assigned to watch a non-partisan video from ProCon.org (which is a real, non-partisan informational website) that described both party’s views on the randomly assigned political topic. Although predictors never actually watched this video, we told them that their partner would be informed that watching the ProCon video caused them to change their position on the randomly assigned topic. Specifically, participants were shown the same seven-point scale item upon which they reported their agreement with their party’s stance on the randomly assigned topic during the pre-screen (1 = *strongly disagree*; 7 = *strongly agree*) and we told them that their partner would learn that their agreement moved one point to the left on the scale after watching the video.

We showed predictors screenshots of the information that their partner in the react condition would receive (which was the exact information their partner actually saw). These instructions implied that the predictor still agreed with the typical view of the party on the issue, but their agreement with this position decreased by the smallest measurable amount. This allowed us to test whether the findings from Study 1 were specific to belief change across party lines, or whether a milder form of dissent—reduced agreement without full reversal—would produce similar effects.

Reactors received the exact information that predictors were told would be shared with their partner. Specifically, we told reactors that their partner was assigned to watch a ProCon.org video on the randomly selected political topic between the pre-screen and the main study session. We told reactors that their partner still agreed with the typical view of the party, but watching this video led the participant to decrease their agreement with this position by one point to the left on the seven-point scale measure compared to how this participant responded to the same item in the pre-screen survey.

All participants were asked to confirm that they read and understood the instructions before responding to the dependent measures in the next part of the survey.

Measures

First, all participants responded to a nine-item scale to measure anticipated social sanctions, which consisted of the same five-item measure of social rejection from Study 1 and four additional items to measure anticipated reputation damage (Levine & Schweitzer, 2015), which asked predictors how another ingroup member would react across the following dimensions after learning that their agreement with the typical view of their party decreased, “how much do you think this person will feel inclined to... (1) exclude you, (2) ignore your input, (3) reject you, (4) disrespect you, (5) criticize you (6) be upset with you, (7) like you, (8) respect you, (9) trust you?” (order randomized; items 7-9 were reverse-coded for analysis). All items were answered on Likert-type measures and the wording was modified for reactors to measure actual reactions (1 = *Not at all*; 4 = *A moderate amount*; 7 = *Very much*).

We found that all nine of these items formed a single factor and showed high reliability ($\alpha = .97$). Thus, we combined all items into a single composite measure, referred to henceforth as

a measure of anticipated and actual “social sanctions”, which we also used in Study 3. Following these survey measures, participants responded to an open-ended response item asking them to describe how they think their partner felt and why (predict condition) or how they felt and why (react condition) after learning the information about belief change.

Next, participants responded to two behavioral measures of social sanctions. Participants first a basic description of the cooperative task they believed they were about to complete, which was described as using communication and cooperation to navigate a virtual maze as a team for a chance to win a financial bonus. We told participants that their bonus payout would depend upon their ability to successfully cooperate with their partner to complete this task. Given that each participant’s financial outcomes were ostensibly linked to their partner, this interaction context was designed to mitigate reactor social desirability bias (i.e., wanting to appear tolerant, open-minded) and encourage them to behave in ways that are consistent with their true feelings toward the target.

After reading the task description, participants responded to a binary-choice behavioral measure of partner choice. Specifically, reactors chose whether they wanted to work on the cooperative task with a different partner or maintain the same partner. We told predictors that their partner would have this choice and asked them to predict whether their partner would choose to work with someone else or with them on the task.

Next, participants completed an adapted version of a dictator game. Participants were told that if they successfully completed the task, they would be awarded a \$0.10 bonus that one person in the dyad would be randomly assigned to distribute between them. In reality, reactors were always assigned the role of bonus distributor. Prior to the ostensible cooperation task,

reactors chose how to distribute the \$0.10 bonus between themselves and their partner and predictors were asked to predict how their partner would distribute the bonus.

After participants responded to all study measures, they reached the end of the study. We told participants that they were disconnected from their partner due to a technical issue that was no fault of their own and that, consequently, they would not be able to attempt the cooperative task. This information came at the very end of the study and did not influence participant survey responses, and all participants were awarded a \$0.10 bonus. To conclude, we gave participants a debriefing form that described the use of deception in the study and provided the rationale behind the study procedure.

Results

Among paired dyads of acquainted ingroup members, did predictors accurately estimate how they would be evaluated? Findings from Study 2 provided further evidence for the perception gap observed in Study 1. Supporting our pre-registered hypothesis, predictors expected significantly greater social sanction ($M = 3.14$, $SD = 1.14$) for dissenting belief change than reactors reported ($M = 2.02$, $SD = 0.84$), $t(276) = 9.32$, $p < .001$, $d = 1.12$. In total, 110 predictors (79%) overestimated social sanctions relative to what their react condition partner reported (7 predictors were accurate; 22 underestimated). Follow up moderation analyses showed that the perception gap between predictors and reactors did not vary meaningfully by topic, partisan identity, or attitude strength (see OSM for full results)¹⁰.

¹⁰In all subsequent studies, we report exploratory tests of moderation and replicate our main analyses controlling for the same set of variables (e.g., topic, party affiliation, political identity strength, believability). Across all studies, the key role (or role \times condition) effect remains statistically significant when controlling for these variables, and no significant moderation effects are detected. See OSM for full results reported for each study.

We note that the means for social sanctions in both the predict and react conditions are slightly lower than we observed in Study 1; however, the effect size of the perception gap is similar. This pattern may suggest that interpersonal contact can buffer absolute anticipated and actual social sanctions, even as the effect size of overestimation remains the same. Alternatively, the lower means in Study 2 may be because participants were predicting and reacting to more modest belief change than in Study 1.

Next, we analyzed the open-ended responses, which asked participants to describe how they thought their partner felt (predictors) or how they felt (reactors) after receiving the information 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 of this information on interpersonal evaluations ranging from -3 (very negative) to +3 (very positive). Zero in this coding scheme represented “neutral / no effect” on the interpersonal evaluation. Coders gave a distinct code for any response that did not answer the question and all rows that received this code were excluded from analysis, yielding a subsample of 219 participants whose responses to these questions were coded.

Participants’ open-ended responses were consistent with what we observed on the survey measures: Participants in the predict condition anticipated significantly worse interpersonal outcomes ($M = -0.83$, $SD = 0.97$) than participants in the react condition reported ($M = 0.00$, $SD = 1.04$), $t(215) = 6.06$, $p < .001$, $d = .83$ (Figure 2; see Study 2 Additional Information in the OSM for the full coding scheme and instructions).

Were participants’ survey and open-ended responses consistent with their financially incentivized behaviors? Yes—we replicated the same pattern of overestimation on both behavioral measures. On average, predictors overestimated how often their partner would choose

to work with someone else on the cooperative task (Predict: 18.71%; React: 7.91%), $\chi^2(1, 278) = 6.11, p = .013, V = .15$, and expected their partner to keep significantly more of the bonus ($M_{predict} = 5.72$ cents, $SD = 1.58$) than their reactor partners kept ($M_{react} = 5.18$ cents, $SD = 1.43$), $t(276) = 2.99, p = .003, d = .36$.

Given the financially incentivized nature of these behavioral measures (i.e., real stakes), we can be more confident that the perception gap findings are not simply due to reactors responding to survey item measures in a socially desirable way (e.g., trying to appear open-minded and accepting¹¹). Specifically, reactors were incentivized to act in ways that reflected their true feelings about their predict condition partner given that (a) any amount of money they gave to their partner in the dictator game would have lowered their own bonus, and (b) their ability to win the bonus hinged upon their ability to cooperate with their partner. Nevertheless, we observed that reactor behaviors were consistent with their self-reported attitudes, distributing the bonus more generously and choosing to reject their partner far less often than predictors expected. Moreover, coded participant open-ended responses were convergent with survey and behavioral findings, providing a comprehensive view of participants' anticipated and actual reactions across a diverse set of measures.

¹¹ An additional pre-registered study reported in the OSM (Supplemental Study 2) tested whether social desirability accounts for the effect. Using a third-party perspective condition to reduce reactors' social desirability concerns, we again replicate predictors' significant overestimation of social sanctions. The full write up of this study and results are reported in the OSM.



Figure 2. Coded open-ended responses revealed that predictors expected their partner would be more upset than their reactors actually were. Note. The plot shows the density distribution of coded scores for predictor and reactor responses. Density reflects the relative concentration of scores at different points on the scale, with higher peaks indicating more frequent or typical responses. The vertical dashed lines represent the mean coded sentiment for each condition. Values to the right of 0 represent an increasingly positive effect on the relationship between partners; values to the left of 0 represent an increasingly negative effect on the relationship.

Discussion

Study 2 replicated Study 1's overestimation effect, this time across survey, behavioral, and open-ended response measures. Using paired dyads and incremental belief change, Study 2 provided panoramic evidence that predictors' estimates about their partner's evaluative and behavioral responses were indeed inaccurate. These findings strengthen the case for a perception gap by showing that it persists even when participants interact directly with an ingroup member and face real financial incentives—conditions that should reduce miscalibration if it were merely an artifact of hypothetical interactions or social desirability

Study 3: Actual Belief Change

Studies 1 and 2 documented that group members overestimate the social costs of dissenting belief change across varying relationship contexts, degrees of belief change, and measures of social sanctions. However, one limitation is that the belief change in those studies was hypothetical. Because people are motivated to view themselves as rational and consistent (Pronin et al., 2002; Ross & Ward, 1996), imagining a shift to a position they do not currently hold may have felt implausible—prompting the inference that such a shift would seem especially deserving of social sanction. In contrast, when partisans actually change their minds on a political issue, perhaps they view that change as thoughtful and justified, thereby reducing anticipated social punishment.

