PREPRINT — This paper has been accepted for publication in Journal of Personality and Social Psychology. This version may differ from the final published article.

Overestimating the Social Costs of Political Belief Change

Trevor Spelman[1]

Abdo Elnakouri[2]

Nour Kteily[1,3]

Eli J. Finkel[1,3,4,5]

[1]Northwestern University, Kellogg School of Management

[2]University of Houston, Department of Psychology

[3]Litowitz Center for Enlightened Disagreement, Northwestern University

[4]Northwestern University, Department of Psychology

[5]Northwestern University, Institute for Policy Research

Corresponding Author: Please address all correspondence for this article to Trevor Spelman at trevor.spelman@kellogg.northwestern.edu.

DOI: https://dx.doi.org/10.1037/pspi0000516

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, may reflect 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 was associated with 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.

Keywords: Social misperceptions; dissent; rejection; self-censorship; social conformity

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 change[1], 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. As a result, 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

[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."

information environment will falsely reflect 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 have been shown to 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; Jordan & Kteily, 2023)—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 and 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 political ingroup members will reject them for dissenting belief change on partisan political issues.*

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 dissenting belief change.

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; see also Supplemental Study 1 and Study 4a in the OSM). Next, we tested whether these expectations align with how ingroup members actually react (Studies 1-5). Finally, we examined whether signal amplification bias—specifically, the tendency to expect that dissenting belief change will send a stronger signal of disloyalty than observers actually perceive—can help to explain 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-registered[2]. 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 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. We surveyed U.S. partisans who reported dissenting belief change on a political issue within the past year (N = 131) and asked them how they expect other political ingroup members would react to their belief change (e.g., "How much do you think another [Democrat/Republican] would reject you?"; 1 = *Not at all,* 4 = *A moderate amount*, 7 = *Very much*) and whether they were more or less likely to self-censor their updated views

[2] Preregistrations can be accessed here: https://researchbox.org/3336

when in conversation with other ingroup members (1 = *I am LESS likely to share my views,* 4 = *I am equally likely to share my views on this topic as I was before I changed my mind,* 7 = *I am MORE likely to share my views*)[3]. We found that participants who anticipated harsher social sanctions for their dissenting belief change (M = 3.61, SD = 1.63) also reported a 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.

Study 1

Given the relationship between anticipated social sanctions and self-censorship of dissenting belief change, an important question is whether individuals' 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

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 interaction effect (f = 0.2) between role and topic conditions in a 2 (Role: predict vs. react) × 3 (Topic: abortion, gun control, immigration) between-subjects ANOVA with 90% power at α = .05. Given the novelty of this research area, and our aim to examine findings across multiple political

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

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) × 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 party[4], 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 party[5].

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 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

[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.

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[6].

Finally, participants responded to attention checks and demographic questions, including age, gender, ethnicity, and U.S. zip code.[7]

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) × 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 1). The main effect of topic was not significant, $F(2, 494) = 2.04$, $p = .131$, $\eta p^2 = .01$, nor was the

[6] Respondents also completed an original four-item measure of social acceptance (e.g., "accept," "include," "treat," and "view you as a member of the group"). These items were exploratory and were included only in Study 1. The results showed the same pattern as the rejection composite—predictors underestimated acceptance relative to reactors—although the gap was smaller in magnitude. Full item wording and results are reported in OSM.

[7] Following the dependent measures, participants responded to additional survey items that were included for a separate research project.

interaction between role and topic, $F(2, 494) = 0.15$, $p = .864$, $\eta p^2 = .000$, suggesting that the effect of role on rejection did not differ significantly by topic condition.

Next, we conducted a 2 (Role: predict vs. react) × 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.

