2023

The Power of Dissent: Mitigating False Polarization and Cross-Party Dislike in Online Interactions

Victoria Parker
park5210@mylaurier.ca

The Power of Dissent: Mitigating False Polarization and Cross-Party Dislike in Online Interactions

by Victoria A. Parker

Honours Bachelor of Arts, Psychology, Wilfrid Laurier University, 2013

Master of Arts, Social Psychology, Wilfrid Laurier University, 2018

A DISSERTATION

Submitted to the Faculty of Psychology at Wilfrid Laurier University

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy in Social Psychology

Wilfrid Laurier University

Acknowledgements

I would like to first formally thank my funding sources for providing the resources needed to conduct this research. The contributions made by the Canadian Institute for Advanced Research (CIFAR), the Social Sciences and Humanities Research Council (SSHRC), and the Foundational Integrity Research Grant provided by Meta, were all integral in the completion of this project, and my doctoral program overall. I would like to express my sincerest gratitude for the guidance, mentorship, and support of my brilliant advisor, Dr. Anne E. Wilson. My development as a scholar, a scientist, and an adult, is largely due to her encouragement and advice, and I cannot thank her enough. I would also like to thank Dr. Frank Kachanoff, and Dr. Christian Jordan for being a part of my defense committee and offering their suggestions and recommendations not only for this document, but for future research projects and my professional development as well. Finally, I would like to thank my loved ones, particularly my husband, who have provided endless emotional support, providing much of the strength and fortitude required to carry one through years of graduate school. And thank you to the littlest scientist who has yet to arrive; you have been a gift to carry during the final stretch of my doctoral program, and I cannot wait to introduce you to the world. My most heartfelt thank you to everyone mentioned; I am forever appreciative of your contributions, and your impact on my academic career.

Abstract

While actual polarization is on the rise in the United States, perceived polarization (i.e., false polarization) is growing at an even faster rate, contributing to increased cross-party hostility. A meaningful amount of out-party dislike may be produced by partisans' dramatic overestimates of the prevalence of extreme, undesirable views among political opponents. In the current research, we examine whether exposing people to out-party dissenters who challenge their copartisans' extreme views might help reduce people's misperceptions of their opponents' extreme views, and possibly mitigate animosity. Across five studies (N = 3789), we explore how seeing public ingroup dissent (in the form of responses to an extreme tweet) changes the (mis)perceived prevalence of the extreme attitude amongst the opponent group. For both liberals and conservatives, seeing an interaction wherein a single political opponent disagreed with a (presumed) widely held extreme tweet lowered their estimates of how prevalent that attitude was, compared to seeing the tweet alone (Studies 1 and 2). This effect was strengthened when participants saw a "dogpile" of dissent, where multiple out-party members dissented against the extreme out-party tweet, compared to a single response, or no response at all (Studies 3, 4, & 5). The dissent condition did not directly affect feelings towards opponents, or willingness to engage with them. However, a serial mediation model revealed that exposure to dissent indirectly affected willingness to engage by reducing prevalence (over)estimates and, in turn, liking. Specifically, participants who saw opponent dissent lowered their estimates of opponent agreement with the extreme tweet; lower estimates were related to more positive feelings towards opponents, and more positive feelings were subsequently related to greater willingness to engage across the aisle.

Keywords: False polarization, misperceptions, dissent, social media, intergroup perceptions

The Power of Dissent: Mitigating False Polarization and Cross-Party Dislike in Online Interactions

America is divided; hostility between political parties in the United States has reached a boiling point after years of steady escalation. Rising polarization is evident to politicians, pundits, and partisans, as commentary on the heated nature of the current political climate is abundant. Rising polarization has also been revealed empirically in research tracking both ideological polarization (i.e., the difference between opponents' policy positions), as well as affective polarization (i.e., intense dislike and dehumanization of the opponent) (Abramowitz & Webster, 2018; Crawford et al., 2013; Pew Research Center, 2014; Westfall et al., 2015). Although the intensification of America's political rift is undeniably real, there is also evidence that at least *some* of this divide are in people's heads, a phenomenon referred to as *false* polarization. False polarization can manifest in a number of ways; people assume that the ideologies of opposing groups are more extreme than they actually are (Chambers et al., 2006; Levendusky & Malhotra, 2016; Westfall et al., 2015), that the identities of those within groups are different than reality (Ahler & Sood, 2018), and that their opponents perceive *them* more negatively than their opponents actually do (Lees & Cikara, 2020, 2021). Further, individuals often believe that far more of their opponents support caricatured, extreme, egregious views than is actually the case (Parker et al., 2023). Indeed, it is often these caricatured representations of liberals and conservatives that are the "version" individuals have in mind when asked what sorts of attitudes their political opponents endorse, even when the majority of them would disagree with this characterization. Further, people's cross-party dislike appears to be more strongly predicted by these extreme views that most opponents do not actually hold than by the more moderate views most opponents actually espouse.

To be sure, dislike for opponents who *actually* hold noxious views may well be warranted. But the evidence suggests that much of the intense animosity held between political rivals may be due to *incorrect* beliefs about what kind of attitudes the opponent endorses. If this is truly the case, then changing these incorrect beliefs about what views are prevalent amongst political adversaries may offer a potential means of mitigating division. Past research has shown that, for example, correcting incorrect

meta-perceptions, which are perceptions regarding the ways your opponent feels about *you* (Lees & Cikara, 2021), or incorrect assumptions about the demographic makeup of a political party (i.e., overestimating the proportion of the party that is part of the LGBTQ community, or making over $200,000 per year, Ahler, 2018) can soften cross-party dislike. While neither of these approaches target the caricatured, stereotyped misperceptions of the opponent specifically, they do demonstrate that misperceptions can be, at least partially, remedied. Extending on this past work, we reasoned that the negative effects of encountering an extreme out-partisan view (like overestimating its prevalence) could be countered by exposing individuals to dissenting out-partisan views (i.e., showing people examples that go *against* the stereotype). We reasoned that this misperception may be mitigated by exposing people to other opponents who hold contrary or dissenting partisan views, challenging the perception that the noxious view is prevalent and underlining the heterogeneity within the group.

Media as a Source of Misperceptions

Partisans believe that extreme and noxious views are far more prevalent amongst their political opponents than is the case (Parker et al., 2023). Where do these beliefs come from? In our previous work (Parker et al., 2023), we found that false polarization arose when participants were asked to estimate the prevalence of the *extreme* views of their opponents, as they believed those views were more widespread than they were in actuality. This was not the case when asked to estimate the prevalence of their opponents' more moderate, policy-type views. We argued that this was, in part, because more extreme, caricatured representations of opponents were most likely to be featured in the mainstream media. Indeed, exposure to partisan media sources was consistently linked to misperceptions; we observed that partisan media consumption (e.g., Fox News and MSNBC) predicted the perception that one's political opponent held egregious views. Specifically, tuning into MSNBC was consistently related to the perception that conservatives were deplorable bigots, while watching Fox News was related to perceiving liberals as free speech hating snowflakes. Partisan media is heavily incentivized to perpetuate these narratives as a means to capture and retain audience engagement. Over time, media outlets have proliferated, competing in an "attention economy" (Hendricks & Vestergaard, 2019) where profit is tied to the capacity to attract

attention, clicks, and engagement time. Because emotionally evocative, often negative content attracts and holds attention, media outlets have come to prioritize reporting stories that incite of fear, anger, and outrage (Brock & Rabin-Havt, 2012; Klein, 2020) all in the name retaining viewership, in a media model described as the "outrage industry" (Berry & Sobieraj, 2013). Partisan news will often frame stories in incendiary and combative ways, like characterizing all liberals as rabidly politically correct, and all conservatives as gun-toting racists in order to elicit audience outrage and maintain attention. While some stories may be fabricated, more often the narrative is supported with highly selective footage of actual (though relatively rare) incidents that exemplify those characteristics.