We addressed this concern in Study 3 by using a two-phase design that captured and tested actual belief change on a partisan issue. In Phase 1, predictors completed a counter-attitudinal writing task intended to induce genuine shifts in their political beliefs. Several months later, in Phase 2, the same participants were randomly assigned to either reflect on this prior belief change or not (a manipulation of the Time 2 salience of the belief change) before reporting how much rejection they anticipated from fellow partisans. This approach allowed us to test whether the perception gap depends not just on whether individuals have changed their beliefs, but on whether that belief change is psychologically salient—helping to rule out alternative explanations based on individual differences in susceptibility to belief change or sensitivity to rejection.

Phase 1

Participants

Our sample size for Phase 1 of Study 3 was determined by logistical constraints in recruiting a specific sub-sample of a population during Phase 1 (i.e., participants whose beliefs changed after writing a counter-attitudinal message). For Phase 1, we recruited a total of 1,524 participants from Amazon's Mechanical Turk and Prolific to take part in a research study¹²; however, only a small subset of participants in the predict condition (25%) reported dissenting belief change. To address this limitation and achieve a sufficient sample size, we combined three samples of participants who completed a nearly identical study procedure. Crucially, combining these samples for Phase 1 created a larger pool of dissenters to re-recruit from for Phase 2 of the study, which occurred two months after Phase 1 concluded. The final sample for Phase 1 consisted of $N = 494$ participants (147 predictors, 347 reactors; 262 Democrats, 232 Republicans; 270 females, 225 males, 5 self-identifying participants; $M_{age} = 42.53$, $SD_{age} = 13.05$).

Procedure

The goal of Phase 1 was to replicate the overestimation of social sanctions for dissenting belief change observed in Studies 1 and 2 among predictors who have actually experienced dissenting belief change on a partisan political topic. This study employed a similar experimental paradigm and design as previous studies in which participants were randomly assigned to either a predict or a react condition. The beginning of Phase 1 was similar to previous studies in which participants reported their partisan identity and their attitudes on central political topics

¹² Due to challenges in collecting a large sample of participants who reported dissenting belief change following the counter-attitudinal writing task, participants in Phase 1 of Study 3 were collected from three separate studies. The methods and findings for each of the three studies are nearly identical and are reported in full in the OSM.

(abortion, gun control, immigration) using the same binary-choice questions from Studies 1 and 2.

After being randomly assigned to consider a single political topic upon which participants reported a view that was consistent with the typical view of their party, predictors were instructed to complete a counter-attitudinal writing task in which they wrote a persuasive message in favor of the opposing party's view on that topic. We employed this procedure because previous research has shown that it produces a relatively consistent effect of belief change outcomes in a similar context (e.g., Briñol et al., 2012; Carlsmith et al., 1966; Greenwald & Albert, 1968). Predictors reported how strongly they agreed with the typical view of their party both before (T1) and after (T2) completing the counter-attitudinal writing task using a 0-100 slider scale (0 = *Strongly Disagree*; 100 = *Strongly Agree*). We determined whether a predictor reported a dissenting shift in their beliefs by calculating a difference score between their T1 and T2 agreement measures.

Any predictor whose T2 agreement score was lower than their T1 score was categorized as a dissenter. On average, dissenters' baseline attitude strength was moderately strong ($M = 85.82$, $SD = 14.38$). At T2 post-treatment, dissenters' attitude strength changed, on average, by 16.74 units on the scale ($M = -16.74$, $SD = 20.91$). A one-sample t-test confirmed that the average change in attitude strength among dissenters was significantly different from zero, $t(146) = 9.71$, $p < .001$, $d = 0.80$.

Dissenting predictors then reported how they expected another ingroup member would react if they learned about their completion of this task and its effects on their beliefs using the same nine-item composite measure of social sanctions from Study 1. Specifically, dissenters were asked to predict how another ingroup member taking the study on Prolific would react

knowing that, “This person will learn that you identify as a [Democrat/Republican] and that your agreement with [the party-typical view on the topic] **DECREASED** after writing an essay in favor of the opposing view.”

We told reactors that they would be randomly assigned to evaluate a participant from the predict condition. Reactors saw screenshots of information that was used to describe the counter-attitudinal writing task to predictors and were told that completing this task led this person to decrease how strongly they agreed with the typical view of the party (i.e., dissenting belief change). This was the same information that predictors were told would be shared with their partner (see OSM Study 3 Additional Information for all study materials).

Phase 1 Results

To examine whether predictors overestimated the social costs of dissenting belief change, we conducted an independent samples t-test comparing social sanction scores between predict and react conditions. Consistent with findings from the previous studies, predictors who reported dissenting belief change anticipated significantly harsher social sanctions ($M = 3.61$, $SD = 1.63$) than reactors reported ($M = 2.74$, $SD = 1.63$), $t(492) = 5.41$, $p < .001$, $d = 0.53$.

We conducted an exploratory multiple regression analysis to examine whether predictors' baseline attitude strength (T1) or post-task attitude strength (T2) moderated the overestimation of social sanctions relative to reactors' actual reports. The model was not significant, $F(2, 144) = 0.033$, $p = .967$, $R^2 = 0.001$, indicating that neither baseline attitude strength ($\beta = 0.001$, $SE = 0.011$, $t(144) = 0.08$, $p = .941$) nor post-task attitude strength ($\beta = 0.001$, $SE = 0.007$, $t(144) = 0.17$, $p = .861$) significantly predicted overestimation.

Next, we conducted an exploratory regression analysis to examine whether the magnitude of belief change (T1 minus T2 attitude strength) predicted the overestimation of social sanctions. The magnitude of belief change was not a significant predictor of overestimation, $\beta = 0.001$, $SE = 0.006$, $t(145) = 0.154$, $p = .878$, $R^2 = 0.0002$. Taken together, these results suggest that predictors overestimate ingroup social sanctions for any amount of dissenting belief change on partisan topics, regardless of their baseline attitudes, their post-task attitudes, or the magnitude of their belief change. This may have been due to the fact that reactors were not given information about the magnitude of predictors' change. Thus, future research should examine whether specific information about baseline attitude strength and the amount of attitude strength change moderates these effects.

Findings from Phase 1 replicate the main effect of overestimation observed in Studies 1 and 2, demonstrating that the effect persists with actual belief change. However, one interpretational challenge is that it is essentially impossible to assign people at random to change (or not change) their attitudes on a divisive moral issue. Therefore, it is possible that some unaccounted-for person-level variable led predictors to both (a) change their beliefs and (b) overestimate social sanctions.

To push toward causal evidence despite the fact that it is impossible to randomly assign people to change their attitudes, Phase 2 randomly assigned the subset of participants who experienced belief change either to a condition in which their dissenting belief change was made salient or to a condition where it was not made salient. This allowed us to ensure that we were making comparisons within the same population of people—that is, those who actually experienced dissenting belief change—helping to rule out person-level confounds. This design also clarifies that overestimation emerges when belief change is *salient*, suggesting the

perception gap is not merely a function of trait-level susceptibility to belief change, but how psychologically present that change is when anticipating social judgment.

In Phase 2 of Study 3, dissenting predictors from Phase 1 were re-contacted approximately two months later—a delay we expected would be sufficient for most to forget the specific details of their earlier participation. Participants were randomly assigned to receive either a detailed reminder about their dissenting belief change or a vague, non-specific reminder about their prior participation (see OSM for full materials). All participants then predicted how an ingroup member would respond if they learned about their participation in Phase 1.

Phase 2 Participants. The sample size for Phase 2 was determined by participant attrition over a delay period of approximately two months between Phase 1 and Phase 2. We invited all 147 predictors from Phase 1 who reported dissenting belief change to participate in Phase 2 of the study. Among these, ninety-three (63%) completed Phase 2 (50% Democrat, 50% Republican; 62% female, 37% male; $M_{age} = 43.73$, $SD_{age} = 12.45$).

Procedure

After a delay period of two months, participants entered the study and were randomly assigned to either have their belief change from Phase 1 made salient (“salient condition”) versus not made salient (“non-salient condition”). An example of the instructions participants saw for the topic of abortion access are shown below. Participants in the control condition only saw the first paragraph. Participants in the salient condition received the full message.

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.

[salient condition only from here on] 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: “abortion access in this country should be **[protected/restricted]**”.

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: “abortion access in this country should be **[protected/restricted]**”.

After you finished writing a message in favor of the opposite view on this topic, you reported that your agreement with the statement, “abortion access in this country should be **[protected/restricted]**” **DECREASED**. That is, your agreement with the majority position of the [ingroup] 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 [outgroup] party.

Following these condition-specific instructions, all participants were then asked to predict how an ingroup member would react upon learning about their participation in the previous study using the same nine-item composite measure of social sanctions from Study 2. To conclude the study, participants responded to a manipulation-check question which asked participants to recall specific details about their belief change during Phase 1 as a test of whether these details were salient.

Phase 2 Results

Manipulation Check. In response to the manipulation check item, 74% of participants in the salient condition accurately recalled details about their belief change, whereas only 22% of

participants in the non-salient condition were accurate. These findings confirm that the belief change was more salient for participants in the salient condition than those in the non-salient condition.

We next tested whether anticipated social sanctions differed between the salient and non-salient conditions. Participants in the salient condition ($M = 3.40$, $SD = 1.45$) anticipated significantly more social sanctions than participants in the non-salient condition ($M = 2.31$, $SD = 1.48$), $t(90) = -3.53$, $p < .001$, $d = .74$. Next, we conducted a one-way ANOVA to compare predictors' social sanction estimates in each Phase 2 condition against reactor reports of social sanctions towards dissenting ingroup members from Phase 1. The main effect of this test was significant, $F(2, 436) = 5.60$, $p = .004$, $\eta p^2 = .025$. Post-hoc comparisons using Tukey's HSD test indicated that predictors in the Phase 2 salient condition ($M = 3.40$, $SD = 1.45$) overestimated social sanctions compared to reactor scores from Phase 1 ($M = 2.74$, $SD = 1.63$), $p = .020$, $d = .41$. In contrast, participants in the Phase 2 non-salient condition did not overestimate reactor social sanctions ($M = 2.31$, $SD = 1.48$), $p = .224$, $d = .27$ (Figure 3).