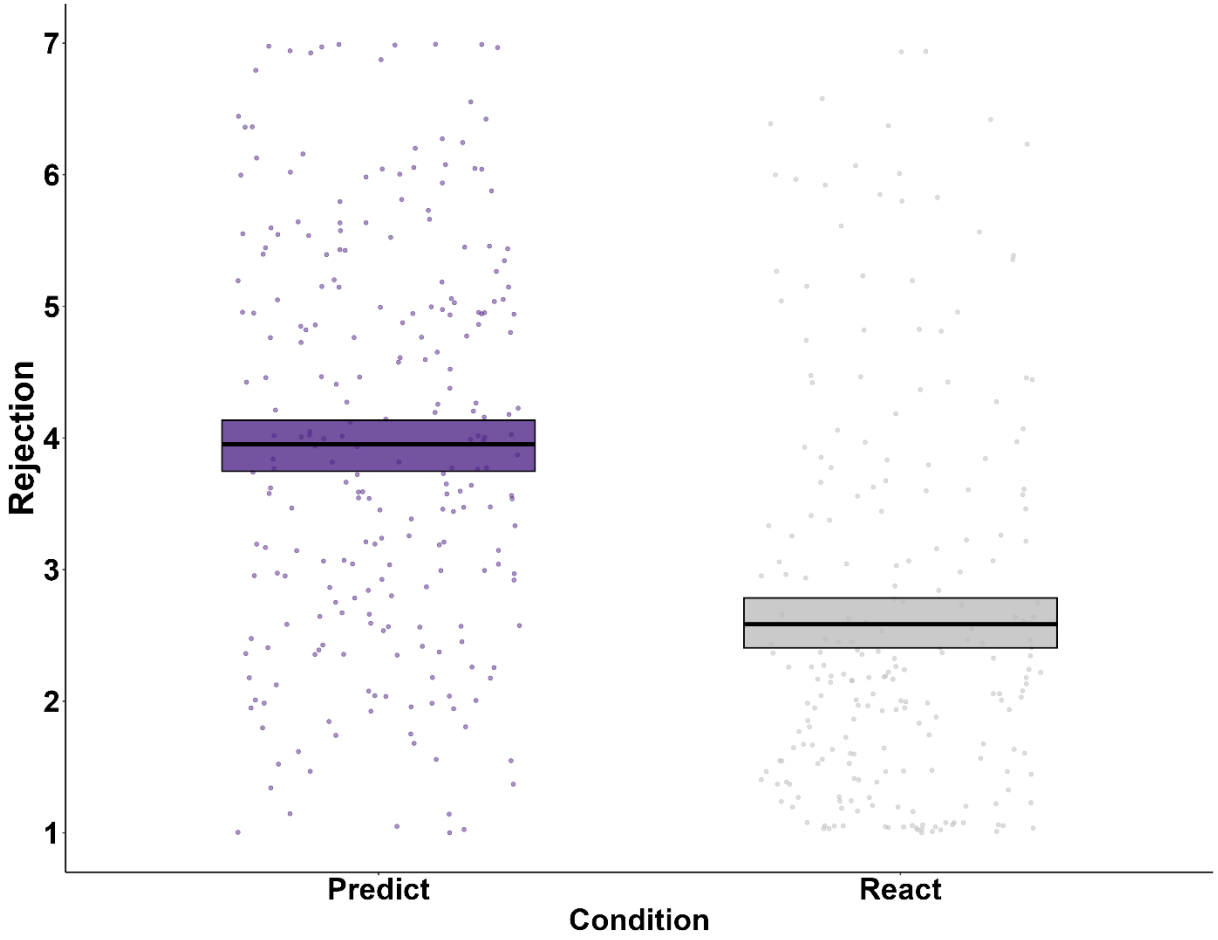


Figure 1. *Predictors overestimate rejection relative to reactors' reports. Note*. The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence

interval. Anticipated and actual social rejection were measured on a seven-point response scale (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*).

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.[8] 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

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

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 using G*Power 3.1 (Faul et al., 2009) indicated that a sample of at least 300 participants was required 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 between predict and react 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)[9] 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 they 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 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

[9] 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.

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 questions[10]. 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 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 parties' 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

[10] 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.

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)[11].

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

[11]In 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 × 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.

(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 sub-sample 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 SUPPLEMENTAL MATERIAL 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 accepting[12]). Specifically, reactors were incentivized to act in ways that reflected

[12] 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.

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.



Figure 2. *Coded open-ended responses revealed that predictors expected their partner to view them more negatively after learning about their dissent than reactors actually reported. Note.* The plot shows the density distribution of coded scores for predictor and reactor responses. Each response was rated on a seven-point response scale (-3 = *Very negative*, 0 = *Neutral,* 3 = *Very positive*). 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 study[13]; 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; 241 females, 242 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

[13] 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 SUPPLEMENTAL MATERIAL 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 political 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 SUPPLEMENTAL MATERIAL 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, $F(2, 436) = 5.60$, $p = .004$, $\eta p^2 = .025$. Follow-up tests of the simple effects 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 ($M = 2.74$, $SD = 1.63$), $t(436) = 2.70$, $p = .007$, $d = .40$. In contrast, participants in the Phase 2 non-salient condition did not overestimate reactor social sanctions ($M = 2.31$, $SD = 1.48$), $t(436) = -1.66$, $p = .099$, $d = .27$ (Figure 3).