Partisan misperceptions are not only disseminated through traditional media sources. These stereotyped characterizations are also coming from, and are being circulated across social media platforms, where individuals espousing incendiary views often get the most attention (Brady et al., 2017; Crockett, 2017). Indeed, some of the most ideologically extreme individuals are among the most active social media users (Gaisbauer et al., 2021), and likely contribute to the impression that more extreme views are widespread as those are the types of views that promote engagement (e.g., liking, sharing, etc.), but which are, in reality, only supported by a minority of actual partisans. This is especially likely given that those with more moderate opinions often choose to stay silent online (Bail, 2021), leaving platforms open to the proliferation of extreme thoughts and opinions which may go unchallenged. Thus, people's miscalibrated prevalence estimates are likely partially derived from their disproportionate exposure to extreme views online. Moreover, previous work on the availability heuristic demonstrates that rare but vivid exemplars are often misjudged as frequent because of their availability in memory (Manis et al., 1993; Tversky & Kahneman, 1973). We reasoned that extreme political exemplars may represent a view held only by a minority of copartisans, yet because of their vividness and salience they may become the most available representation of the group as a whole, inflating the perceived prevalence of those attitudes. If these are the views an individual sees the most often online, and these are the views they are predisposed to remember, we speculate that multiple factors may then converge to reinforce misperceptions of the prevalence of opponents' worst views. We know from our previous work that there

are, in fact, individuals within groups that do endorse extreme attitudes, and it is likely those extreme individuals are sharing opinions on social media, highlighting themselves as memorable, vivid exemplars. Is it possible to prevent those extreme users from fostering the perception that their attitudes are widespread? We looked to past interventions on the corrections of misperceptions to inform the current approach.

Targeting and Correcting Opponent Misperceptions

The growing divide between liberals and conservatives appears to be, at least partially, driven by illusory differences (in perceived extremity, in party make-up, etc.) perpetuated across both traditional and social media platforms, but is this fissure irreparable? Some researchers have sought to improve the current state of affective polarization between parties by specifically targeting misperceptions (e.g., the degree to which they are disliked and dehumanized by opponents), providing individuals with a more accurate picture of their political opponents. This type of intervention, wherein participants are provided with accurate information regarding the thoughts and feelings of their opponents, appears promising. For example, partisans often have misperceptions regarding how much their opponents dislike (Lees & Cikara, 2020; Moore-Berg et al., 2020) and dehumanize them (Landry et al., 2022), a type of misperception referred to as a meta-misperception. Correcting these meta-misperceptions, by giving participants accurate information regarding how their opponents view them, has successfully reduced negative outgroup perceptions, as well as support for partisan violence (Mernyk et al., 2022).

However, this intervention approach is not always successful (Parker & Wilson, unpublished data). In one of our previous studies, participants were asked to estimate the percentage of opponents who endorsed an egregious attitude, and, as expected, vastly overestimated the proportion of opponents who did, compared to the number who actually reported agreeing. Participants were then given a graph that represented the proportion of opponents who had endorsed the attitude in past studies, as a means of correcting their misperception and improving opponent dislike. However, showing participants the actual percentage of the opponent group that endorses extreme attitudes had no impact on their reported opponent dislike. In fact, participants reported a fair amount of skepticism that the statistics we were

providing were legitimate, highlighting the entrenched nature of these beliefs. It appears that sometimes, statistics are an effective intervention strategy, and other times they are not. The mixed effectiveness of the approach may be partly because individuals can be more strongly influenced by vivid exemplars (i.e., an individual saying or doing something emotionally evocative or memorable) compared to similar issues conveyed in statistical form (De Wit et al., 2008; Kogut & Ritov, 2005), with individuals often viewing personal experiences as more true than facts (Kubin et al., 2021). Perhaps a more consistently effective approach to correcting opponent-targeted misperceptions could be achieved by presenting participants with a vivid exemplar who actively *counters* an exemplar that reflects an extreme view that is typically associated with the opponents' side. Indeed, one of the most effective approaches in changing individuals' stereotypical beliefs about outgroup members (e.g., negative prejudices against Black people) is by providing them with exemplars who go *against* the prevailing stereotype (e.g., thinking about President Obama as a positive Black exemplar; Columb & Plant, 2011; see FitzGerald et al., 2019 for review).

If social media is, indeed, contributing to the false perception that political adversaries are monstrous, perhaps it can also be used to help correct the misperceptions they may be exacerbating. What if, instead of just seeing an incendiary post from an opponent on social media, individuals were also exposed to opponents expressing explicit disagreement with the post? It is possible that seeing dissent from other opponents may make participants perceive the attitude expressed in the post as less prototypical of the opponent group, and thus less representative of the group's views. Seeing other members of the opponent party speak up against egregious views (specifically over social media) might encourage individuals to see their opposing group as more heterogeneous (Brauer & Er-rafiy, 2011; Ostrom & Sedikides, 1992), as the counter-stereotypical examples might force them to consider that perhaps not everyone in the outgroup endorses the same attitudes and beliefs.

How might an extreme exemplar be countered in a way that is realistic and models believable intragroup behaviour? One way is to depict instances of within group dissent. Although there can often be pressures to conform with ingroup views (see Jetten & Hornsey, 2014, and Packer, 2007 for review), ingroup dissenters serve an important function of course-correction, often protecting the group from

engaging in harmful behaviours, or endorsing harmful attitudes (Packer, 2007). Another (to date unexplored) function of ingroup dissenters is their effect on group perception - the tendency to paint an entire party with the same (extreme) brush is likely to be challenged when dissenters challenge the group's extreme views and highlight the heterogeneity of opinion within a group.

Goals of the Current Research

Across five studies, we investigated a number of aspects of this online dissent dynamic, specifically wherein individuals dissented against the extreme views of a copartisan. In study 1 and 2, we examined how encountering a vivid exemplar of an extreme view from an opponent (specifically in the form of a tweet) changed the perceived prevalence of the view. To do this, we compared the extreme tweet on its own to the same tweet followed by either agreement or dissent from another copartisan. In the next set of studies, we investigated the effect of introducing a single versus multiple dissenters following an extreme tweet on the perceived prevalence of the tweeted view, specifically to strengthen the effects of the dissenting outgroup members. In a final study, we controlled for baseline liking prior to the presentation of stimuli. Across the five studies, we examined several different political issues to increase generalizability.

We hypothesize that exposing participants to opponents dissenting against extreme views being posted on Twitter by another opponent (compared to extreme posts without dissent), would reduce how prevalent participants believed the egregious view was amongst the outgroup. For example, we predict that a liberal participant who is shown a racist conservative tweet where another conservative publicly responds with disagreement, will feel that the racist view is less prevalent among conservatives than if they saw a racist conservative tweet with no responses at all. Further, we have no specific reason to expect the effect to differ for liberals versus conservatives; thus we expect the effects of opponent dissent to be equally effective regardless of partisanship.

Further, we hypothesize that exposing participants to posts containing opponent dissent (compared to posts without dissent) would improve feelings towards opponents, as the misperceptions that drive dislike would be immediately countered. This correction, we speculate, may subsequently

encourage engaging with opponents in various ways (i.e., having a political discussion; Parker et al, 2023). In these ways, merely seeing opponent dissent could begin to bridge the widening divide between parties.

Study 1 and Study 2

The first two studies were similar in design and will be discussed together. We explored how responses to an extreme tweet from a political opponent affected partisans' estimates of the prevalence with which opponents hold those extreme views.[1] Specifically, we sought to investigate whether the perceived prevalence of opponents' extreme views (expressed in the form of an extreme tweet) can be altered by exposure to a *second* tweet from an opponent group member that either supported – or dissented from -- the original extreme view. Further, we speculated that when participants were exposed to an outgroup member dissenting against the extreme tweet, they would feel more positively towards them (liking and trusting them more) than when exposed to an outgroup member expressing agreement (or seeing no response at all).

Methods

Participants

American participants from Amazon Mechanical Turk (MTurk) were tested in both studies 1 and 2; a breakdown of the demographics can be found in Table 1. We aimed to recruit approximately equal numbers of liberals and conservatives in both studies. Participants who failed the attention check (S1: n = 30, S2: n = 27) or identified as "both equally" politically (i.e., identified as both liberal and conservative equally) (S1: n = 24, S2: n = 58) were excluded from analysis. In the current manuscript, the research design of studies 1 and 2 was a 2 (partisanship: liberal or conservative) x 3 (condition: agreement response, dissenting response, or no response) factorial design, in which partisans were always exposed to tweets and responses of political outgroup members. These data were collected as part of a larger research

[1] There was also a third study, run just after the first two, wherein the civility of the dissenting response was also manipulated (to be civil vs. uncivil). However, as our later research did not explore that variable further, it has been dropped to retain the focus and flow in this current paper.

design in studies 1,2, 4, and 5, in which participants were exposed to either in-party or out-party tweets in a fully crossed design. Because exposure to in-party tweets addresses a separate set of research questions they are not analyzed here.[2] Therefore, the demographics reported in the table represent only the participants used in analyses.

Given the sample size used in analyses, an alpha level of .05, and minimum power of .80 (Cohen, 1992), sensitivity analyses using GPower indicated that, for our primary method of analysis, an ANOVA with main effects and interactions, we could detect an effect size of at least $f = .19$ for study 1, and $f = .16$ for study 2 (with both being a small-to-medium effect size).