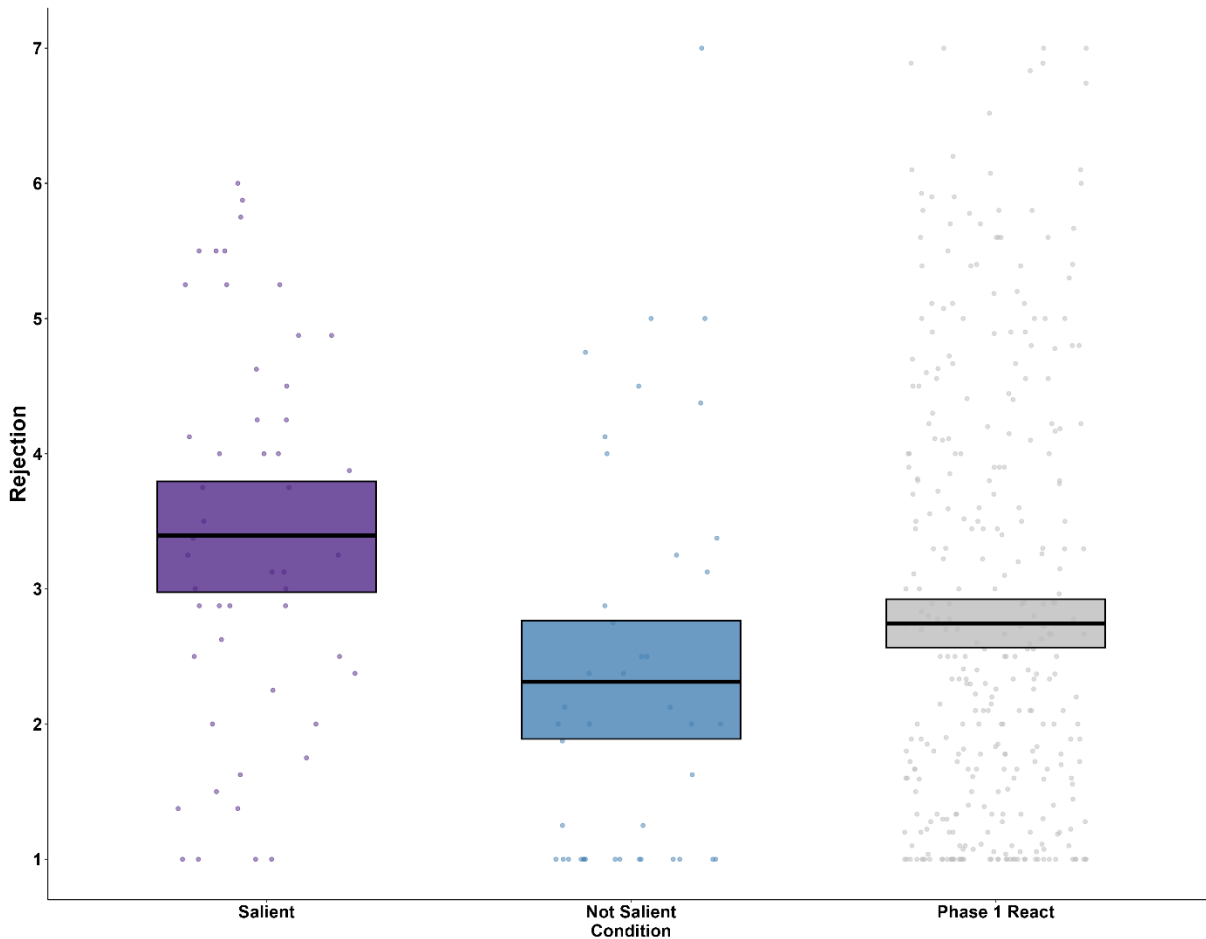


Figure 3. Predictors reminded of their belief change overestimated social sanctions. Those not reminded did not. Dissenting predictors whose belief change was not made salient did not. Note. The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval.

Discussion

Study 3 extended the present research by replicating the perception gap among participants who actually changed their minds on a partisan topic—rather than imagining belief change, as in Studies 1 and 2. The study also showed that the perception gap depends not just on whether belief change occurred, but whether that change was psychologically salient. Only predictors who were reminded of their belief change overestimated social sanctions in Phase 2 of the study, suggesting that the effect is not driven by stable individual differences in susceptibility

to belief change or other person-level factors, but by the salience of reputational threat when dissent is top-of-mind.

At the same time, we acknowledge a limitation of this design: Participants in the non-salient condition were not given a psychologically plausible basis for rejection beyond participation in an earlier study. We address this concern in Studies 4a-c and Study 5 by maintaining a plausible basis for rejection across experimental conditions. Overall, by combining actual belief change with experimental manipulation, Study 3 provided further evidence that dissenting belief change is a key psychological driver of the perception gap.

Study 4a-c: Boundary Conditions and Mechanism

Studies 1-3 showed that individuals systematically overestimate the social costs of dissenting belief change. What explains this misperception, and under what conditions is it most likely to occur? In Studies 4a-c, we address these questions by testing both the boundary conditions and psychological mechanism underlying the perception gap.

We begin in Study 4a by testing whether this gap is uniquely tied to belief change on partisan political topics—issues that clearly signal group affiliation and loyalty (e.g., abortion, gun control)—or whether it instead reflects a more general form of social pessimism. Prior research has demonstrated a tendency to expect more negative evaluations during social interactions than is warranted (Boothby et al., 2018; Bruk et al., 2018; Elsaadawy & Carlson, 2022; Savitsky et al., 2001; See Epley et al., 2022, for review). From this perspective, the perception gap observed in Studies 1-3 may not be specific to dissenting belief change, but rather a manifestation of a more global bias whereby people expect to be disliked or rejected more than they actually are. If this is the case, individuals should overestimate social sanctions to a similar

degree for both partisan and non-partisan dissenting belief change. In contrast, if the gap is significantly larger for partisan topics—where belief change may signal ingroup disloyalty—this would provide evidence that the perception gap is not merely an instance of a general tendency to expect negative evaluations during social interactions, but a distinct phenomenon rooted in reputational concern in polarized contexts.

To understand why partisan belief change may trigger this perception gap, we turn to research on loyalty as a central dimension of group life. In group settings, loyalty functions as a signal of commitment and dependability (Zdaniuk & Levine, 2001; Van Vugt & Hart, 2004), and partisan attitudes often serve as identity-relevant markers of group membership (Connors, 2019; Funkhouser, 2020; Galak & Critcher, 2023; Kahan, 2013; Van Bavel & Pereira, 2018). As a result, changing one's mind on a partisan issue may be seen as a potential act of disloyalty—making dissent feel reputationally risky. Drawing on the concept of signal amplification bias (Vorauer et al., 2003), we propose that dissenters—motivated to avoid social rejection—may overestimate how strongly others will interpret their belief change as a signal of disloyalty, leading to inflated expectations of social sanctions. Studies 4a-c test this account directly, examining whether predictors' meta-perceptions of loyalty mediate the perception gap—and whether this mediation effect is stronger for partisan versus non-partisan belief change.

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 driven by 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 mediated by loyalty meta-perceptions, and that this indirect effect would be stronger for partisan (vs. non-partisan) topics.

We also tested whether predictors in the partisan (vs. non-partisan) condition reported greater self-censorship, and whether this effect was sequentially mediated by loyalty meta-perceptions and anticipated social sanctions.

Method

Participants

We recruited 400 participants to yield approximately 100 participants per cell after making exclusions for failed attention and comprehension checks. This sample size was determined based on an a priori power analysis that indicated 92 participants per cell would provide sufficient power to detect the smallest hypothesized effect size ($f = .17$) at 90% power in a two-way type III ANOVA with alpha set to .05. Our final sample consisted of $N = 393$ participants (240 Democrats, 153 Republicans; 253 females, 134 males, 6 self-identifying participants), after applying all exclusions.

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*¹³). 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 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 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.

¹³ 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.

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 questions¹⁴. 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 ($\alpha = .96$). 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.

To clarify the nature of the interaction, we conducted Tukey's HSD post-hoc tests to compare predictors and reactors within each topic condition. In the partisan condition, predictors anticipated significantly more social sanctions ($M = 4.68, SD = 1.45$) than reactors reported ($M = 2.87, SD = 1.74$), $t(389) = 8.64, p < .001, d = 1.23$. In the non-partisan condition, predictors again anticipated more rejection ($M = 2.48, SD = 1.45$) than reactors reported ($M = 1.67, SD = 1.17$), $t(389) = 3.84, p < .001, d = 0.55$. 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

¹⁴ 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.

social evaluations and may instead stem from additional psychological mechanisms specific to politically polarized contexts.

To test whether the larger perception gap observed in the partisan condition was driven by 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 \times 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 appears to stem from predictors' belief that their dissent will signal greater disloyalty than it actually does—fueling inflated expectations of rejection.

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 in the OSM.