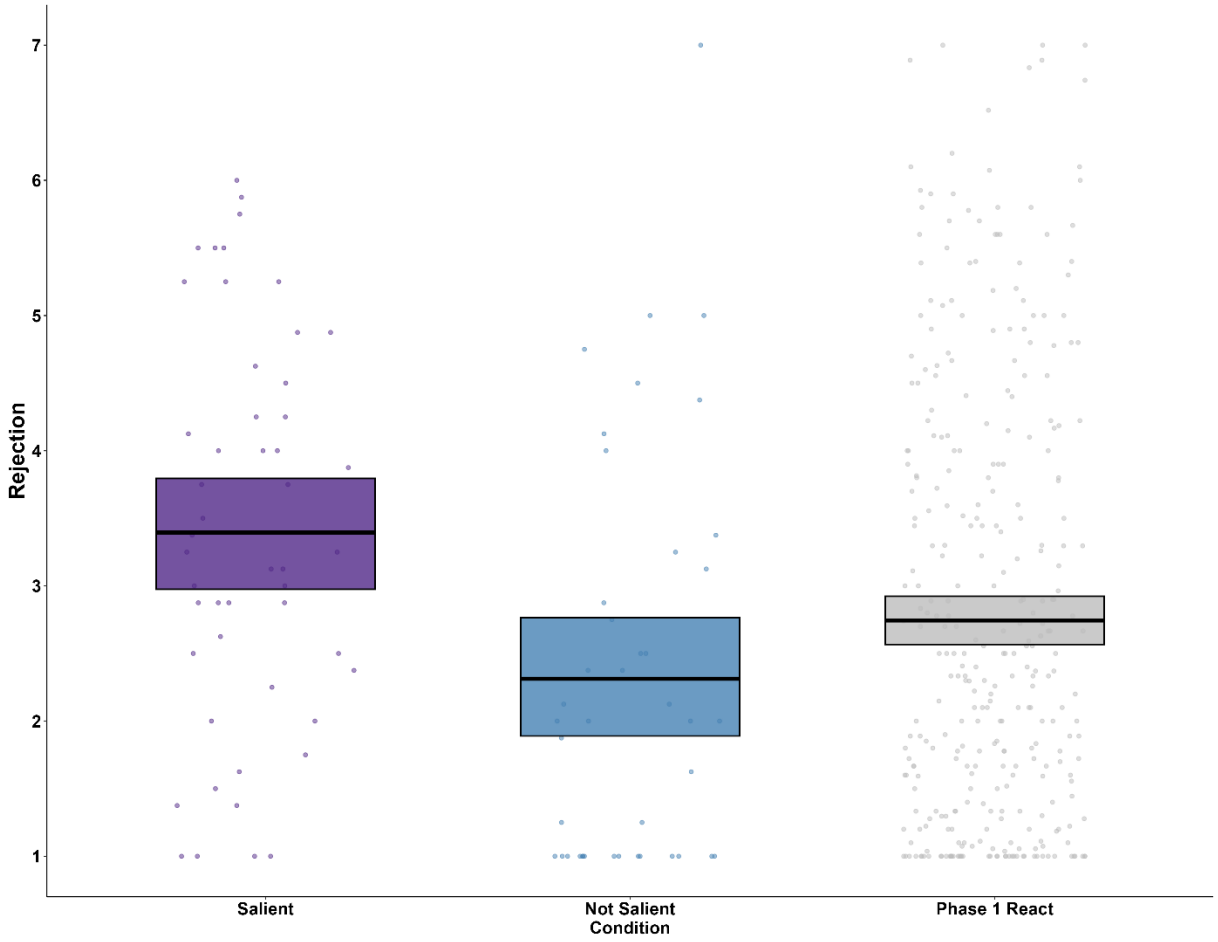


Figure 3. *Predictors reminded of their belief change overestimated social sanctions. Those not reminded did not. Dissenting predictors whose belief change was not made salient did not. Note.* The data in this plot are slightly jittered to avoid overlap. The height of the boxes represents the 95% confidence interval. Anticipated and actual social rejection were measured on a seven-point response scale (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*).

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 potential psychological mechanisms.

We begin by testing whether this misperception 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; Wald et al., 2024; 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.

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, 2020; Funkhouser, 2022; Galak & Critcher, 2023). 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, which could partly explain their 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.

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 these studies 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).

Methods

Across Studies 4a–4c, we tested when and why individuals overestimate social sanctions for dissenting belief change. Each study employed the same 2 (Role: predictor vs. reactor) × 2 (Topic: partisan vs. non-partisan) between-subjects design and tested the same two hypotheses: (i) that predictors would anticipate stronger social sanctions than reactors actually report, and that this difference (the "perception gap") would be significantly larger for partisan topics compared to non-partisan topics; and (ii) that this interaction would be explained by perceptions of ingroup loyalty through a moderated mediation model in which role (predictor = 1, reactor = 0) predicted social sanctions through loyalty judgments, with topic condition (partisan = 1, non-partisan = 0) moderating the indirect effect.

The paradigm was consistent across studies. Predictors were paired with a political ingroup member and imagined changing their mind on a political issue to adopt an opposing view, thereby resulting in disagreement with their assigned partner. They then estimated how their ingroup partner would evaluate them. Reactors evaluated a described ingroup member who had changed their mind to adopt the opposing view on the same issue.

All studies measured anticipated and actual social sanctions using the same five-item measure from Study 1. They also assessed perceived ingroup loyalty, capturing either predictors' meta-perceptions of how loyal they expected to be seen or reactors' actual judgments of the target's loyalty. Predictors rated two items assessing how their belief change would affect an ingroup peer's view of their loyalty (e.g., "This person will think I am loyal to the [ingroup] Party"; "This person will think I am likely to vote for [ingroup]s in future elections", 1 = *Definitely not*, 7 = *Definitely yes*). Reactors completed parallel items referring to their own judgments of the target's loyalty.

Design Differences Across Studies. Although all three studies used the same 2 × 2 design and tested identical hypotheses, Study 4a differed from Studies 4b and 4c (which were nearly identical) in how topic partisanship was operationalized. In Study 4a, participants considered either a partisan or a non-partisan issue drawn from distinct topic sets, allowing us to compare belief change across issues that naturally varied in political divisiveness while holding the act of dissent constant. 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").

Whereas Study 4a relied on different topics that were either partisan or non-partisan, Studies 4b and 4c, by contrast, held the topic constant and experimentally framed it as either partisan or non-partisan. The issues used in Studies 4b and 4c ("The use of artificial intelligence in government surveillance programs should be [**banned / permitted]**" and "Digitized gambling should be [**banned / permitted]** nationwide") were selected based on pre-test data which suggested that Americans do not have strong priors about whether Democrats and Republicans hold a typical position on these issues (full pre-test details are reported in the SUPPLEMENTAL MATERIAL). We experimentally manipulated perceived partisanship by telling participants that the issue is either highly polarized or not very polarized between Democrats and Republicans. This framing manipulation isolated the perceived partisanship of the topic while keeping its

substantive content fixed. We included a manipulation check item, which asked participants how much other ingroup members support a given position on the issue (1 = *Strongly oppose;* 4 = *Neither oppose nor support;* 7 = *Strongly support)*, and an item at the end of the survey which asked participants how much they believed the manipulation information (1 = *Not at all;* 4 = *Moderately*; 7 = *Very much*). The full item wording and results are reported in the SUPPLEMENTAL MATERIAL.

Finally, Studies 4a and 4b were not pre-registered, whereas Study 4c was pre-registered. Study 4c served as a direct preregistered, higher-powered replication of Study 4b, save for some minor procedural adjustments described in the SUPPLEMENTAL MATERIAL. Each of these three studies is presented independently in full detail in the SUPPLEMENTAL MATERIAL.

Power Analysis. Across Studies 4a–4c, sample sizes were determined using a priori power analyses targeting the 2 × 2 interaction between role (predictor vs. reactor) and topic conditions (partisan vs. non-partisan) on the social-sanctions composite, tested via two-way ANOVA using Type III sum of squares. We used G*Power 3.1 (Faul et al., 2009) to calculate required sample sizes, specifying a small-to-moderate interaction effect size of $f = 0.17$ ($\eta p^2 \approx .03$), with $\alpha = .05$ and power = .90. This effect size was selected based on pilot data, theoretical expectations, and conventional benchmarks for interaction effects in social psychological research (Cohen, 1988).