Table 1

Demographics for Study 1 and 2

Factor	S1	S2
N	265	365
Age		
Mean (SD)	41.28 (13.78)	39.66 (13.15)
Gender		
% Female	54.5	52.1
% Male	45.3	46.8
% Other	.4	.5
Race		
% White	82.3	81.6
% Black	5.3	6.6
% Hispanic/ Latino	6.0	3.6
% Other	6.3	7.4
Political Orientation		
Liberal	147	204
Conservative	118	161
Political Party		

[2] We report the ingroup estimations in the supplemental materials; these results are also broken down by political affiliation, and include a column with participants' actual agreement with the issues in each study.

Democrat	130	172
Republican	115	155
Libertarian	12	21
Green	3	6
Other	5	11

Note: Demographics are reported after all exclusions were in effect.

Procedure

Participants were first shown a brief definition of "liberal" and "conservative" before providing their political affiliation (see Appendix A for wording). The definitions were relatively minimal, associating liberals with voting Democrat [and conservatives with voting Republican] but also extending it to those holding liberal [conservative] views who don't necessarily vote. Participants were then randomly assigned to one of six conditions in study 1; three of these conditions showed participants tweets regarding a conservative issue, and three showed tweets regarding a liberal issue. Because we created caricatured "extreme tweets" meant to reflect fringe views in the culture war, but were not necessarily reflective of policies supported by either party, we use the terms "liberal and "conservative" rather than "Democrat" and "Republican" (though the majority of the latter self-identified into the expected liberal/conservative categories).[3] For each type of tweet, participants were either shown an image of the tweet on its own, the tweet followed by a supporting response, or the tweet followed by a dissenting response (see Figure 1 for study 1 tweets, and Figure 2 for study 2 tweets); the responder always identified as a member of the same political group as the initial tweeter. This set up allowed us to examine whether seeing evidence of either consensus or opposition to an extreme view (one that is assumed to be prevalent among members of that group) changed readers' estimations of what percentage of outgroup members might endorse those extreme attitudes, when compared to estimates in the control condition. In study 2, the design was almost identical, except that it included a second control condition where participants were only shown an extreme *statement* made by a liberal or conservative, not in the

[3] The "culture war" issues that were selected were ones meant to reflect the kind of discourse we observed occurring on social media. These topics were frequently referenced, and often appeared to incite strong emotions (like outrage).

tweet format; this was done to determine whether participants thought an attitude was *more* prevalent seeing it expressed in the social media context. However, as participants' estimations of opponent agreement did not differ between control conditions (the initial extreme view presented as a tweet or statement with no response) in study 2 (*p*s > .468), they were combined for analysis.

Participants were then asked to estimate the percentage of both liberals and conservatives that agreed with the initial extreme tweet, as well as their personal agreement with the tweet. Next, participants rated their liking and trust of liberals and conservatives, as well as their willingness to engage with political opponents in several activities. Finally, they provided general demographic information. They were compensated $1.50 USD upon completion.

Figure 1

Study 1 Tweets

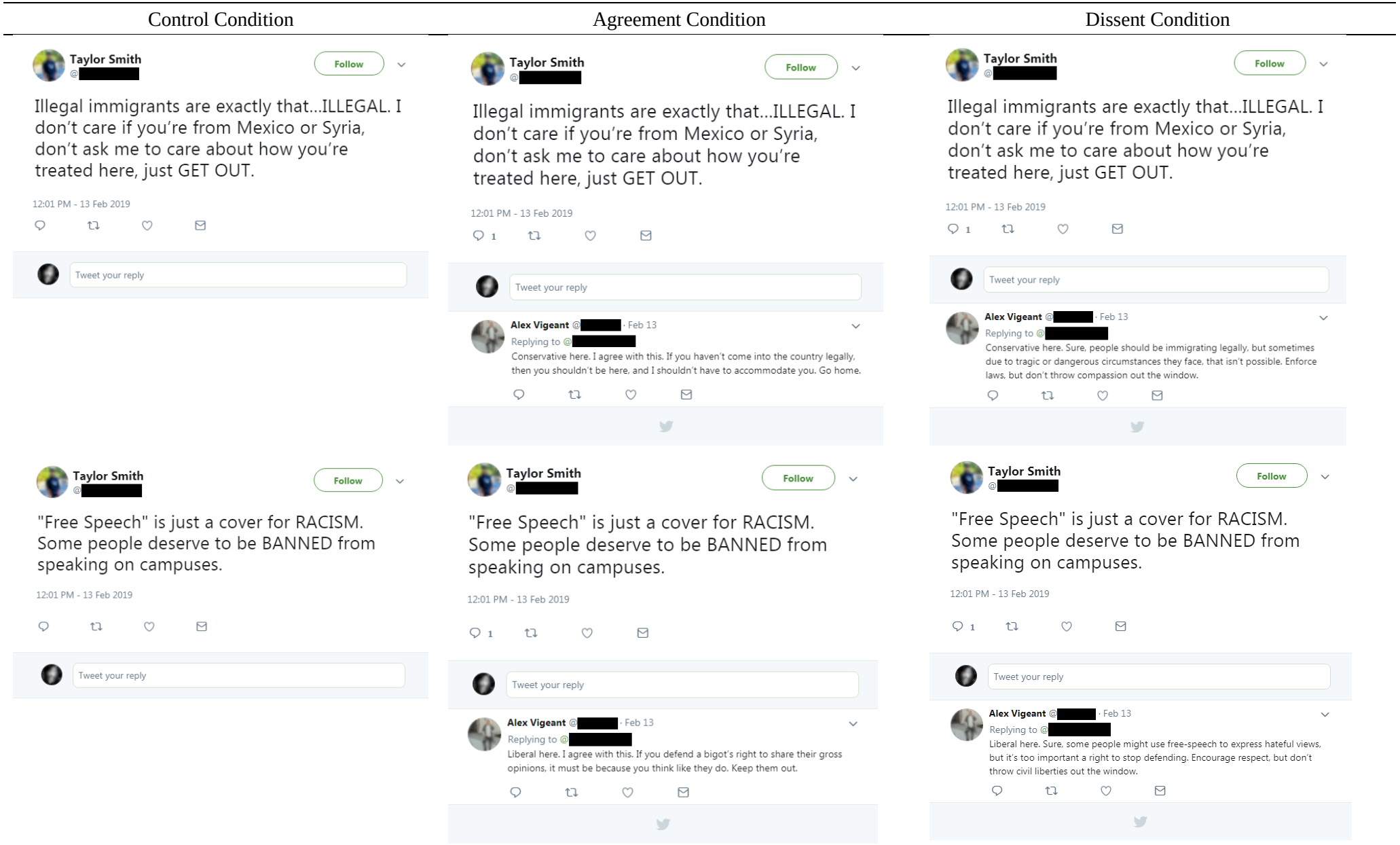
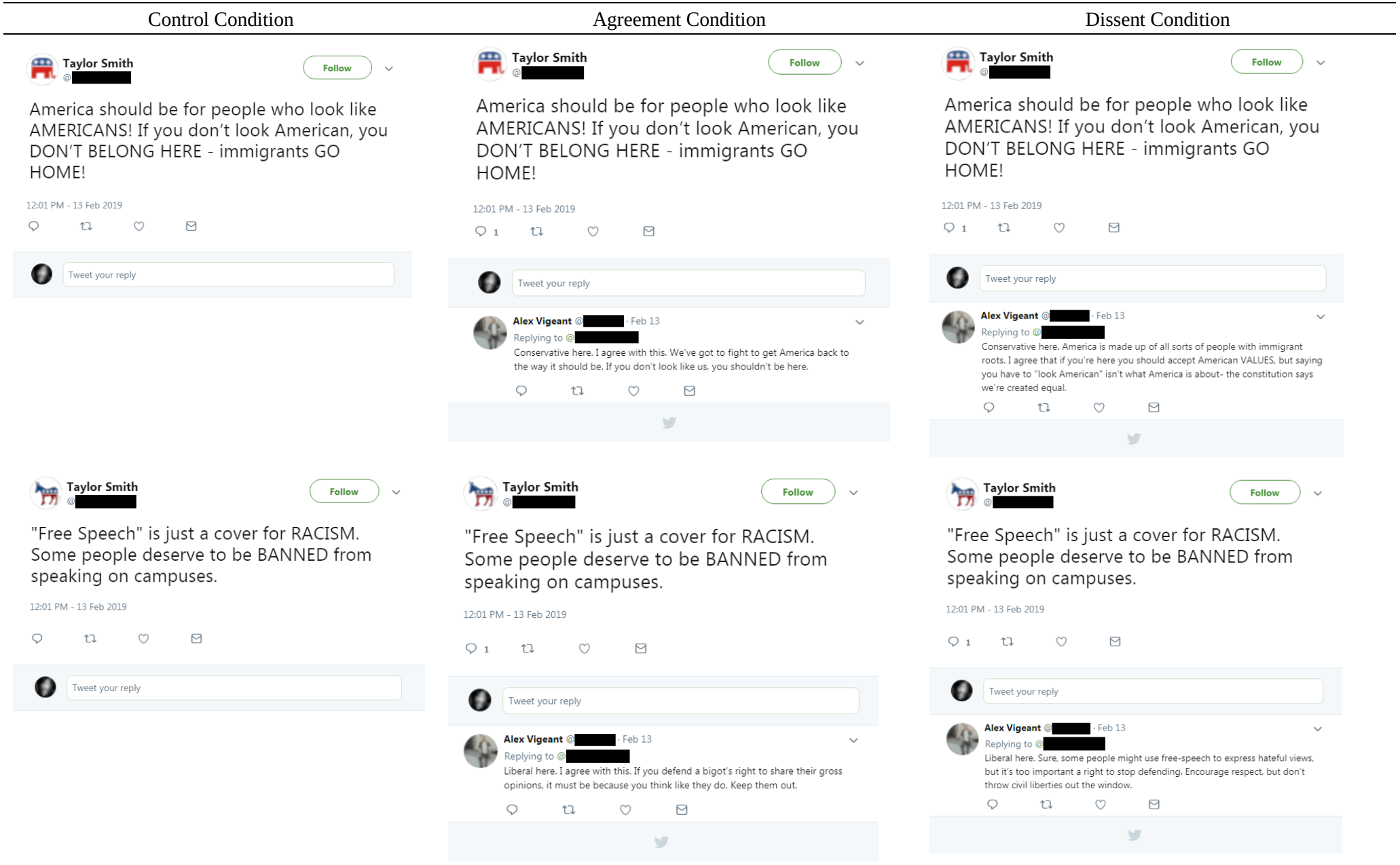