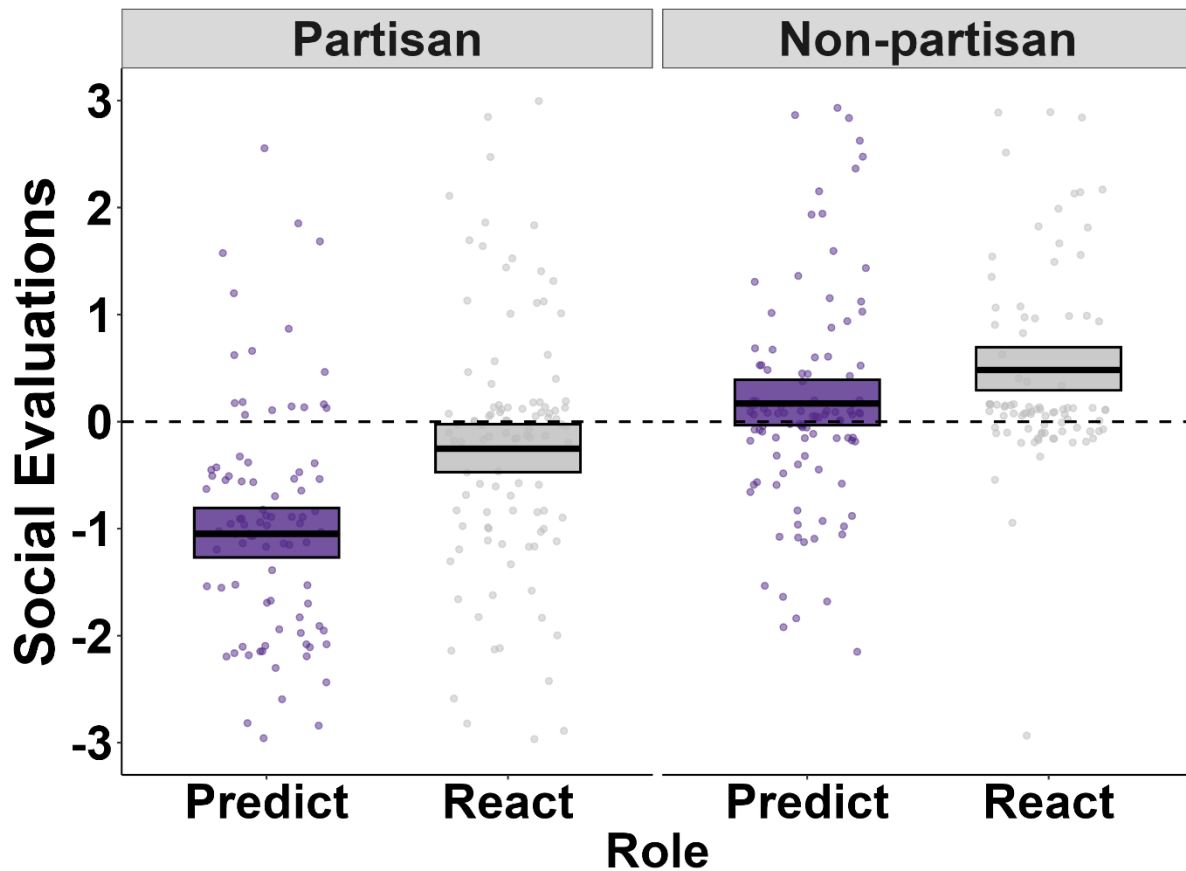


Figure 4. Comparing predictor estimates of social evaluations against reactors' actual evaluations within each topic condition. *Note.* The bi-polar composite has been re-scaled on the y-axis to center the midpoint on zero, which represents a neutral evaluation. Positive scores represent a positive social 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.

We next examined whether these distorted expectations had downstream consequences for self-censorship. That is, we tested whether heightened concerns about loyalty and anticipated rejection in partisan contexts would reduce 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 (indirect₁ = 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 (indirect₂ = 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 suggest 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 at play.

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 is not merely a reflection of more global bias leading to negative social evaluations, but instead reflects a group-specific reputational concern when such belief change might signal disloyalty to the group. It further demonstrated that this gap was driven, 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 ensure adequate sample size. Based on the interaction effect observed in Study 4a ($\eta p^2 = .03$), a sample of 342 participants would be required to achieve 90% power in a 2×2 between-subjects design. Although the recruited sample fell slightly below this threshold after excluding participants who failed comprehension and attention

checks, a post hoc power analysis using the observed effect size ($\eta p^2 = .029$) indicated achieved power of 83.7%, suggesting that the study retained sufficient power to detect the targeted interaction effect.

Procedure. Participants were randomly assigned to a 2 (Role: predict vs. react) \times 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 ($M = 6.78, SD = 0.56$) significantly more than participants in the non-partisan topic condition ($M = 4.09, SD = 0.98$). 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 scale¹⁵. 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).

To clarify the nature of the interaction, we conducted Tukey's HSD post-hoc 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

¹⁵ 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.

reported ($M = 2.18$, $SD = 1.62$), $t(278) = 3.36$, $p = .005$, $d = 0.58$. Although the overestimation effect was present across both framing conditions, it was substantially larger when belief change occurred in the context of ideologically polarized topics. This pattern is consistent with the hypothesis that dissent on politicized topics is seen as especially reputationally risky, potentially triggering 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 \times 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 driven 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 in turn leads them 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. The sample size for Study 4c was determined based on an a priori power analysis, which indicated that 610 participants would be required to detect a partial eta squared effect size of .017 for a two-way interaction with 90% power. We initially recruited 670 Democrat and Republican participants from Prolific Academic to participate in the research study. In accord with our pre-registration, we included the same attention and comprehension checks in the survey from Study 4b that participants had to pass in order to participate in the study. Our final sample size consisted of $N = 596$ participants (272 Republicans, 324 Democrats; 247 males, 343 females, 6 self-identifying participants; $M_{\text{age}} = 41.46$, $SD_{\text{age}} = 13.34$).

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 \times 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 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 exploratory post-hoc 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 ($M = 2.10$, $SD = 1.37$), $t(592) = 5.80$, $p < .001$, $d = 0.72$. 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.

Mediation Analysis. 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 4a-c Discussion

Studies 4a-4c were designed to both replicate the perception gap in new contexts and extend prior findings by testing its boundary conditions and underlying psychological mechanisms. Across all three studies, the overestimation of rejection was substantially larger for partisan (vs. non-partisan) issues—suggesting that this perception gap is not merely a product of general pessimistic expectations about social interactions but is heightened when belief change may signal diagnostic information about group loyalty. In Studies 4a and 4b, predictors in the partisan condition expected to be seen as more disloyal than reactors actually judged, and this exaggerated expectation mediated the overestimation of rejection. Additionally, Study 4a found that these distorted expectations have downstream consequences for self-censorship.

Study 4c replicated the interaction of role and topic condition for rejection and again found that loyalty mediated the link between role and anticipated rejection across conditions. However, the moderated mediation effect did not replicate. Several contextual factors may account for this discrepancy. In Study 4c, predictors rated the framing manipulation significantly less believable in the non-partisan than the partisan condition ($p < .001$, $d = .44$), potentially leading participants to infer partisanship even when it was not explicitly stated. Moreover, the partisan framing manipulation was less effective overall, as reflected in smaller manipulation check effect size ($\eta p^2 = .37$ in Study 4c vs. $\eta p^2 = .57$ in Study 4b) and lower believability ratings in the non-partisan condition—which may have attenuated the role \times condition interaction on loyalty ($\eta p^2 = .002$ vs. $\eta p^2 = .048$ in Study 4b). In the final study, with the cognizance that the loyalty moderated mediation effect was not statistically significant in Study 4c, we build on this evidence by experimentally manipulating the conditions theorized to give rise to signal amplification bias.

Study 5: Manipulating Loyalty

Study 5 sought to experimentally manipulate how secure participants felt in their loyalty-based standing within the political ingroup—targeting a key antecedent theorized to contribute to signal amplification bias in the context of dissenting belief change. Building on Studies 4a-4c, which showed that lower loyalty meta-perceptions predicted higher anticipated rejection, Study 5 examined whether reminding predictors of their past loyalty would ease concerns about being perceived as disloyal—thereby reducing expectations of social sanctions for belief change. This approach draws on self-affirmation and trait activation frameworks (Dunning, 2007; Steele, 1988), which suggest that activating valued aspects of one’s identity can reduce defensiveness in identity-relevant domains. In this case, participants were primed to reflect on past demonstrations of loyalty to affirm that aspect of their self-concept before evaluating the reputational consequences of dissent.

Method

Participants. We recruited participants who reported changing their mind on one of five pre-specified partisan political topics¹⁶ in a way that placed them at odds with the majority view of their political ingroup. To identify eligible participants, Study 5 began with a two-question pre-screen asking participants to report their political party and whether they had changed their mind on any of the five topics in the past year in a way that put them at odds with their party. If participants had changed their mind on more than one topic, they were instructed to select the

¹⁶ These five topics were selected based on pilot data identifying the most common belief change topics for Democrats (abortion, immigration, gun control, gender issues, Israel/Palestine) and Republicans (abortion, immigration, gun control, economic policy, climate change). Although this design introduces different topic sets by party, this approach ensured that we could achieve approximately equal numbers of Democrats and Republicans across predict and react conditions for each topic—enabling balanced comparisons between groups. We judged this tradeoff acceptable given that our hypotheses focus on role and loyalty effects rather than cross-party comparisons. We do not observe any effects of topic or partisan identity in the models, suggesting that this decision did not significantly impact the findings. Additional details on the topic-matching procedure, descriptive statistics about topic distributions, and analyses considering topic and partisan identity are reported in the OSM.

one that stood out most. In line with our pre-registered recruitment plans, participants who selected a political affiliation other than Democrat or Republican, or indicated “None of the above” to the belief change item, were screened out and compensated \$0.20 for completing the pre-screen.

A priori power analysis revealed that we would need a sample size of 252 participants to detect a medium-sized effect ($d = .50$) in two primary pre-registered planned contrasts with 90% power at $\alpha = 0.05$. We detail the two primary pre-registered contrasts in the Procedure section below. However, given the complexity of the recruitment scheme and the need for matched topic-distributions across conditions (described in detail in the procedure), we over-recruited heavily to ensure sufficient power for all hypothesis tests. After excluding participants based on pre-registered attention and comprehension check items, the final sample consisted of 620 participants (367 Republicans, 253 Democrats; 301 males, 313 females, 5 self-identifying participants; $M_{age} = 38.47$, $SD_{age} = 13.27$).