In Study 4a, the observed interaction effect was $\eta p^2 = .03$ ($f \approx .17$), which was used as the target for Study 4b. For the preregistered replication in Study 4c, we adopted a more conservative approach by powering to detect the lower bound of the confidence interval around the effect observed in Study 4b ($\eta p^2 = .017$). Final sample sizes for studies 4a-c were $N = 393$,

282, and 596, respectively. Additional information about power calculations is described in the SUPPLEMENTAL MATERIAL.

Participants. Participants across Studies 4a–4c were U.S.-based Democrats and Republicans recruited from Prolific Academic. Demographic information for each sample is presented in the SUPPLEMENTAL MATERIAL.

Analytic Approach. Analyses followed a plan that was identical across studies. We first tested the Role × Topic Condition interaction on the social-sanctions composite using two-way ANOVAs with Type III sums of squares followed by tests of simple effects. We then conducted moderated mediation analyses using structural equation modeling (SEM) in lavaan (R), with 5,000 bootstrap samples to estimate indirect effects. All models used the same variable coding and parameterization: role (predictor = 1, reactor = 0) as the independent variable, topic condition (partisan = 1, non-partisan = 0) as the moderator, loyalty as the mediator, and social sanctions as the dependent variable.

Results

Manipulation and Believability Checks. In both Studies 4b and 4c, the partisan framing manipulation was effective: participants in the partisan condition perceived their ingroup's position as more politically divided than did those in the non-partisan condition (Study 4b: M = 6.78 vs. 4.09, $F(1, 278) = 375.19$, $p < .001$, $\eta p^2 = .57$; Study 4c: $M = 6.43$ vs. 3.37, $F(1, 592) = 340.33$, $p < .001$, $\eta p^2 = .37$). The manipulation's believability was generally high across studies (Study 4b: $M = 5.00$, $SD = 1.70$; Study 4c: $M = 4.77$, $SD = 1.71$). In Study 4b, believability did not differ across conditions ($ps > .65$), whereas in Study 4c, predictors in the non-partisan condition rated the manipulation as less believable than those in the partisan condition (M = 4.37

vs. 5.11, $F(1, 592) = 14.03$, $p < .001$, $\eta p^2 = .02$). Full results and additional analyses using believability as a control variable for our main hypothesis tests are reported in the SUPPLEMENTAL MATERIAL.

Main Analyses. Across studies 4a-c, the 2 (Role: predictor vs. reactor) × 2 (Topic: partisan vs. non-partisan) interaction on social sanctions was significant in all three datasets (see Table 1). Predictors consistently anticipated stronger social sanctions than reactors reported, and this perception gap was significantly larger in the partisan topic condition.

Table 1. *Two-Way ANOVA Results for the Effects of Role and Topic Condition on Social Sanctions in Studies 4a–4c.*

				Partisan		Non-Partisan	
Study	***F*(1, *df*)**	***p***	***ηp²***	**Predict *M* (*SD*)**	**React *M* (*SD*)**	**Predict *M* (*SD*)**	**React *M* (*SD*)**
4a	11.21 (1, 389)	< .001	0.03	4.68 (1.45)	2.87 (1.74)	2.48 (1.45)	1.67 (1.17)
4b	8.34 (1, 278)	.004	0.029	4.54 (1.49)	2.56 (1.64)	3.08 (1.47)	2.18 (1.62)
4c	7.14 (1, 592)	.008	0.012	4.42 (1.58)	2.74 (1.64)	3.13 (1.48)	2.10 (1.37)

Note. Social sanctions were assessed using a five-item composite (αs ≥ .96). All studies used 2 (Role: predictor vs. reactor) × 2 (Topic: partisan vs. non-partisan) between-subjects ANOVAs with Type III sums of squares. The *F* statistics, *p*-values, and *ηp²* statistics refer to the interaction effect. In all studies, predictors anticipated greater social sanctions than reactors reported, and this difference was larger in the partisan condition. Anticipated and actual social sanctions were measured on a seven-point response scale (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*). The full model output for each study is reported in the SUPPLEMENTAL MATERIAL.

We next examined whether these effects were explained by perceived ingroup loyalty. Moderated mediation analyses tested whether role (predictor = 1, reactor = 0) predicted social sanctions through loyalty, with topic condition (partisan = 1, non-partisan = 0) moderating the indirect path (see Figure 4). Results for each study are shown in Table 2 below.

In Studies 4a and 4b, the results supported the proposed loyalty-based signal amplification account: predictors in the partisan condition expected to be seen as less loyal than reactors actually judged, and we found evidence for a significant indirect effect from the larger perception gap for partisan topics to predicted social sanctions via loyalty misperceptions. In

both studies, the indirect effect of role on sanctions through loyalty was significantly stronger in the partisan condition, consistent with a loyalty-focused form of signal amplification bias.

In Study 4c, which replicated the design of Study 4b with greater statistical power, the role × condition interaction on loyalty did not reach significance. As shown in Table 2, the magnitude of this interaction has varied across Studies 4a–4c. It was modest in Study 4a, substantially larger in Study 4b, and absent in Study 4c—even as the downstream effect of loyalty on social sanctions remained consistently strong across studies. Although the moderated mediation was not supported in Study 4c, the basic mediation pathway remained robust: Across both topic conditions, predictors expected greater social sanction than reactors reported, with a significant indirect effect linking role to social sanction via lower perceived loyalty scores.