Figure 2

Study 2 Tweets



Materials

Materials were the same between studies 1 and 2 unless otherwise indicated. For the full list of materials and questionnaire items, please refer to Appendix A.

Political Affiliation. Participants categorized themselves as either "Mostly Conservative," "Mostly Liberal," or "Both Equally." We used a dichotomous measure of partisanship for two reasons. First, we assigned people to view an out-party tweet, so we needed a way to categorize everyone as either an in-party member, or out-party member. Secondly, we needed to calculate an 'actual prevalence' score (i.e., actual personal agreement with the tweet) from anyone categorized as liberal or conservative, which also required a dichotomy.

Other Agreement. Participants estimated the percentage of both liberals and conservatives who agreed with the extreme tweets on a scale from 0-100.

Personal Agreement. Participants rated how much they personally agreed with the extreme tweet on a scale from *(1) Strongly Disagree*, to *(6) Strongly Agree.*

Feelings Towards Opponents. Participants reported how much they both like and trust liberals and conservatives on a scale from *(0) Not At All*, to *(100) Very Much*. As both variables were highly correlated (r = .83) we report them as a composite, referred to as "feelings towards opponents."

Engagement. Participants reported how willing they would be to engage with political opponents on a 5-item, 6-point scale, which included items such as having a political discussion, sharing a taxi, or shaking hands (S1: α = .74, S2: α = .73).

Results

Effect of Dissent

To compare estimates of opposition agreement between conditions, and to determine whether these effects were the same across political affiliations, we ran a 3x2 ANOVA, using condition and partisanship as the predictors, and opposition agreement estimates as the outcome. As predicted, for both study 1 and study 2, there was a significant main effect of condition, S1: $F(2, 262) = 7.47$, $p < .001$, $\eta^2 = .054$, S2: $F(2, 359) = 3.62$, $p = .028$, $\eta^2 = .020$. LSD post-hoc tests revealed that, across political

affiliations, seeing outgroup members dissent against opponents' extreme tweets lowered participants' estimates for how prevalent that attitude was amongst opponents compared to both seeing outgroup members agree, and seeing no response at all (see Table 2).

As expected, for both study 1 and study 2, there was no significant interaction between condition and partisanship, suggesting that the intervention was functioning the same way for liberals and conservatives, S1: $F(2,262) = .188$, $p = .829$, $\eta^2 = .001$, S2: $F(2, 359) = 1.15$, $p = .317$, $\eta^2 = .006$. Additionally, in study 1, there was a main effect of partisanship on agreement estimations, such that liberals ($M = 80.17$, $SD = 17.68$) had greater estimations of opponent agreement than conservatives ($M = 66.75$, $SD = 23.80$), $F(1, 262) = 28.71$, $p < .001$, $\eta^2 = .099$.

Table 2

Opponent Agreement Estimations by Condition for Study 1 and 2

		Agree	Disagree	Control	Actual Opponent Agreement
		M (SD)	M (SD)	M(SD)	
Overall	S1	77.67_a (19.31)	67.85_b (22.96)	77.46_a (20.79)	52%
	S2	66.03_a (26.53)	55.73_b (26.22)	65.15_a (26.00)	32%

Note: Differing subscripts denote differences between conditions at the $p < .05$ level; LSD post-hoc tests were used. "Actual Opponent Agreement" refers to the percentage of liberals who scored between 4-6 on their personal agreement with the liberal tweet, and the percentage of conservatives who scored between 4-6 on their personal agreement with the conservative tweet.

Table 2 includes estimates by condition as well as the actual percentage of participants who selected "at least somewhat agree" or above with each item in the "actual opponent agreement" column. These estimates allow us to examine whether participants' estimates of their opponents' agreement were roughly accurate, over-estimates, or under-estimates; over-estimates reflect false polarization. Although false polarization was not the central focus of these analyses, we observed it in most instances. As expected, fewer participants actually agreed with the extreme attitude than their opponents estimated, with one exception. In study 1, unexpectedly, the majority of conservatives agreed with the conservative tweet (76%), leading us to alter its wording in study 2.

Liking, Trust, and Engagement

We ran a 3x2 ANOVA using condition and partisanship as the predictors, and feelings towards opponents, and willingness to engage as separate outcomes to determine whether these variables differed between conditions, and further, whether these differences were the same across political affiliations (see Table 3 for descriptives and mean differences). In study 1, there was a significant main effect of condition on feelings towards opponents, $F(2, 259) = 3.57$, $p = .030$, $\eta^2 = .027$. LSD post-hoc tests revealed that participants in the agreement condition liked and trusted the opponents less than participants in the dissent and control conditions. There was no main effect of condition on willingness to engage in study 1, $p = .276$. There were also no main effects of condition on feelings towards opponents ($p = .227$), or willingness to engage with opponents ($p = .202$) in study 2.

In study 1 and study 2, there were no significant interactions for either feelings towards opponents (S1: $p = .679$, S2: $p = .828$) or willingness to engage (S1: $p = .925$, S2: $p = .067$). For study 2, there was a main effect of partisanship, such that conservatives ($M = 4.11$, $SD = .86$) were more willing to engage with opponents than liberals ($M = 3.96$, $SD = .89$), $F(1, 359) = 6.14$, $p = .014$, $\eta^2 = .017$.

In both study 1 and study 2, feelings towards opponents and willingness to engage with the opponent were highly correlated (see Table 4); the more participants liked and trusted their opponents, the more willing they were to engage with them. Further, the opponent agreement estimations were negatively correlated with feelings towards opponents; the more participants thought opponents agreed with the tweet, the less they liked and trusted them, and in study 2, the less they wanted to engage with them. There was no correlation between estimations and engagement in study 1.

Table 3

Participant Liking of, Trust of, and Engagement with the Opponent for Study 1 and 2

	Agree	Disagree	Control
	M (SD)	M (SD)	M(SD)

Overall	S1	Opponent Feelings	19.22_a (18.69)	28.21_b (23.16)	27.23_b (25.04)
		Engage	4.01 (.90)	4.22 (.81)	4.10 (.89)
	S2	Opponent Feelings	23.44 (23.93)	28.30 (21.00)	24.86 (22.33)
		Engage	4.04 (.91)	4.13 (.86)	3.96 (.90)

Note: Differing subscripts denote differences between conditions at the $p < .05$ level; LSD post-hoc tests were used. High "feelings" scores correspond to more positive feelings (more liking and trust).