Procedure

Study 5 employed a three-cell between-subjects design. Following the initial screening, participants were randomly assigned to a high-loyalty predict, low-loyalty predict, or react condition.

Participants were informed that they qualified for the study because they reported changing their views on a political issue in a way that placed them at odds with their party’s majority stance. They were also told that (a) all participants in the study met the same recruitment criteria, and (b) they would not be paired with someone who changed their mind on the same topic.

This information was included to ensure that all participants understood they were paired with another political dissenter—someone who had changed their mind on a different partisan topic. This step was necessary because, unlike previous studies, which used randomized topic assignment (Studies 1, 2, 4a-c) or experimental procedures to induce belief change (Study 3), Study 5 relied on real instances of belief change. Without this clarification, predictors might assume their partner had not changed their mind and therefore expect harsher judgment than reactors would actually deliver—which would have introduced a potential confound.

To further prevent potential bias in predictors' estimates, we ensured that no participant was paired with an ingroup member who had changed their mind on the same topic. Without this safeguard, predictors might have unknowingly made predictions about someone who changed their mind on the same topic, which could alter the accuracy of their expectations. To eliminate this concern, we applied a weighted-randomization procedure (described below) to prevent topic-matched predictor–reactor pairs and informed all participants that such matching would not occur¹⁷.

Participants completed a comprehension check to ensure understanding of the pairing instructions. Ninety-five percent passed on the first attempt. Of those who initially failed, only three failed a second attempt and were excluded. Participants then advanced to condition-specific instructions.

¹⁷ In a separate, pre-registered study (Supplemental Study 1), we tested the possibility that overestimation effect could be explained by predictors exaggerating the strength of reactors' political attitudes. We tested this idea using a pre-registered 2 (Role: predict vs. react) × 2 (Extremity: moderate vs. extreme) between-subjects design in which predictors were given accurate information about their partner's political attitude strength. Knowledge of reactors' attitude strength did not attenuate the overestimation effect. The results of this study are reported in full in the OSM.

Predictors began the study by completing either a loyalty affirmation intervention (referred to as the high loyalty condition) or a loyalty disaffirmation intervention (referred to as the low loyalty condition). Those in the high loyalty condition reflected on their relationship with their political party and listed the three most significant actions they had taken to support it (e.g., voting for an ingroup candidate, donating to a campaign, advocating for party values). Those in the low loyalty condition listed three significant actions they had taken that went against their party (e.g., voting for an out-party or third-party candidate, criticizing ingroup values). Following this manipulation, participants completed a manipulation check question asking them, “To what extent do you feel like you are a loyal supporter of the [Democrat/Republican] Party?” (1 = *Not at all*; 4 = *Moderately*; 7 = *Very much*).

As noted in the study introduction, this manipulation was designed to influence predictors’ loyalty meta-perceptions by shifting how secure they felt in their group standing. From a signal amplification perspective, individuals often default to a self-protective posture—anticipating that even subtle signs of disloyalty may be exaggerated and socially costly in politically polarized environments. By prompting reflection on either loyalty-affirming or disloyal behaviors, the manipulation was intended to shape how secure participants felt in their loyalty-based standing at the moment of judgment, and, in turn, how vulnerable they felt to being perceived as disloyal.

This approach is consistent with self-affirmation theory (Steele, 1988), which shows that affirming a valued aspect of the self can buffer against identity-relevant threat and reduce defensive responding. In the present case, the intervention focused on affirming dissenters’ sense of group loyalty by having them reflect on past expressions of (dis)loyalty to one’s political ingroup. Supporting this interpretation, research on trait activation suggests that making trait-

relevant behaviors more accessible can shift self-judgments and expectations of how others will evaluate us (Dunning, 2007; Higgins, 1996). Recent evidence also shows that reducing identity threat increases openness in politically fraught contexts (Argyle & Freeze, 2024), suggesting that this type of affirmation may recalibrate predictors' expectations about how they will be socially evaluated after expressing dissent.

For the remainder of the study, participants in both of the predictor conditions followed the same procedure. First, following the same general paradigm as in previous studies, predictors were informed they would be paired with another participant on Prolific who was also participating in the study and who identified with the same political party. Predictors were told that this participant would be informed that they identify as a [Democrat/Republican], and that they had recently changed their mind on the selected topic—resulting in a current view less aligned with typical ingroup beliefs.

Next, predictors responded to dependent measures, in which they estimated how this ingroup member would evaluate them. Predictors were asked to estimate how much ingroup participants would view them as (a) loyal to the [Democratic/Republican] party, and (b) likely to vote for [Democrats/Republicans] in future elections (1 = *Definitely Not*; 7 = *Definitely Yes*; the party name that was shown was always the in-party; $r = .67$). These two items were combined into a single composite measure which served as the measure of the theorized mediator: loyalty meta-perceptions. Next, predictors responded to the central dependent measure of anticipated social sanctions using a six-item scale that included measures from Study 2 about exclusion, rejection, disrespect, upsetting their partner, and losing trust (1 = *Not very much*; 4 = *A moderate amount*; 7 = *Very much*; $\alpha = .94$) and the same bi-polar measure of interpersonal liking from Study 4a.

Finally, participants responded to control measures, including: (a) a predicted alignment item—how much they thought the ingroup member would perceive their views as (mis)aligned with typical party beliefs—and (b) a self-reported alignment item assessing how aligned with the ingroup their own views actually were (1 = *Not at all aligned*; 5 = *Completely aligned*). We included these items to ensure that the loyalty manipulation did not have any spillover effects to other outcomes that could potentially mediate the effect of condition on anticipated social sanctions. The second alignment item also served as a comprehension check. Per our pre-registration, we excluded thirty-six participants who selected “5 – Completely Aligned”, which indicated a failure to comprehend instructions.

Reactors followed a similar procedure as in previous studies. To ensure a uniform experience across all participants, reactors began by completing the loyalty manipulation check item. This allowed us to (a) maintain procedural consistency across conditions and (b) collect a comparison measure of perceived loyalty in the absence of any intervention (i.e., by measuring a sample of participants’ feelings of loyalty absent any intervention).

Next, reactors were told they would be paired with another participant on Prolific who was also taking part in the study. The information presented to reactors mirrored exactly what predictors were told their partner would learn. Namely, that the paired participant supported the same political party and had recently changed their views on a partisan political issue—resulting in a current position less aligned with typical ingroup beliefs. To avoid topic-matching, the issue assigned to reactors’ partner was randomly selected from the remaining four topics not chosen by the reactor during the pre-screen.

We used a weighted-random selection process that—based on base rates observed in Pilot Study 5 (reported in the OSM)—sought to randomly assign reactors to topics with the same

frequency that predictors in the study would list them. For example, in the pilot data we saw that roughly thirty percent of Democrat participants reported belief change on the topic of immigration. Therefore, a Democrat reactor in the sample (who reported belief change on a topic *other than* immigration) had roughly a thirty percent chance of being randomly assigned to evaluate an ingroup member who changed their mind on immigration. We did this to increase the likelihood that we would have balanced numbers of predictors who listed belief change on a given topic and reactors who were randomly assigned to that topic condition to allow for “apples-to-apples” comparisons between predict and react conditions.

The study successfully achieved a balanced distribution of topics. This allowed us to pool variance by condition and test for topic effects in our analyses. The pilot data, a detailed description of the weighted random selection process, and chi-square tests to examine condition-level topic distributions are reported in full detail in the OSM.

Reactors read about the ingroup member who changed their mind on the randomly selected topic and then reported their evaluations using modified items to assess their evaluations on the same dimensions. These items were presented in the same order as the predict condition. At the end of the study, all participants completed attention checks and responded to demographic questions.

Results

Manipulation check.

To assess the effectiveness of the loyalty affirmation manipulation, we first examined whether participants’ self-reported feelings of loyalty to their political ingroup differed across the high loyalty, low loyalty, and react conditions. A one-way ANOVA revealed a significant effect

of condition on loyalty, $F(2, 617) = 23.50, p < .001, \eta p^2 = .071$. Tukey's post-hoc tests revealed that participants in the high-loyalty condition reported significantly stronger feelings of loyalty ($M = 5.70, SD = 1.02$) than those in the low-loyalty condition ($M = 4.91, SD = 1.32$), $t(617) = 6.68, p < .001, d = 0.67$. These results suggest that the intervention successfully influenced participants' self-perceived loyalty in the intended direction.

As an exploratory comparison, we also examined reactor loyalty scores, treating it as a baseline control condition in which loyalty was measured without any experimental treatment. Reactors reported feelings of loyalty in between the mean scores for the two treatment conditions, suggesting that the manipulation successfully shifted self-reported loyalty in both directions relative to the untreated reactor group. Reactors reported significantly greater feelings of loyalty ($M = 5.45, SD = 1.18$) than those in the low-loyalty condition, $t(617) = 4.70, p < .001, d = 0.43$. Participants in the high-loyalty condition reported greater feelings of loyalty than reactors, but this difference did not reach significance, $t(617) = 2.15, p = .081, d = 0.23$.

Loyalty (Meta)Perceptions. To examine the downstream effects of the loyalty affirmation intervention on the proposed mediator—predictors' perceptions of how loyal another group member would perceive them to be—we used a one-way ANOVA to compare loyalty meta-perceptions across the two predictor conditions and the reactor condition. This test revealed a significant effect of condition on loyalty meta-perceptions, $F(2, 617) = 13.80, p < .001, \eta p^2 = .043$. Next, we ran Tukey-adjusted post-hoc tests to account for multiple comparisons, which revealed that predictors in the high loyalty condition believed they would be seen as significantly more loyal to their ingroup ($M = 4.64, SD = 1.40$) than those in the low loyalty condition ($M = 4.13, SD = 1.44$), $t(617) = 3.62, p = .001, d = .36$.