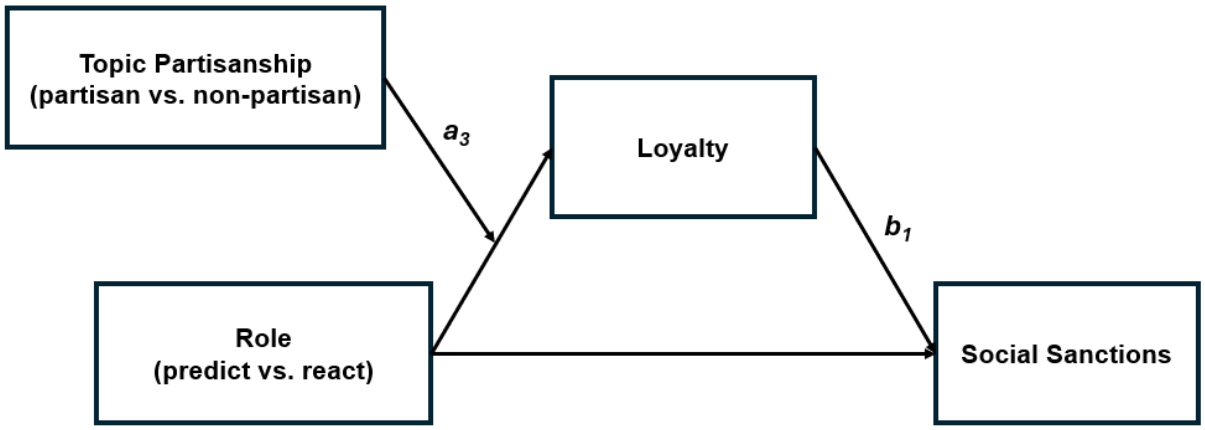


Figure 4. Conceptual Model of the Moderated Mediation Tested in Studies 4a–4c.
The model tests whether the indirect effect of role (predictor = 1, reactor = 0) on social sanctions through loyalty varies as a function of topic partisanship (partisan = 1, non-partisan = 0). The interaction of role × topic partisanship (path a_3) predicts loyalty, and loyalty in turn predicts social sanctions (path b_1). Unstandardized coefficients reported in Table 2 were estimated with 5,000 bootstrap samples in *lavaan* (R).

Table 2. *Moderated Mediation of Role on Social Sanctions Through Loyalty as a Function of Topic Partisanship Condition (Studies 4a–4c).*

Study	Path	*b*	*SE*	*z*	*p*	95% CI
4a	a_3: Role × Topic → Loyalty	–0.58	0.27	–2.11	0.035	[–1.11, –0.04]
	b_1: Loyalty → Sanctions	–0.55	0.05	–10.28	< .001	[–0.65, –0.45]
	Indirect (partisan context)	0.37	0.12	3.22	0.002	[0.14, 0.61]
	Indirect (non-partisan context)	0.06	0.10	0.58	0.563	[–0.12, 0.25]
	Index of moderated mediation	0.32	0.15	2.11	0.035	[0.02, 0.62]
4b	a_3: Role × Topic → Loyalty	–1.26	0.34	–3.75	< .001	[–1.93, –0.59]
	b_1: Loyalty → Sanctions	–0.52	0.07	–7.76	< .001	[–0.65, –0.39]
	Indirect (partisan context)	0.70	0.15	4.83	< .001	[0.40, 1.00]
	Indirect (non-partisan context)	0.05	0.13	0.40	0.687	[–0.20, 0.29]
	Index of moderated mediation	0.65	0.19	3.44	0.001	[0.27, 1.03]
4c	a_3: Role × Topic → Loyalty	0.23	0.23	1.00	0.317	[–0.22, 0.68]
	b_1: Loyalty → Sanctions	–0.47	0.05	–10.11	< .001	[–0.56, –0.38]
	Indirect (partisan context)	0.33	0.09	3.80	< .001	[0.16, 0.51]
	Indirect (non-partisan context)	0.44	0.08	5.21	< .001	[0.28, 0.60]
	Index of moderated mediation	–0.11	0.11	–0.97	0.331	[–0.33, 0.11]

Note. All indirect effects were estimated using 5,000 bootstrap samples in lavaan (R). Role (predictor = 1, reactor = 0) served as the independent variable; topic partisanship condition (partisan = 1, non-partisan = 0) was the moderator; loyalty was the mediator; and social sanctions were the dependent variable. The conditional indirect effects are computed as ($a_1 \times b_1$) for non-partisan topics and ($(a_1 + a_3) \times b_1$) for partisan topics; the index of moderated mediation equals $a_3 \times b_1$. The index of moderated mediation tests whether the indirect effect differs significantly between topic conditions (Hayes, 2015). Loyalty meta-perceptions and perceptions were measured on a seven-point response scale (1 = *Definitely not*, 7 = *Definitely yes*), as were anticipated and actual social sanctions (1 = *Not at all,* 4 = *A moderate amount,* 7 = *Very much*). For tests of the distal interaction effect (Role × Topic Partisanship → Social Sanctions), see Table 1.

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.

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 was not replicated. Several factors may account for this discrepancy. In Study 4c, predictors rated the framing manipulation significantly less believable in the non-partisan condition compared to predictors in 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 a smaller manipulation check effect size ($\eta p^2 = .37$ in Study 4c vs. $\eta p^2 = .57$ in Study 4b)—which, in combination with the discrepant believability ratings, may have attenuated the role × topic condition interaction on loyalty ($\eta p^2 = .002$ vs. $\eta p^2 = .048$ in Study 4b).

Our indirect effect analyses should be interpreted with caution. Whereas role and topic conditions were experimentally manipulated, loyalty and sanction expectations were measured and were assessed concurrently in these studies. Thus, as is true of all indirect effect analyses relying on cross-sectional correlational data, the causal ordering between these constructs cannot be definitively determined. Although our findings are consistent with the hypothesis that concerns about appearing disloyal contribute to inflated expectations of rejection, it is also plausible that anticipating rejection could reduce perceived loyalty. In the final study, we build on this evidence by experimentally manipulating the conditions theorized to give rise to signal amplification bias, providing a causal test of our theoretically proposed mechanism.