Table 4

Correlations Between Opponent Agreement Estimations, Liking and Trust, and Engagement

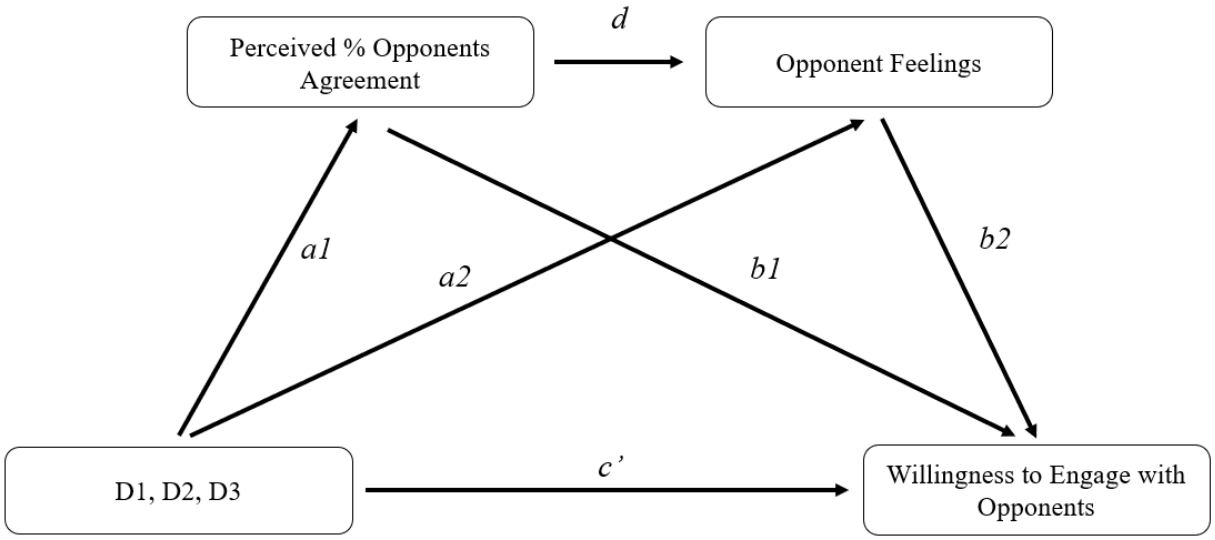
	Study 1			Study 2		
	1	2	3	1	2	3
1. Opponent Estimations						
2. Opponent Feelings	-.24**			-.34***		
3. Engagement	-.05	.43***		-.18***	.42***	

Note: *** indicates $p < .001$, ** indicates $p < .01$, * indicates $p < .05$.

Indirect effect of response condition on engagement. Although the tweet response condition (dissent, agree, no response) did not consistently affect feelings towards opponents or willingness to engage with opponents, it did affect prevalence estimates for the extreme view, which based on our past theory and research (wherein we tested an SEM model with similar parameters; Parker et al., 2023), should predict liking and, in turn, engagement. This is also replicated and supported by the previous correlations. As such, we tested a serial mediation model (using PROCESS) in both study 1 and study 2 (see Figure 3 for conceptual model and pathway labels, and Table 5 for all pathway coefficients between variables.). We predicted that exposure to dissent would reduce estimates of opponent agreement (*a1* path), and that lower prevalence estimates would predict more positive feelings towards opponents (*d* path), which in turn would predict a greater willingness to engage with opponents (*b2* path).

Figure 3

Conceptual Serial Mediation Model with Pathway Labels



Note: Condition was categorical, so three contrasts were used in the analysis; D1 represents Control vs. Agree, D2 represents Control vs. Disagree, and D3 represents Agree vs. Disagree.[4]

In both studies, we found a significant indirect effect of D2 (Control vs. Disagree) on willingness to engage through opponent estimations and the composite of liking and trust (S1: effect = .046, 95% CI [.013, .089], S2: effect = .054, 95% CI [.015, .102]), as none of the bootstrapped confidence intervals contain 0. Additionally, we find a significant indirect effect of D3 (Agree vs. Disagree) on willingness to engage through opponent estimations and the composite of liking and trust (S1: effect = .047, 95% CI [.012, .095], S2: effect = .061, 95% CI [.014, .117]). This was not the case for D1 (Control vs. Agree). All effects reported are partially standardized.[5]

Table 5

Pathway Coefficients for Study 1 and Study 2 Serial Mediations

	a1	*a2*	*b1*	*b2*	*d*	*c*	*c'*

[4] Two dummy coded condition variables were created within the data set, and two serial mediations were run to obtain the results for all three contrasts.
[5] As the predictor is categorical, and all other variables are continuous, the output is "partially standardized", as only three of the four variables in the model can be standardized.

S1	Opponent Feelings	D1	.22	1.89				.09	.06
		D2	-9.61**	-1.18	.002	.02***	-.24***	.08	.08
		D3	-9.82**	-3.08				-.02	.02
S2	Opponent Feelings	D1	1.32	-1.04				.08	.11
		D2	-9.43**	.71	-.001	.02***	-.29***	.13	.06
		D3	-10.75**	1.75				.05	-.05

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05. D1 represents the Control vs. Agree contrast, D2 represents the Control vs. Disagree contrast, and D3 represents the Agree vs. Disagree contrast.

Study 1 and Study 2 Discussion

These first two studies were an initial test to determine if seeing ingroup dissent following the expression of an extreme attitude on Twitter changed participants' perception that the attitude was widespread amongst political opponents. Overall, participants who were shown an opponent dissenting against an extreme tweet changed their perceptions of opponent attitudes; specifically, they thought fewer of their opponents held those attitudes, compared to when they were shown an opponent agreeing, or no response at all.

We did note in study 1 that the conservative tweet had elicited a much higher degree of conservative agreement than expected. We speculated that this may have been due to the tweet's wording, as it mentioned "illegal immigrants" which we (incorrectly) assumed would be interpreted as overtly racist. It is possible that, instead, conservatives interpreted this tweet as a national security issue (with a focus on opposition to the "illegal" component), not a racist one. Because of the ambiguity of the tweet's meaning in study 1, in study 2 we altered the wording to more unambiguously reflect explicit racist attitudes.

The Twitter interactions did, in one case, influence participants' feelings towards their opponents; overall in study 1, seeing agreement with the extreme tweet did increase dislike and distrust compared to seeing dissent, or no response at all. While seeing a dissenter did not directly reduce liking/trust or engagement, tweet response condition did affect willingness to engage indirectly through prevalence estimates and feelings towards opponents. Seeing dissent lowered estimates of opponent agreement with extreme tweets, and these lowered estimates predicted greater liking and trust of opponents, which

subsequently, predicted a greater willingness to engage with them. This suggests that, at least indirectly, seeing opponent dissent may, in the long run, encourage greater cross-party discussion and engagement.

One of the goals of the intervention was, specifically, to directly improve feelings towards political opponents through exposure to opponent dissent, though in both studies 1 and 2, it was unsuccessful. We speculated that, while the single dissenting response was enough to shift perceptions of attitude prevalence, perhaps more evidence was needed for participants to shift their feelings as well. As such, in study 3 and study 4, we removed the "agreement" condition, and instead included a condition wherein participants saw *multiple* instances of opponent dissent.

Study 3 and Study 4

Although the single dissenting response reduced participants' estimates of opponent agreement with the extreme tweet, these estimates were still considerably higher than the attitude's actual prevalence; partisans were still estimating that a majority of their opponents held the extreme views. It may be that the correction was not strong enough to alter liking or engagement since a particularly egregious tweet was still salient and estimated to be common. The extreme tweet may also have evoked additional beliefs about the immoral nature of the out-party, making animosity difficult to shift. As such, we tested whether one way to strengthen the effects of dissent on opponent perceptions would be to show opponents *multiple* dissenters, or what we refer to as a "dogpile." Our approach shared some similarities to past work that exposed participants to an online pile-on (Sawaoka & Monin, 2018), examining how the virality of negative responses to offensive internet posts (i.e., a single response versus many responses) changed perceptions of both the poster of the offensive content, and the initial responder. However, in our study design, we constructed a different kind of pile-on. Participants were, again, exposed to an extreme tweet from a political opponent, and were then exposed to outgroup dissent (i.e., dissent from an ingroup member of the original tweeter). We experimentally varied whether people were exposed to a single dissenting tweet or a "dogpile." However, unlike the approach taken by Sawaoka and Monin (2018) in which the pile-on was full of outrage and derision, we created dissenting tweets that disagreed in a civil manner with the original position. Here, we tested dissent phrased only in civil ways to ensure that any

reduction in opponent agreement estimates could be attributed to the dissent specifically, and not the tone of the response.