How accurate were these meta-perceptions? Compared to reactors' evaluations of dissenting ingroup members ($M = 4.82$, $SD = 1.36$) predictors in the low loyalty condition expected to be seen as significantly less loyal than they actually were, $t(617) = 5.10$, $p < .001$, $d = .50$. There was no significant difference between high loyalty condition predictor expectations and reactors' actual evaluations, $t(617) = 1.36$, $p = .362$, $d = .14$. Taken together, these results suggest that the loyalty affirmation intervention helped participants feel more secure in their loyalty-based standing—reducing the expectation that others would view them as disloyal.

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.

Having confirmed the effectiveness of the manipulation, we next turned to the primary analyses. We tested two pre-registered orthogonal contrasts using a linear model with contrast-

coded condition variables. Contrast 1 compared predictors' anticipated rejection to reactors' reports of rejection, testing the overestimation hypothesis. For this contrast, both of the predictor conditions were coded as +0.5, and the reactor condition was coded as -1. This contrast compares the average of the predictor conditions against the reactor condition. For Contrast 2, the high loyalty condition was coded as +1, the low loyalty condition was coded as -1, and the reactor condition was coded as 0, allowing us to test whether affirming loyalty reduced anticipated rejection among predictors.

Contrast 1: Predictors vs. Reactors. Contrast 1 compared predictors' anticipated rejection against reactors' reported rejection. Consistent with the overestimation hypothesis, predictors anticipated significantly more social rejection ($M = 3.61$, $SD = 1.50$) than reactors actually reported ($M = 2.36$, $SD = 1.31$), $b = 0.83$, $SE = 0.08$, $t(618) = 10.30$, $p < .001$, $d = 0.83$, 95% CI [0.66, 0.99]. This effect replicates the key finding from earlier studies and demonstrates its robustness in the context of actual dissenting belief change on a partisan political topic.

Next, we explored whether differences in loyalty judgments across roles helped explain the overestimation of rejection observed in Contrast 1. A mediation analysis revealed a significant indirect effect through loyalty (meta)perceptions, $ACME = 0.13$, 95% CI [0.06, 0.21], $p < .001$, indicating that a key reason predictors overestimated rejection was that they expected to be seen as less loyal than reactors actually judged them to be. The direct effect remained significant, $ADE = 0.70$, 95% CI [0.55, 0.84], $p < .001$, suggesting that loyalty (meta)perceptions partially, but not fully, accounted for the overestimation effect.

Contrast 2: High vs. Low Loyalty (within predictors). Next, we tested whether affirming loyalty to one's political ingroup reduces anticipated rejection among predictors. As predicted, participants in the high loyalty condition ($M = 3.33$, $SD = 1.50$) anticipated significantly less

rejection than those in the low loyalty condition ($M = 3.87$, $SD = 1.47$), $b = -0.28$, $SE = 0.08$, $t(618) = -3.61$, $p < .001$, $d = -0.29$, 95% CI $[-0.45, -0.13]$. These results reveal that participants in the high loyalty condition exhibited a significantly smaller misprediction error ($M = 0.97$) than participants in the low loyalty condition ($M = 1.51$). This finding suggests that reminding individuals of their prior loyalty to the group can reduce the signal amplification bias that gives rise to inflated expectations of social sanctions by easing the psychological concern that dissent will be seen as diagnostic of disloyalty.

To test whether the reduction in anticipated rejection was mediated by shifts in loyalty meta-perceptions, we conducted a causal mediation analysis in which condition predicted loyalty meta-perceptions, which in turn predicted rejection expectations. The indirect effect was significant, ACME = -0.12 , 95% CI $[-0.20, -0.06]$, $p < .001$, indicating that affirming loyalty led predictors to expect to be seen as more loyal by ingroup members, which in turn reduced anticipated rejection. The direct effect of condition on anticipated rejection remained significant after accounting for the mediator, ADE = -0.16 , 95% CI $[-0.29, -0.02]$, $p = .021$, suggesting partial mediation. These findings suggest that loyalty meta-perceptions are not only predictive of anticipated rejection but are also responsive to interventions that affirm participants' confidence in how loyal they are seen within the group. Reflecting on past acts of loyalty may have helped reduce the social evaluation threat that gives rise to signal amplification bias, thereby leading participants to expect less rejection from others¹⁸.

¹⁸ As a robustness check, we re-ran the primary Contrast 1 and Contrast 2 models, as well as their respective mediation analyses, controlling for belief change topic, gender, age, political identification strength, party affiliation, meta-perceptions of ingroup alignment, and self-reported alignment. All results remained statistically significant and replicated the same pattern of effects, supporting our main hypothesis tests. We also note that all of these analyses replicate the same results when using the bi-polar measure of social evaluations in place of the rejection composite measure. See Study 5 Additional Analyses in the OSM for full model output.

Next, we conducted parallel mediation analyses to assess whether the mediational effects of loyalty meta-perceptions in both Contrast 1 and Contrast 2 remained significant after controlling for three competing mediators: self-reported loyalty, meta-perceptions of ingroup alignment, and self-reported ingroup alignment. For Contrast 1 (Predictors vs. Reactors), the indirect effect through loyalty meta-perceptions was significant, $b = 0.15$, 95% CI [0.07, 0.23], $p < .001$. For Contrast 2 (High vs. Low Loyalty), the indirect effect on anticipated rejection through loyalty meta-perceptions was again the only significant indirect path, $b = -0.14$, 95% CI [-0.22, -0.06], $p = .001$. In both models, none of the three competing mediators showed significant indirect effects. These findings suggest that loyalty meta-perceptions uniquely accounted for the overestimation of social rejection in this context. Full models and results are reported in the Online Supplementary Materials (OSM).

Together, these mediation analyses are consistent with the idea that insecurity about one's group loyalty—and the resulting concern about being perceived as disloyal—may contribute to the overestimation of social sanctions for dissenting belief change. By increasing participants' felt security in their loyalty-based standing, the intervention appeared to buffer against signal amplification bias and reduce the gap between anticipated and actual social sanctions.

Discussion

Affirming one's loyalty to the group reduced the social evaluative concerns that give rise to signal amplification bias—specifically, the expectation that others will interpret one's dissenting belief change as a stronger signal of disloyalty than they actually do. Participants in the high loyalty condition anticipated significantly less social rejection than those in the low loyalty condition—suggesting that affirming one's loyalty-based standing can curb signal

amplification and foster more accurate expectations about social reactions to dissent. This study demonstrates that signal amplification bias is not inevitable—it can be attenuated through a brief, targeted intervention. By connecting this dynamic to broader literatures on self-affirmation (Sherman & Cohen, 2006; Critcher & Dunning, 2012) and identity buffering (Walton & Cohen, 2011), Study 5 points to a promising pathway for reducing misperceptions in politically polarized environments.

General Discussion

Across five main studies and five supplemental studies (reported in the OSM), we consistently found that U.S. partisans significantly overestimate how negatively their political ingroup will react to dissenting belief change (Hypothesis 1). The perception gap was large (weighted average effect size across studies 1-5: $d = 0.85$, 95% CI [0.77, 0.94]) and robust across various contexts: It occurred between strangers (Studies 1, 2, 4, 5) and acquaintances (Study 2), and for hypothetical (Studies 1, 2, 3, 4a-c) and actual belief change (Studies 3 and 5). We found evidence for the perception gap across an array of measures, including survey items (Studies 1-5), behavioral measures (Studies 2, 3, and Supplemental Study 1), and coded responses (Study 2). These systematically miscalibrated expectations predicted communication behavior: The more participants anticipated rejection, the more likely they were to self-censor dissenting belief change (Hypothesis 2: Pilot Study 1, Study 4a, Supplemental Study 1).

Why do partisans overestimate social sanctions for dissent? We adapt established theorizing on signal amplification bias (e.g., Vorauer et al., 2003) to the domain of partisan dissent. In this context, dissenters anticipate that their belief change will send a stronger signal of disloyalty than ingroup members actually perceive. Studies 4a-c supported this account by showing that loyalty meta-perceptions mediated predictors' overestimation of rejection—

particularly for partisan topics, where loyalty is especially salient (Hypothesis 3). Study 5 provided experimental support for this mechanism: Affirming group loyalty reduced concerns about appearing disloyal and, in turn, attenuated predictors' overestimation of rejection (Hypothesis 4). Together, these findings underscore partisans' heightened sensitivity to the possibility of signaling disloyalty—reflecting a “better-safe-than-sorry” orientation toward social rejection from their political ingroup in polarized political contexts.

A Distorted Public Sphere

Whereas classic theories of conformity emphasize the role of actual social responses to dissent—such as norm enforcement or social sanctions—as the primary forces driving group conformity (e.g., Jetten & Hornsey, 2014; Kruglanski & Webster, 1991), our findings build on research showing that individuals also conform in anticipation of sanctions (e.g., Miller & Ratner, 2001)—even when those expectations are inaccurate. This pattern builds upon classic theories of pluralistic ignorance (Fields & Schuman, 1976; Miller & Prentice, 1994) and the spiral of silence (Noelle-Neumann, 1993), which describe how people conceal dissenting views out of the mistaken belief that they are alone in holding them. These dynamics can produce a powerful feedback loop: As fewer people voice dissent, the group appears more ideologically uniform—making dissent feel increasingly risky. Recent work underscores the relevance of these processes, showing that partisans vastly underestimate the diversity and strength of attitudes within their own party (Dias, Lelkes, & Pearl, 2024; Brady et al., 2023). In such environments, individuals who privately change their minds may remain silent—not because dissent is necessarily punished, but because of exaggerated expectations that it will be.