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 were associated with higher anticipated rejection, Study 5 examined whether reminding predictors of their past loyalty would mitigate 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, 2005; Steele, 1988), which suggest that activating valued aspects of one's identity can reduce defensiveness in identity-relevant domains.

Method

Participants. We recruited participants who reported changing their mind on one of five pre-specified partisan political topics[14] 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 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

[14] 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.

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 to compare social sanction scores between groups with 90% power at $\alpha = 0.05$ (Contrast 1: Predict versus React; Contrast 2: Predict High versus Predict Low). 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 occur[15].

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.

[15] 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 from 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 SUPPLEMENTAL MATERIAL)—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 SUPPLEMENTAL MATERIAL.

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, $F(2, 617) = 23.50$, $p < .001$, $\eta p^2 = .071$. Follow-up

tests of the simple effects 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 post-hoc pairwise comparison, we also examined reactor loyalty scores, treating them as a baseline control condition in which loyalty was measured without any experimental treatment. Reactors reported feelings of loyalty that were 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, $t(617) = 4.76$, $p < .001$, $d = 0.46$. Participants in the high-loyalty condition reported significantly greater feelings of loyalty than reactors, $t(617) = 2.15$, $p = .032$, $d = 0.21$.

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, $F(2, 617) = 13.80$, $p < .001$, $\eta p^2 = .043$. Follow-up tests of the simple effects 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.

Auxiliary analyses revealed that predictors expected to be seen as less loyal than they felt in both the high and low loyalty conditions. In the high-loyalty condition, participants reported feeling significantly more loyal ($M = 5.70$, $SD = 1.02$) than they expected to be seen by others ($M = 4.64$, $SD = 1.40$), $t(195) = 9.61$, $p < .001$, 95% CI [0.85, 1.28], $d = 0.69$. Similarly, in the low-loyalty condition, participants also felt more loyal ($M = 4.91$, $SD = 1.32$) than they expected to be seen ($M = 4.13$, $SD = 1.44$), $t(203) = 6.59$, $p < .001$, 95% CI [0.55, 1.02], $d = 0.46$. Although predictors in both conditions felt more loyal than they expected to be seen, the magnitude of this self–other loyalty gap did not differ significantly by condition (M_{high} = 1,06; $M_{low} = 0.78$), Welch's $t(396.88) = 1.72$, $p = .086$. Together, these results suggest that the loyalty affirmation intervention was associated with more accurate loyalty meta-perceptions in the high-loyalty condition relative to the low-loyalty condition, despite a persistent gap between felt loyalty and loyalty meta-judgments in both groups.

Main Analyses. 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 predictors' overestimation of rejection was explained in part by their expectations of being 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 is associated with less 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 may 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 could be accounted for by shifts in loyalty meta-perceptions, we conducted a mediation analysis in which the experimental condition (high vs. low loyalty) predicted loyalty meta-perceptions, which in turn were associated with rejection expectations. The indirect effect was significant, ACME = –0.12, 95% CI [–0.20, –0.06], $p < .001$, suggesting that affirming feelings of ingroup loyalty increased participants' expectations of being seen as loyal by others, which was in turn linked to lower 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], $p = .021$, indicating a residual effect of the manipulation beyond changes in loyalty meta-perceptions. Together, these findings are consistent with the interpretation that loyalty meta-perceptions not only relate to anticipated rejection but also may shift in response to interventions that affirm feelings of group loyalty. Reflecting on past acts of loyalty may have reduced the perceived social-evaluation threat associated with dissent, helping to explain why participants in the high-loyalty condition anticipated less rejection from others.[16].

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 help to explain the overestimation of social sanctions for dissenting belief change. By increasing

[16] 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. We also conducted several competing mediation models to test the robustness of loyalty meta-perceptions against other theoretically derived mediators. See Study 5 Additional Analyses in the OSM for all additional analyses.

participants' felt security in their loyalty-based standing, the intervention appeared to buffer against signal amplification bias and was associated with a narrower gap between anticipated and actual social sanctions.

Discussion

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 may help to attenuate signal amplification, which in turn is associated with 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, 2014) 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 SUPPLEMENTAL MATERIAL), 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 = 0.87, 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 findings reported in the SUPPLEMENTAL MATERIAL under Additional Analyses, and Supplemental Study 1).

Why do partisans overestimate social sanctions for dissent? We adapt established theorizing on signal amplification bias to the domain of partisan dissent. Our use of the term signal amplification bias extends previous applications of the concept—typically focused on exaggerated positive signals of social interest (e.g., initiating friendship, romantic overtures; Vorauer et al., 2003; Vorauer & Sakamoto, 2006)—to contexts involving negative signaling concerns. In the present 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 indirect effects via loyalty meta-perceptions helped to explain 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, was associated with a narrower 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. 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 may be explained in part by a deeper misperception about how dissent will be interpreted.

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., Lees & Cikara, 2020; 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-based 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 helps to explain why the overestimation is amplified in partisan political contexts.

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. That said, reinforcing a sense of group loyalty may have unintended side effects, such as increasing ingroup identification, which could potentially reinforce partisan hostility and bias (e.g., Iyengar & Westwood, 2015; Van Bavel & Pereira, 2018). Future research should examine

whether similar accuracy gains can be achieved through interventions that reduce defensive threat without simultaneously heightening group attachment.

Limitations and Future Research

Despite the robustness and consistency of the present findings, this research has 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; Elsaadawy & 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.