We expected that seeing a dogpile of dissent would more consistently and effectively reduce participants' estimates of the prevalence of those attitudes amongst opponents, compared to a single dissenting response, or no response at all. Seeing *more* dissenters should, in theory, provide participants with a greater pool of vivid exemplars to counter the initial tweet. Further, while the stronger intervention was intended to test whether we could directly mitigate dislike and disengagement, in the event that feelings towards or willingness to engage with opponents did not differ by condition, we expected to replicate the serial mediation; seeing greater opponent dissent in response to the tweet will reduce estimations of opponent agreement, which will subsequently predict greater positive feelings towards opponents, which will in turn predict greater willingness to engage with them.

Methods

Participants

American participants from Cloud Research were tested in both studies 3 and 4; a full breakdown of the demographics can be found in Table 6.[6] Participants who failed the attention check (S3: n = 52, S4: n = 64), or identified as "both equally" politically (S3: n = 104, S4: n = 52) were excluded. Given our sample size used in analyses, an alpha level of .05, and minimum power of .80 (Cohen, 1992), sensitivity analyses using GPower indicated that, for an ANOVA with main effects and interactions, we could detect an effect size of at least f = .12 (a small effect size) for study 3, and f = .16 (a small-to-medium effect size) for study 4.

Table 6

Demographics for Study 3 and 4

Factor	S3	S4

[6] In study 3, as opposed to previous studies, participants were only shown opponent tweets; this was because personal agreement on the tweets used in study 3 had already been collected in study 2, thereby providing the false polarization information. This was reverted in study 4, in order to obtain new false polarization results for the updated tweets.

N	687	362
Age		
Mean (SD)	37.95 (11.87)	41.48 (14.11)
Gender		
% Female	43.5	56.1
% Male	55.6	43.6
% Other	0.4	0
Race		
% White	73.4	75.7
% Black	9.5	7.5
% Hispanic/ Latino	5.8	5.8
% Other	10.9	10.5
Political Orientation		
Liberal	456	212
Conservative	231	150
Political Party		
Democrat	404	189
Republican	209	146
Libertarian	43	18
Green	14	5
Other	16	4

Note: Demographics are reported after all exclusions were in effect.

Procedure

The procedure was almost identical to studies 1 and 2. However, instead of seeing responses that either agreed or disagreed with the initial tweet, participants saw either a single dissenting response from an ingroup member of the political opponent who wrote the initial tweet, or a "dogpile" of dissenting responses from multiple ingroup members of the political opponent who wrote the initial tweet (see Appendix B for tweets from each condition). The agreement condition was dropped. Additionally, we changed the liberal tweet in study 4; whereas in earlier studies it referred to "banning free speech on

college campuses", in study 4, we altered the wording to refer to "white toxicity". This was done to better mirror the conservative tweet, which was also related to race (see Figure 4 for image).

In study 4, we also included some additional items measuring liking or animosity to better capture variance across the measure (as our previous measure was a composite of two single items). We also included questions to gauge participants' feelings towards Democrat and Republican political elites, as well as towards regular citizens who are liberal or conservative, as some research has suggested that partisans do not feel the same way about political elites compared to ordinary party members (Kingzette, 2020). It is difficult, when asking about liking and trust of opponents, to know exactly whether participants are thinking of the average citizen, or other more prolific exemplars like political elites. It is possible that participants' dislike of opponents is focused on specific individuals who are much more deserving of dislike compared to regular partisans. One way to address this is to be more specific about who is the target of their dislike (in order to avoid response substitution), so we first asked about their feelings towards opponent elites to ensure that, when asking about feelings towards partisans, that was the only target they had in mind (Gal & Rucker, 2011; Schwarz, 1999). By asking participants *first* about political elites, this allows them to separate their elite-targeted feelings from their citizen-targeted feelings and helps to avoid response substitution (wherein their responses to "opponent feeling" questions are actually expressing their feelings towards elites).

Participants were paid $1.50 USD upon completion.

Figure 4

Updated Liberal Tweet

Materials

Materials in studies 3 and 4 were similar to previous studies, with a few changes. Personal agreement, estimations of agreement, and engagement (S3: α = .77, S4: α = .73) remained the same. In study 3, ratings of opponent liking and trust were highly correlated (*r* = .88), so we report them here as a single composite.

Feelings Towards Opponents. Participants rated their agreement that Republican elites (α = .85) and conservative partisans (α = .83), and Democratic elites (α = .83) and liberal partisans (α = .79) were: worthy of respect, likable, worthy of hatred (reverse coded), and trustworthy from *(1) Strongly Disagree*, to *(6) Strongly Agree*. Additionally, they reported whether Republicans and conservatives, and Democrats and liberals were "a threat to democracy" on the same scale; as this is a measure that bodes particularly poorly for democracy if individuals are seeing their opponents as existential threats, we kept this item separate, and analysed it as a stand-alone variable. To note, as the tweets are written by, and the estimations are targeted towards partisans specifically, we did not analyze liking questions targeting elites. The inclusion of elite-targeted questions was to allow participants to express their dislike for elites

separately, ensuring that participants were thinking exclusively of partisans when answering partisan-targeted questions.

Results

Effect of Dissent

As in earlier studies, we ran a 3x2 ANOVA, using condition and partisanship as the predictors, and opposition agreement estimates as the outcome (see Table 7 for descriptives and post-hoc comparisons). As predicted, in both study 3 and study 4, we found a significant main effect of condition, S3: $F(2, 681) = 25.50$, $p < .001$, $\eta^2 = .070$, S4: $F(2, 356) = 4.50$, $p = .012$, $\eta^2 = .025$. LSD post-hoc tests revealed that seeing a dogpile of dissent reduced participants prevalence estimates of opponent agreement when compared to seeing only a single response, or no response at all. Additionally, in study 3, seeing a single dissenting response also lowered prevalence estimates compared to seeing no response.

In both studies, there were significant interactions between condition and partisanship, S3: $F(2, 681) = 3.01$, $p = .050$, $\eta^2 = .009$, S4: $F(2,356) = 7.24$, $p < .001$, $\eta^2 = .039$ (see Table 8 for means and simple effect comparisons). In study 3, for both liberals and conservatives, the conditions all differed in the expected direction, but liberals appeared to show a stronger effect than conservatives in the dogpile condition, as their estimates were significantly lower than conservatives in this condition only. In study 4, liberals in the dogpile condition had lower prevalence estimates than the single response and control conditions, however there were no condition differences for conservatives at all. Liberal estimations were significantly higher than conservatives in the single response and control conditions, but not in the dogpile condition.

Additionally, in study 4, a main effect of partisanship emerged, such that conservatives, across all conditions, had lower opponent agreement estimates than liberals, $F(1,356) = 46.29$, $p < .001$, $\eta^2 = .115$.

Table 7

Opponent Agreement Estimations by Condition for Study 3 and 4

		Single Response	Dogpile	Control	Actual Opponent Agreement
		M (SD)	M (SD)	M(SD)	
Overall	S3	53.92_{b} (25.34)	41.99_{a} (27.97)	62.14_{c} (25.78)	-
	S4	48.59_{b} (27.96)	39.55_{a} (26.93)	51.57_{b} (29.25)	14%

Note: Differing subscripts denote differences between conditions at the $p < .05$ level; LSD post-hoc tests were used. Actual opponent agreement is not available for S3, as participants were only shown opponent tweets.

Table 8

Opponent Agreement Estimations by Condition for Study 3 and 4, for Liberals and Conservatives

		Control	Single	Dogpile	Total
		M(SD)			
S3	Liberals	62.47_{aa} (25.84)	53.92_{ba} (24.65)	38.36_{ca} (26.10)	51.72 (27.36)
	Conservatives	61.49_{aa} (25.82)	53.92_{ba} (26.91)	48.85_{cb} (30.20)	54.77 (28.10)
S4	Liberals	60.77_{aa} (26.31)	60.81_{aa} (22.83)	41.42_{ba} (25.30)	54.49 (26.41)
	Conservatives	37.49_{b} (28.11)	31.63_{b} (25.63)	37.06_{a} (29.01)	35.43 (27.59)

Note: Subscripts denote interaction simple effects. The first subscript denotes differences between conditions (for liberals and conservatives separately), the second subscript denotes differences between liberals and conservatives (for each condition). Differing subscripts denote differences between conditions at the $p < .05$ level

Feelings Towards Opponents, Threat, and Engagement

We ran 3x2 ANOVAS, using condition and partisanship as predictors, and feelings towards opponents, threat (for study 4), and willingness to engage as outcomes to examine both the main effect of condition, and possible interactions. For both study 3 (ps < .470) and study 4 (ps < .332), there were no main effects of condition on feelings towards opponents, threat, or willingness to engage.