Our findings build on these theories and suggest that this self-reinforcing silence is fueled in part by a deeper misperception about how dissent will be interpreted—as a signal of

disloyalty. Study 5 demonstrated that even a brief intervention can reduce reputational concerns and recalibrate expectations of social rejection, offering a foundation for the development of future interventions that promote dissent and foster more open discourse within political groups.

Social Misperceptions

Our findings contribute to a growing literature on political misperceptions by identifying a novel and consequential social misperception at the intragroup level. Whereas most work on political misperceptions has focused primarily on inaccurate beliefs about the outgroup (e.g., Mernyk et al., 2022; Ruggeri et al., 2021), recent evidence suggests that partisans also misperceive norms within their own group—underestimating the diversity and overestimating the extremity of ingroup member beliefs (Brady et al., 2023; Dias et al., 2024). Given that American partisans tend to socialize with (Iyengar et al., 2012; Mason & Wronski, 2018), live near (Brown & Enos, 2021; Motyl et al., 2014), and discuss politics (Mutz, 2006) with fellow ingroup members significantly more often than with outgroup members, the study of intragroup misperceptions is increasingly important. We build on this burgeoning area of research by showing that partisans systematically overestimate how negatively ingroup members will react to dissenting belief change, potentially reinforcing misperceptions through self-censorship.

Our findings also contribute to a broader literature showing that individuals tend to overestimate how negatively they will be evaluated by others in social contexts (e.g., Savitsky et al., 2001; Bruk et al., 2018; Kardas et al., 2024; Ratner & Miller, 2001). We build on this literature by identifying a distinct, group-specific mechanism—signal amplification bias in loyalty meta-perceptions—that helps explain why this miscalibration may be especially acute in contexts of partisan dissent. Unlike prior work, which has largely focused on generalized interpersonal dynamics or intergroup disagreement, our studies draw from research on group

psychology to examine how fear of being perceived as disloyal by one's political ingroup inflates expectations of rejection. In this way, we not only replicate a well-documented social bias but also extend it to a novel group context and identify a psychological mechanism that amplifies it.

These theoretical insights also informed our intervention design. Prior efforts to reduce misperceptions have typically involved providing corrective information—such as revealing actual attitudes or highlighting misperceptions (e.g., Collins et al., 2022; Kardas et al., 2024, Ruggeri et al., 2021). Our studies take a different approach, targeting the underlying psychology that gives rise to misperceptions. Drawing on research from intergroup conflict and attitude change (e.g., Halperin et al., 2011, 2013), we show that a targeted psychological intervention—in our case, affirming group loyalty—can recalibrate misperceptions without the need for external correction. This psychological approach complements existing strategies and offers a promising tool to mitigate misperceptions.

Limitations and Future Research

Despite the robustness and consistency of the present findings, this research possesses several limitations. First, it was conducted in the specific political and cultural context of the contemporary United States—a setting marked by high affective polarization. Generalizability to other political systems or cultural environments remains an open question. Second, these studies primarily involved rejection dynamics between strangers or loosely connected individuals. Although this design afforded strong experimental control, future research should explore these dynamics within close relationships, where reputational stakes are higher and communication patterns more complex. Third, our studies focused on the specific context of dyadic interactions, thereby limiting the generalizability of these findings to larger group settings.

Promising next steps include exploring these dynamics in high-stakes settings such as organizations, families, or online communities. Although increased familiarity can reduce meta-perceptual bias by increasing cue availability and shared interpretive frameworks (Funder, 1995; Elswaady & Carlson, 2023), it is also possible that the perception gap could intensify in these contexts due to heightened concern about relational fallout. Additionally, future work could explore whether belief change prompted by different sources (e.g., emotion vs. reason) or targeting different levels of identity (e.g., issue-level vs. partisan identity-level change) elicits different patterns of anticipated and actual rejection. Finally, future research could examine whether similar overestimation patterns emerge for dissent within other groups formed around a shared belief system (e.g., religious, military), and whether loyalty-affirmation interventions are similarly effective in these settings.

Conclusion

Our research sheds light on the dynamics of social judgment and conformity within political groups, revealing a robust perception gap between expected and actual social reactions to dissenting belief change. We show that heightened concern about how dissent will be interpreted—as a signal of disloyalty—can inflate expectations of social rejection and, in turn, stifle dissent before it is ever voiced. These findings offer new insight into the ways that covert social pressures reinforce ideological conformity, even in the absence of explicit enforcement. By addressing the psychological roots of these misperceptions, organizations and communities may be better positioned to encourage dissent, reduce false consensus, and foster political discourse that is both more inclusive and more representative.

References

- Allport, F. H. (1924). *Social Psychology*. Houghton Mifflin.
- Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23(4), 363–377.
- Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1.
- Bakker, B. N., Lelkes, Y., & Malka, A. (2020). Understanding partisan cue receptivity: Tests of predictions from the bounded rationality and expressive utility perspectives. *The Journal of Politics*, 82(3), 1061–1077.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529.
- Baumeister, R. F., & Tice, D. M. (1990). Anxiety and social exclusion. *Journal of Social and Clinical Psychology*, 9(2), 165-.
- Beisswanger, A. H., Stone, E. R., Hupp, J. M., & Allgaier, L. (2003). Risk taking in relationships: Differences in deciding for oneself versus for a friend. *Basic and Applied Social Psychology*, 25(2), 121–135.
- Boothby, E. J., Clark, M. S., & Bargh, J. A. (2014). Shared experiences are amplified. *Psychological Science*, 25(12), 2209-2216.
- Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M., & Crockett, M. J. (2023). Overperception of moral outrage in online social networks inflates beliefs about

intergroup hostility. *Nature Human Behaviour*, 7(6), 917–927.

Briñol, P., McCaslin, M. J., & Petty, R. E. (2012). Self-generated persuasion: effects of the target and direction of arguments. *Journal of Personality and Social Psychology*, 102(5), 925.

Brooks, A. W., & Schweitzer, M. E. (2011). Can Nervous Nelly negotiate? How anxiety causes negotiators to make low first offers, exit early, and earn less profit. *Organizational Behavior and Human Decision Processes*, 115(1), 43–54.

Brown, J. R., & Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*, 5(8), 998–1008.

Bruk, A., Scholl, S. G., & Bless, H. (2018). Beautiful mess effect: Self–other differences in evaluation of showing vulnerability. *Journal of Personality and Social Psychology*, 115(2), 192.

Carlsmith, J. M., Collins, B. E., & Helmreich, R. L. (1966). Studies in forced compliance: The effect of pressure for compliance on attitude change produced by face-to-face role playing and anonymous essay writing. *Journal of Personality and Social Psychology*, 4(1), 1.

Cavazza, N., Pagliaro, S., & Guidetti, M. (2014). Antecedents of concern for personal reputation: The role of group entitativity and fear of social exclusion. *Basic and Applied Social Psychology*, 36(4), 365–376.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55(1), 591–621.

Collins, H. K., Dorison, C. A., Gino, F., & Minson, J. A. (2022). Underestimating counterparts' learning goals impairs conflictual conversations. *Psychological Science*, 33(10), 1732–1752.

Connors, E. C. (2020). The social dimension of political values. *Political Behavior*, 42(3), 961–982.

Correll, J., & Park, B. (2005). A model of the ingroup as a social resource. *Personality and Social Psychology Review*, 9(4), 341–359.

Critcher, C. R., & Dunning, D. (2015). Self-affirmations provide a broader perspective on self-threat. *Personality and Social Psychology Bulletin*, 41(1), 3-18.

Dias, N. C., Aarslew, L. F., Frederiksen, K. V. S., Lelkes, Y., Pradella, L., & Westwood, S. J. (2024). Correcting misperceptions of partisan opponents is not effective at treating democratic ills. *PNAS Nexus*, 3(8), pgae304. <https://doi.org/10.1093/pnasnexus/pgae304>

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. A. (2017). *The social psychology of prosocial behavior* (1st ed.). Psychology Press.
<https://doi.org/https://doi.org/10.4324/9781315085241>

Druckman, J. N., & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1), 114–122.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science¹. *Behavioral and Brain Sciences*, 38, e130.

Elsaadawy, N., & Carlson, E. N. (2022). Is meta-accuracy consistent across levels of acquaintanceship?. *Social Psychological and Personality Science*, 13(1), 178-185.

Epley, N., Kardas, M., Zhao, X., Atir, S., & Schroeder, J. (2022). Undersociality: miscalibrated social cognition can inhibit social connection. *Trends in Cognitive Sciences*, 26(5), 406–

418.

Epley, N., Keysar, B., Boven, L. Van, & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339.

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. *Journal of Experimental Psychology: General*, *143*(5), 1980.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–210.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160.

Fields, J. M., & Schuman, H. (1976). Public beliefs about the beliefs of the public. *Public Opinion Quarterly*, *40*(4), 427-448.

Fernbach, P. M., & Van Boven, L. (2022). False polarization: Cognitive mechanisms and potential solutions. *Current Opinion in Psychology*, *43*, 1–6.

<https://doi.org/https://doi.org/10.1016/j.copsyc.2021.06.005>

Funder, D. C. (Ed.). (1999). *Personality judgment: A realistic approach to person perception*. Elsevier.

Funkhouser, E. (2022). A tribal mind: Beliefs that signal group identity or commitment. *Mind & Language*, *37*(3), 444-464.

Gaertner, S. L., Dovidio, J. F., Guerra, R., Hehman, E., & Saguy, T. (2015). A common ingroup

identity: Categorization, identity, and intergroup relations. In *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 433–454). Psychology Press.

Galak, J., & Critcher, C. R. 2022. Who sees which political falsehoods as more acceptable and why: A new look at in-group loyalty and trustworthiness. *Journal of Personality and Social Psychology*, 124(3): 593-619.

Gilbert, D. T. (2002). Inferential correction. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 167–184). Cambridge University Press.