Another important direction will be to test whether the attenuation of overestimation observed in Study 5 is unique to loyalty affirmation or reflects a broader affirmation-based

process. Although affirming one's loyalty was associated with less anticipated rejection, other forms of affirmation (e.g., value or self-integrity affirmation) may similarly reduce defensive concern about social evaluation. 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—is associated with inflated expectations of social rejection, which, in turn, predicts the self-censorship of dissent. 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 pluralistic ignorance, and foster political discourse that is both more inclusive and more representative.

References

Allport, F. H. (1924). Social psychology. Houghton Mifflin.

Argyle, L. P., & Freeze, M. (2024). The role of self-threat and self-affirmation in initiation of political conversations. American Politics Research, 52(6), 655–670. https://doi.org/10.1177/1532673X241263079

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. https://doi.org/10.1177/0146167297234003

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. https://doi.org/10.1086/707616

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. https://doi.org/10.1037/0033-2909.117.3.497

Baumeister, R. F., & Tice, D. M. (1990). Anxiety and social exclusion. Journal of Social and Clinical Psychology, 9(2), 165–195. https://doi.org/10.1521/jscp.1990.9.2.165

Berry, Z., Lewis, N. A., & Sowden, W. J. (2021). The double-edged sword of loyalty. Current Directions in Psychological Science, 30(4), 321–326. https://doi.org/10.1177/09637214211010759

Boothby, E. J., Cooney, G., Sandstrom, G. M., & Clark, M. S. (2018). The liking gap in

conversations: Do people like us more than we think? Psychological Science, 29(11), 1742–1756. https://doi.org/10.1177/0956797618783714

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. https://doi.org/10.1038/s41562-023-01582-0

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), 925–940. https://doi.org/10.1037/a0027231

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. https://doi.org/10.1016/j.obhdp.2011.01.008

Brown, J. R., & Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. Nature Human Behaviour, 5(8), 998–1008. https://doi.org/10.1038/s41562-021-01066-z

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), 192–205. https://doi.org/10.1037/pspa0000120

Carlsmith, J. M., Collins, B. E., & Helmreich, R. L. (1966). Studies in forced compliance. I. 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), 1–13. https://doi.org/10.1037/h0023507

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. https://doi.org/10.1080/01973533.2014.925453

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum.

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. https://doi.org/10.1177/09567976221085494

Connors, E. C. (2020). The social dimension of political values. Political Behavior, 42(3), 961–982. https://doi.org/10.1007/s11109-019-09530-3

Correll, J., & Park, B. (2005). A model of the ingroup as a social resource. Personality and Social Psychology Review, 9(4), 341–359. https://doi.org/10.1207/s15327957pspr0904_4

Critcher, C. R., & Dunning, D. (2014). Self-affirmations provide a broader perspective on self-threat. Personality and Social Psychology Bulletin, 41(1), 3–18. https://doi.org/10.1177/0146167214554956

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), Article 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. Psychology Press. https://doi.org/10.4324/9781315085241

Dunning, D. (2005). Self-insight: Roadblocks and detours on the path to knowing thyself. Psychology Press. https://doi.org/10.4324/9780203337998

Elsaadawy, N., & Carlson, E. N. (2022a). Do you make a better or worse impression than you think? Journal of Personality and Social Psychology, 123(6), 1407–1420. https://doi.org/10.1037/pspp0000434

Elsaadawy, N., & Carlson, E. N. (2022b). Is meta-accuracy consistent across levels of acquaintanceship? Social Psychological and Personality Science, 13(1), 178–185. https://doi.org/10.1177/19485506211018151

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. https://doi.org/10.1016/j.tics.2022.02.007

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, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

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. https://doi.org/10.3758/BRM.41.4.1149

Fields, J. M., & Schuman, H. (1976). Public beliefs about the beliefs of the public. Public Opinion Quarterly, 40(4), 427–448. https://doi.org/10.1086/268330

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of

personality: Interpersonal relations, similarity, and acquaintanceship. Journal of Personality and Social Psychology, 69(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Funkhouser, E. (2022). A tribal mind: Beliefs that signal group identity or commitment. Mind & Language, 37(3), 444–464. https://doi.org/10.1111/mila.12326

Gaertner, S. L., Dovidio, J. F., Guerra, R., Hehman, E., & Saguy, T. (2015). A common ingroup identity: Categorization, identity, and intergroup relations. In T. D. Nelson (Ed.), Handbook of prejudice, stereotyping, and discrimination (pp. 433–454). Psychology Press. https://doi.org/10.4324/9781003399162

Galak, J., & Critcher, C. R. (2023). 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. https://doi.org/10.1037/pspi0000264

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), 211–222. https://doi.org/10.1037/0022-3514.78.2.211

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, 61(3), 452–463. https://doi.org/10.1086/297808

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(1), 31–34. https://doi.org/10.1037/h0021237

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

Halperin, E., Porat, R., Tamir, M., & Gross, J. J. (2012). Can emotion regulation change political attitudes in intractable conflicts? From the laboratory to the field. Psychological Science, 24(1), 106–111. https://doi.org/10.1177/0956797612452572

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. https://doi.org/10.1126/science.1202925

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. More in Common. https://doi.org/10.31234/osf.io/xz25v

Hayes, A. F. (2015). An index and test of linear moderated mediation. Multivariate Behavioral Research, 50(1), 1–22. https://doi.org/10.1080/00273171.2014.962683

Heltzel, G., & Laurin, K. (2021). Seek and ye shall be fine: Attitudes toward political-perspective seekers. Psychological Science, 32(11), 1782–1800. https://doi.org/10.1177/09567976211011969

Henrich, J. (2015). Culture and social behavior. Current Opinion in Behavioral Sciences, 3, 84–89. https://doi.org/10.1016/j.cobeha.2015.02.001

Higgins, E. T. (1996). Activation: Accessibility, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), Social psychology: Handbook of basic principles (pp. 133–168). Guilford Press.