For study 3, there were no interactions between condition and partisanship for either feelings towards opponents ($p = .670$) or willingness to engage ($p = .153$). There was a main effect of partisanship on feelings towards opponents, such that across all conditions, conservatives ($M = 35.56$, $SD = 28.39$) liked and trusted opponents significantly more than liberals ($M = 25.37$, $SD = 21.72$), $F(1, 681) = 25.99$, $p < .001$, $\eta^2 = .037$.

Similarly, for study 4, there were no interactions between condition and partisanship for feeling towards opponents (p = .118), opponent threat (p = .590), or willingness to engage (p = .739). There was a main effect of partisanship for both feelings towards opponents, $F(1,356) = 5.73$, p = .017, η^2 = .016, and for willingness to engage with opponents, $F(1,356) = 7.83$, p = .005, η^2 = .022 (see Table 10 for means), but not for threat, $F(1,356) = 2.72$, p = .100, η^2 = .008. Conservatives had more positive feelings toward and were more willing to engage with opponents than liberals.

See Table 11 and Table 12 for the full overall correlations between opponent agreement estimates, feelings towards opponents, and engagement for study 3 and study 4. As in previous studies, the more participants thought the opponent agreed with the extreme tweet, the less positively they felt about them, the more threatened they felt by them, and the less they wanted to engage with them.

Table 9

Participant Feelings Towards the Opponent for Study 3 and 4

			Single Response	Dogpile	Control
			M (SD)	M (SD)	M(SD)
Overall	S3	Opponent Feelings	27.88 (24.60)	30.68 (24.78)	28.79 (24.62)
		Engage	4.07 (.92)	4.19 (.87)	4.09 (.92)
	S4	Opponent Feelings	3.37 (1.04)	3.39 (1.00)	3.52 (.99)
		Opponent Threat	3.88 (1.43)	3.83 (1.44)	3.72 (1.40)
		Engage	4.12 (.88)	4.12 (.81)	4.13 (.90)

Table 10

Liberal and Conservative Feelings Towards Opponents, Threat, and Engagement

	Liberals	Conservatives
	M (SD)	M (SD)
Opponent Feelings	3.05 (.96)	3.28 (.95)
Opponent Threat	3.99 (1.36)	4.23 (1.25)
Engagement	3.99 (.93)	4.24 (.78)

Note: Higher "feelings" scores correspond to more positive feelings, and higher threat scores correspond to higher feelings of threat.

Table 11

Correlations between Opponent Agreement Estimations, Liking and Trust, and Engagement for Study 3

	1	2	3
1. Opponent Estimations			
2. Opponent Feelings	-.14***		
3. Engagement	-.15***	.44***	

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05.

Table 12

Correlations between Opponent Feelings, Threat, and Engagement for Study 4

	1	2	3	4
1. Opponent Estimations				
2. Opponent Feelings	-.29***			
3. Opponent Threat	.21***	-.60***		
4. Engagement	-.24***	.50***	-.34****	

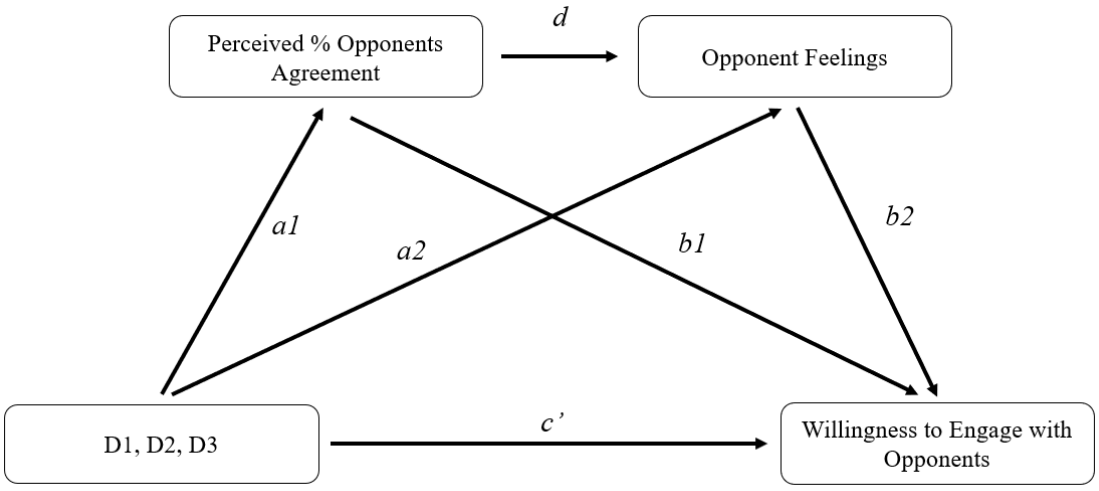
Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05.

Indirect effect of response condition on engagement. We then tested the serial mediation model from previous studies; for study 3, the conceptual model is the same as Figure 3, while the conceptual models for study 4 are depicted in Figure 5 and Figure 6. For study 4, one model was run

using partisan feelings as the second mediator, and one was run using opponent threat as the second mediator. All effects reported are partially standardized.[7]

Figure 5

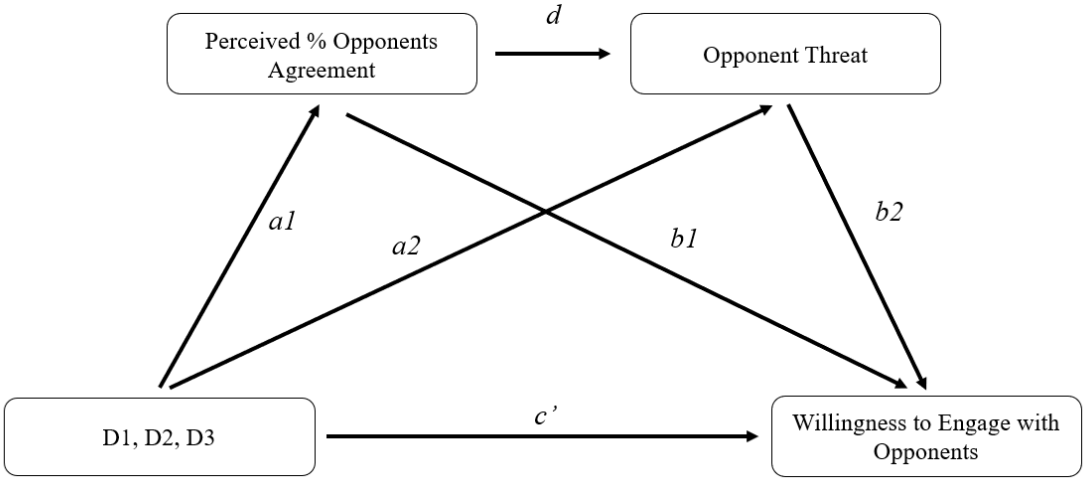
Serial Mediation of Condition, Agreement Estimates, Opponent Feelings, and Engagement for S4



Note: D1 represents Control vs. Single Response, D2 represents Control vs. Dogpile, and D3 represents Single Response vs. Dogpile.

Figure 6

Serial Mediation of Condition, Agreement Estimates, Opponent Threat, and Engagement for S4



[7] To note, because of the significant interactions between condition and partisanship on opponent agreement estimates, we also ran a moderated serial mediation model, which included partisanship as the moderator. As these two studies were the only ones where this interaction was present, these results have been included in the supplemental material.

Note: D1 represents Control vs. Single Response, D2 represents Control vs. Dogpile, and D3 represents Single Response vs. Dogpile.

Table 13 displays all pathway coefficients between variables. In study 3, for all contrasts, two indirect effects emerged as significant (see Table 14). Estimations of opponent agreement significantly mediated the relationship between condition and engagement. Additionally, as predicted, we found a significant indirect effect of condition on willingness to engage through opponent estimations and liking and trust of opponents.

For study 4, for D2 and D3, we found the predicted significant indirect effect of condition on willingness to engage through opponent estimations and feelings towards opponents (see Table 15). We also found a significant indirect effect of condition on willingness to engage through opponent estimations and opponent threat (see Table 16). Additionally, there was an indirect effect of condition on willingness to engage through just opponent estimations for both D2 and D3. Lastly, there was an indirect effect of condition on willingness to engage through just feelings towards opponents for D2.