Gilovich, T., Kruger, J., & Medvec, V. H. (2002). The spotlight effect revisited: Overestimating the manifest variability of our actions and appearance. *Journal of Experimental Social Psychology*, 38(1), 93–99.

Gilovich, T., Kruger, J., & Savitsky, K. (1999). Everyday egocentrism and everyday interpersonal problems. In *The social psychology of emotional and behavioral problems: Interfaces of social and clinical psychology* (pp. 69–95). American Psychological Association.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, 78(2), 211.

Glynn, C. J., Hayes, A. F., & Shanahan, J. (1997). Perceived support for one's opinions and willingness to speak out: A meta-analysis of survey studies on the "spiral of silence". *Public Opinion Quarterly*, 452–463.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person

perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148–168.
<https://doi.org/10.1037/a0034726>

Greenwald, A. G., & Albert, R. D. (1968). Acceptance and recall of improvised arguments. *Journal of Personality and Social Psychology*, 8.

Gunther Moor, B., Crone, E. A., & van der Molen, M. W. (2010). The heartbrake of social rejection: Heart rate deceleration in response to unexpected peer rejection. *Psychological Science*, 21(9), 1326–1333.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon/Random House.

Halperin, E., Pliskin, R., Saguy, T., Liberman, V., & Gross, J. J. (2014). Emotion regulation and the cultivation of political tolerance: Searching for a new track for intervention. *Journal of Conflict Resolution*, 58(6), 1110-1138.

Halperin, E., Russell, A. G., Trzesniewski, K. H., Gross, J. J., & Dweck, C. S. (2011). Promoting the Middle East peace process by changing beliefs about group malleability. *Science*, 333(6050), 1767-1769.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, 78(1), 81–91.
<https://doi.org/10.1037/0022-3514.78.1.81>

Hawkins, S., Yudkin, D., Juan-Torres, M., & Dixon, T. (2018). *Hidden Tribes: A Study of America's Polarized Landscape*. New York: More in Common.

- Heltzel, G., & Laurin, K. (2021). Seek and ye shall be fine: Attitudes toward political-perspective seekers. *Psychological Science*, 32(11), 1782–1800.
- Henrich, J. (2015). Culture and social behavior. *Current Opinion in Behavioral Sciences*, 3, 84–89.
- Higgins, E. T. (1996). Activation: Accessibility, and salience. *Social psychology: Handbook of basic principles*, 133-168.
- Hogg, M. A. (2016). *Social identity theory*. Springer.
- Hussein, M. A., & Wheeler, S. C. (2024). Reputational costs of receptiveness: When and why being receptive to opposing political views backfires. *Journal of Experimental Psychology: General*.
- Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405–431.
- Jetten, J., & Hornsey, M. J. (2014). Deviance and dissent in groups. *Annual Review of Psychology*, 65(1), 461–485.
- Kardas, M., Kumar, A., & Epley, N. (2024). Let it go: How exaggerating the reputational costs of revealing negative information encourages secrecy in relationships. In *Journal of Personality and Social Psychology* (Vol. 126, Issue 6, pp. 1052–1083). American Psychological Association. <https://doi.org/10.1037/pspi0000441>
- Kossowska, M., Czarnek, G., & Szwed, P. (2023). Political ideology and belief change in the face of counterevidence. *European Journal of Social Psychology*, 53(6), 1157–1171.
- Kruglanski, A. W., Pierro, A., Mannetti, L., & Grada, E. De. (2006). Groups as Epistemic

- Providers: Need for Closure and the Unfolding of Group-Centrism. *Psychological Review*, *113*(1), 84–100.
- Kruglanski, A. W., & Webster, D. M. (1991). Group members' reactions to opinion deviates and conformists at varying degrees of proximity to decision deadline and of environmental noise. *Journal of Personality and Social Psychology*, *61*(2), 212–225.
- Kunst, J. R., Thomsen, L., & Dovidio, J. F. (2019). Divided loyalties: Perceptions of disloyalty underpin bias toward dually-identified minority-group members. *Journal of Personality and Social Psychology*, *117*(4), 807–838.
- Leary, M. R., & Downs, D. L. (1995). Interpersonal functions of the self-esteem motive: The self-esteem system as a sociometer. In *Efficacy, agency, and self-esteem* (pp. 123–144). Springer.
- Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279–286.
- Levendusky, M. S., & Malhotra, N. (2016). (Mis) perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, *80*(S1), 378–391.
- Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88–106.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267.
- MacDonald, G., & Leary, M. R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, *131*(2), 202–223.

<https://doi.org/10.1037/0033-2909.131.2.202>

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology, 39*, 257–277.

Matthes, J., Knoll, J., & von Sikorski, C. (2017). The “spiral of silence” revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research, 45*(1), 3–33.

<https://doi.org/10.1177/0093650217745429>

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 31*(3), 227-242.

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans’ support for partisan violence. *Proceedings of the National Academy of Sciences, 119*(16), e2116851119.

Miller, D. T. (2023). A century of pluralistic ignorance: what we have learned about its origins, forms, and consequences. *Frontiers in Social Psychology, 1*, 1260896.

Miller, D. T., & Prentice, D. A. (1994). Collective errors and errors about the collective. *Personality and Social Psychology Bulletin, 20*(5), 541-550.

Miller, D. T., & Ratner, R. K. (2001). The norm of self-interest and its effects on social action. *Journal of Personality and Social Psychology, 81*(1), 5–16.

Moore, M., Dorison, C. A., & Minson, J. A. (2023). The contingent reputational benefits of selective exposure to partisan information. In *Journal of Experimental Psychology: General* (Vol. 152, Issue 12, pp. 3490–3525). American Psychological Association.

<https://doi.org/10.1037/xge0001463>

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology, 51*, 1–14.

Moy, P., Domke, D., & Stamm, K. (2001). The spiral of silence and public opinion on affirmative action. *Journalism & Mass Communication Quarterly, 78*(1), 7–25.

Mutz, D. C. (2006). Hearing the other side: Deliberative versus participatory democracy. *Cambridge University*.

Nemeth, Charlan Jeanne, Connell, J. B., Rogers, J. D., & Brown, K. S. (2001). Improving decision making by means of dissent. *Journal of Applied Social Psychology, 31*(1), 48–58.

Nickerson, R. S. (1999). How we know and sometimes misjudge what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*(6), 737–759.

Noelle-Neumann, E. (1993). *The spiral of silence: Public opinion--Our social skin*. University of Chicago Press.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

OpenAI. (2024). *ChatGPT* (July 2024). <https://chat.openai.com/chat>

Piaget, J. (1926). *The child's conception of the world*. Routledge and Keegan Paul.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381.

Ratner, R. K., & Miller, D. T. (2001). The norm of self-interest and its effects on social

- action. *Journal of Personality and Social Psychology*, 81(1), 5-16.
- Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge* (1st ed., pp. 103–135). Psychology Press.
- Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalioğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., ... Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*, 5(10), 1369–1380. <https://doi.org/10.1038/s41562-021-01092-x>
- Savitsky, K., Epley, N., & Gilovich, T. (2001). Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps. *Journal of Personality and Social Psychology*, 81(1), 44.
- Schachter, S. (1951). Deviation, rejection, and communication. *The Journal of Abnormal and Social Psychology*, 46(2), 190.
- Scheufele, D. A., Shanahan, J., & Lee, E. (2001). Real talk: Manipulating the dependent variable in spiral of silence research. *Communication Research*, 28(3), 304–324.
- Shaw, A., DeScioli, P., Barakzai, A., & Kurzban, R. (2017). Whoever is not with me is against me: The costs of neutrality among friends. *Journal of Experimental Social Psychology*, 71, 96–104
- Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, 38, 183-242.
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self.

In *Advances in experimental social psychology* (Vol. 21, pp. 261-302). Academic Press.

Swann, W. B., Jetten, J., Gómez, Á., Whitehouse, H., & Bastian, B. (2012). When group membership gets personal: A theory of identity fusion. *Psychological Review*, *119*(3), 441–456.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984-16.

Tappin, B. M., Berinsky, A. J., & Rand, D. G. (2023). Partisans' receptivity to persuasive messaging is undiminished by countervailing party leader cues. *Nature Human Behaviour*, *7*(4), 568–582.

Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. In H. Høgh-Olesen (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 91–234). Palgrave Macmillan.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Bavel, J., & Packer, D. J. (2021). *The Power of Us: Harnessing Our Shared Identities for Personal and Collective Success*. Hachette UK.

Van Bavel, J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224.

Van Vugt, M., & Hart, C. M. (2004). Social identity as social glue: the origins of group loyalty. *Journal of Personality and Social Psychology*, *86*(4), 585.

Vorauer, J. D., Cameron, J. J., Holmes, J. G., & Pearce, D. G. (2003). Invisible overtures: Fears

of rejection and the signal amplification bias. *Journal of Personality and Social Psychology*, 84(4), 793.

Vorauer, J. D., & Sakamoto, Y. (2006). I thought we could be friends, but... Systematic miscommunication and defensive distancing as obstacles to cross-group friendship formation. *Psychological Science*, 17(4), 326-331.

Wald, K. A., Kardas, M., & Epley, N. (2024). Misplaced divides? Discussing political disagreement with strangers can be unexpectedly positive. *Psychological Science*, 35(5), 471–488.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, 331(6023), 1447-1451.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, 58, 425–452.

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.

Xu, M., & Petty, R. E. (2022). Two-sided messages promote openness for morally based attitudes. *Personality and Social Psychology Bulletin*, 48(8), 1151–1166.

Yudkin, D., Hawkins, S., & Dixon, T. (2019). *The perception gap: How false impressions are pulling Americans apart*. More In Common.

Zdaniuk, B., & Levine, J. M. (2001). Group loyalty: Impact of members' identification and contributions. *Journal of Experimental Social Psychology*, 37(6), 502-509.