Hildreth, J. A. D., Gino, F., & Bazerman, M. (2016). Blind loyalty? When group loyalty makes us see evil or engage in it. Organizational Behavior and Human Decision Processes, 132, 16–36. https://doi.org/10.1016/j.obhdp.2015.10.001

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, 153(6), 1425–1448. https://doi.org/10.1037/xge0001579

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. Public Opinion Quarterly, 76(3), 405–431. https://doi.org/10.1093/poq/nfs038

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. American Journal of Political Science, 59(3), 690–707. https://doi.org/10.1111/ajps.12152

Jetten, J., & Hornsey, M. J. (2014). Deviance and dissent in groups. Annual Review of Psychology, 65(1), 461–485. https://doi.org/10.1146/annurev-psych-010213-115151

Jordan, J. J., & Kteily, N. S. (2023). How reputation does (and does not) drive people to punish without looking. Proceedings of the National Academy of Sciences of the United States of America, 120(28), Article e2302475120. https://doi.org/10.1073/pnas.2302475120

Kardas, M., Kumar, A., & Epley, N. (2024). Let it go: How exaggerating the reputational costs of revealing negative information encourages secrecy in relationships. Journal of Personality and Social Psychology, 126(6), 1052–1083. 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. https://doi.org/10.1002/ejsp.2965

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. Psychological Review, 113(1), 84–100. https://doi.org/10.1037/0033-295X.113.1.84

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. https://doi.org/10.1037/0022-3514.61.2.212

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. https://doi.org/10.1037/pspi0000168

Lasswell, H. D. (1936). Politics: Who gets what, when, how. McGraw-Hill.

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. Nature Human Behaviour, 4(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. Organizational Behavior and Human Decision Processes, 126, 88–106. https://doi.org/10.1016/j.obhdp.2014.10.007

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(S1), 257–277. https://doi.org/10.1111/pops.12485

Matthes, J., Knoll, J., & von Sikorski, C. (2018). 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. https://doi.org/10.1177/0956797619900321

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 of the United States of America, 119(16), Article e2116851119. https://doi.org/10.1073/pnas.2116851119

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

https://doi.org/10.3389/frsps.2023.1260896

Miller, D. T., & Prentice, D. A. (1994). Collective errors and errors about the collective. Personality and Social Psychology Bulletin, 20(5), 541–550. https://doi.org/10.1177/0146167294205011

Moore, M., Dorison, C. A., & Minson, J. A. (2023). The contingent reputational benefits of selective exposure to partisan information. Journal of Experimental Psychology: General, 152(12), 3490–3525. 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. https://doi.org/10.1016/j.jesp.2013.10.010

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. https://doi.org/10.1177/107769900107800102

Mutz, D. C. (2006). Hearing the other side: Deliberative versus participatory democracy. Cambridge University. https://doi.org/10.1017/CBO9780511617201

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

OpenAI. (2024, July). ChatGPT. https://chat.openai.com/chat

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. https://doi.org/10.1177/0146167202286008

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. https://doi.org/10.1037/0022-3514.81.1.5

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed, & E. Turiel (Eds.), Values and knowledge (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., Bursalıoğlu, 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), 44–56. https://doi.org/10.1037/0022-3514.81.1.44

Schachter, S. (1951). Deviation, rejection, and communication. Journal of Abnormal Psychology, 46(2), 190–207. https://doi.org/10.1037/h0062326

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. https://doi.org/10.1177/009365001028003003

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. https://doi.org/10.1016/j.jesp.2017.03.002

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. Advances in Experimental Social Psychology, 38, 183–242. https://doi.org/10.1016/S0065-2601(06)38004-5

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 21, pp. 261–302). Academic Press. https://doi.org/10.1016/S0065-2601(08)60229-4

Stokes, D. E. (1963). Spatial models of party competition. In A. Campbell, P. E. Converse, W. E. Miller, & D. E. Stokes (Eds.), The American voter (pp. 368–377). Wiley.

Swann, W. B., Jetten, J., Gómez, A., Whitehouse, H., & Bastian, B. (2012). When group membership gets personal: A theory of identity fusion. Psychological Review, 119(3), 441–456. https://doi.org/10.1037/a0028589

Tajfel, H., & Turner, J. C. (2000). An integrative theory of intergroup conflict. In M. J. Hatch & M. Schultz (Eds.), Organizational identity: A reader (pp. 56–65). Oxford University Press. https://doi.org/10.1093/oso/9780199269464.003.0005

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. https://doi.org/10.1038/s41562-023-01551-7

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. 191–234). Palgrave Macmillan. https://doi.org/10.1007/978-1-137-05001-

4_8

Van Bavel, J., & Packer, D. J. (2021). The power of us: Harnessing our shared identities for personal and collective success. Hachette.

Van Bavel, J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. Trends in Cognitive Sciences, 22(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

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), 585–598. https://doi.org/10.1037/0022-3514.86.4.585

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), 793–812. https://doi.org/10.1037/0022-3514.84.4.793

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. https://doi.org/10.1111/j.1467-9280.2006.01706.x

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. https://doi.org/10.1177/09567976241230005

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.

https://doi.org/10.1126/science.1198364

Williams, K. D. (2007). Ostracism. Annual Review of Psychology, 58(1), 425–452. https://doi.org/10.1146/annurev.psych.58.110405.085641

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), 1–14. https://doi.org/10.1177/25152459231213802

Xu, M., & Petty, R. E. (2022). Two-sided messages promote openness for morally based attitudes. Personality and Social Psychology Bulletin, 48(8), 1151–1166. https://doi.org/10.1177/0146167220988371

Zdaniuk, B., & Levine, J. M. (2001). Group loyalty: Impact of members' identification and contributions. Journal of Experimental Social Psychology, 37(6), 502–509. https://doi.org/10.1006/jesp.2000.1474