To note, in study 4, one pathway became significant unexpectedly. Specifically, for D2, being in the dogpile condition was negatively related to liking; seeing dissent appeared to lower positive feelings towards partisans. While this effect is not replicated in any other model across studies, it is worth noting that the relationship is not in the expected direction.[8]

Table 13

Pathway Coefficients for Study 3 and Study 4 Serial Mediations

	Mediator		*a1*	*a2*	*b1*	*b2*	*d*	*c*	*c'*
S3	Opponent Feelings	D1	-8.22***	-2.03				-.02	-.03
		D2	-20.15***	-.61	-.003*	.02***	-.13***	.10	.01
		D3	-11.93***	1.42				.12	.04
S4		D1	-3.03	-.15				-.03	-.00

[8] To note, as there were no differences in feelings towards opponents between conditions, and because the a2 path was non-significant in previous studies, it is possible that the negative coefficient that emerged here for the D2 a2 pathway is a statistical artifact, and should be interpreted with caution.

Partisan Feelings	D2	-12.35***	-.31*	-.004*	.39***	-.01***	-.15	-.12
	D3	-9.04*	-.15				-.11	-.12
Partisan Threat	D1	-3.24	.08				-.03	-.04
	D2	-12.15***	.25	-.006***	-.18***	.01***	-.14	-.19
	D3	-8.90*	.17				-.12	-.16

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 14

Indirect Effects of Condition Contrasts on Engagement for Study 3

		95% CI	
	Effect	Lower	Upper
Condition – Estimate – Engage			
D1	**.027**	.005	.057
D2	**.066**	.015	.121
D3	**.039**	.009	.078
Condition – Opponent Feelings – Engage			
D1	-.035	-.115	.044
D2	-.011	-.095	.077
D3	.025	-.056	.107
Condition – Estimate – Opponent Feelings - Engage			
D1	**.018**	.005	.036
D2	**.045**	.016	.078
D3	**.026**	.009	.048

Note: Bolded effects are significant at the p < .05 level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 15

Indirect Effects of Condition Contrasts through Opponent Feelings on Engagement for Study 4

		95% CI	
	Effect	Lower	Upper
Condition – Estimate – Engage			
D1	.013	-.015	.056
D2	**.050**	.003	.117
D3	**.037**	.001	.089
Condition – Partisan Feelings – Engage			

D1	-.070	-.178	.045
D2	**-.138**	-.255	-.024
D3	-.068	-.188	.042
Condition – Estimate – Partisan Feelings - Engage			
D1	.017	-.020	.058
D2	**.063**	.025	.113
D3	**.046**	.009	.091

Note: Bolded effects are significant at the $p < .05$ level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 16

Indirect Effects of Condition Contrasts through Opponent Threat on Engagement for Study 4

		95% CI	
	Effect	Lower	Upper
Condition – Estimate – Engage			
D1	.022	-.026	.081
D2	**.084**	.027	.163
D3	**.061**	.011	.126
Condition – Partisan Threat – Engage			
D1	-.016	-.093	.060
D2	-.050	-.128	.022
D3	-.034	-.111	.037
Condition – Estimate – Partisan Threat - Engage			
D1	.008	-.010	.027
D2	**.028**	.009	.054
D3	**.021**	.004	.043

Note: Bolded effects are significant at the $p < .05$ level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Study 3 and 4 Discussion

Seeing a greater degree of dissent (in the form of an ingroup "dogpile") did have a more consistent effect on participants' estimates of opponent agreement with extreme tweets compared to a single response, or no response at all. Unlike a single dissenting tweet, exposure to multiple out-party members dissenting against the out-party extremist reduced prevalence estimates to below the 50% mark; this means the dogpile of dissent successfully convinced people that the extreme view was not held by a *majority* of out-party members. Although the dogpile of dissenters seems promising with regard to impact

on misperceptions, a condition x party interaction effect emerged in both studies. In Study 3, conservatives showed the identical effect but more weakly than liberals in the dogpile condition. In Study 4, conservatives were not significantly impacted by *any* degree of liberal dissent. Looking at their estimates of liberal agreement, it appears that the tweet we chose in Study 4 (about white toxicity) may have been part of the issue; agreement estimates were low across all conditions, suggesting that conservatives may not have believed that this particular attitude was one that liberals commonly endorsed in any condition.

While the effects of the conditions on opponent agreement estimations were overall stronger in studies 3 and 4 (with one exception for conservatives), these stronger effects still did not directly improve feelings towards, or willingness to engage with political opponents. However, we replicated our proposed serial mediation model observed in study 1 and 2; seeing greater dissent lowered participants estimates of opponent agreement with egregious tweets, which predicted more positive feelings toward opponents, which in turn predicted more willingness to engage with the opponent. The same indirect effect was observed for partisans' belief that opponents are a threat to democracy; we examined this variable separately because the rise of this sentiment has serious implications for democratic functioning.

Our approach to exploring the effects of dissent, specifically creating stimuli based on real-world issues, does have some inherent limitations. While using these culture-war issues ensures high ecological validity, as they are based on comments seen and read online, they do require a sacrifice in experimental control; it is difficult to ensure that issues are equally impactful or extreme across both parties. This is part of the reason why, beyond changing social circumstances, personal agreement and opponent estimations occasionally jump around in ways we did not anticipate.

Study 5

In a final study, we first updated the tweets used in the manipulation; this was done to increase the generalizability of our findings. Further, we wanted to enhance the parallel between the liberal and conservative tweets by creating more parallel "culture war" positions on opposite ends of a particular extreme. We conducted this study during a time when police reform and accountability were a large part

of public discourse (February 2021). As such, the tweets reflect the kind of fringe attitudes from either side on the topic of police conduct specifically (see Figure 7; for full tweets, see Appendix B). This also allowed us to address an issue in study 4, wherein conservative participants did not assume that liberals endorsed the ostensibly 'extreme' tweet we presented them.[9]

The mediation analyses we use to examine indirect effects of condition on engagement are cross-sectional, and thus we cannot make claims about causality. Agreement estimates could be predicted *by* liking, rather than producing changes in liking; perhaps those who dislike opponents more already are willing to assign higher percentage estimates to their opponents' agreement with egregious attitudes. Although this cannot account for the causal effects of the manipulation on prevalence estimates, we can't rule out the possibility that much of the association between prevalence and liking is due to pre-existing levels of liking. Thus, in study 5, we partially address this concern by including a measure of opponent dislike at the outset of the survey, in order to control pre-existing opponent liking in later analyses. If patterns of post-manipulation liking still emerge controlling for baseline liking, it provides some additional evidence consistent with our theorized account.

[9] An additional goal of study 5 was to include an exploratory moderator, namely, Need for Cognitive Closure (NCC) that could help explain participants' resistance to shifting their feelings towards opponents in a more positive direction. However, as this moderator did not add much useful clarity in answering this question, the findings are reported in the supplemental materials only.

Figure 7

Updated Tweets for Study 5

Participants

American participants from Cloud Research were tested; full demographics can be found in Table 17. Participants who failed the attention check (n = 71), or who identified as "both equally" (n = 56) were excluded. Given our sample size used in analyses, an alpha level of .05, and minimum power of .80 (Cohen, 1992), sensitivity analyses using GPower indicated that for an ANOVA with main effects and interactions, we could detect an effect size of f = .13 (a small effect size).

Table 17

Demographics for Study 5

Factor	
N	550

Age	
Mean (SD)	40.05 (12.98)
Gender	
% Female	52.2
% Male	46.5
% Other	1.1
Race	
% White	80.3
% Black	7.0
% Hispanic/ Latino	4.1
% Other	8.0
Political Orientation	
Liberal	316
Conservative	234
Political Party	
Democrat	283
Republican	207
Libertarian	30
Green	15
Other	15

Note: Demographics are reported after all exclusions were in effect.

Procedure

Participants, as in previous studies, began by seeing the definition of liberal and conservative, and then provided their political affiliation. Following this, they reported how they felt about several groups, including liberals and conservatives.[10]

The rest of the study was the same as studies 3 and 4, wherein participants were shown either a liberal or conservative tweet, with either no response, a single response, or a "dogpile", as well as all the same questions regarding opponent and own agreement, liking, and willingness to engage.

[10] Participants also completed the NCC scale, as well as the Social Vigilantism scale (used as filler and will not be reported).

Participants were paid $2.00 USD upon completion.

Materials

Questions regarding own and opponent agreement with tweets, feelings towards opponent partisans (liberals: α = .86, conservatives: α = .90), and willingness to engage with opponents (α = .74) were the same as studies 3 and 4. As patterns were similar between opponent elites and partisans, questions regarding elites were dropped.

Feeling Thermometer. Participants rated how they felt about a number of groups on a scale from (0) Very Cold, to (100) Very Warm. These groups included: liberals, conservatives, people who earn $250,000 a year or more, members of the LGBTQ community, Evangelical Christians, and union members. This measure served as our baseline for feelings towards the opponent; as such, we were primarily interested in using this measure as a control for later analyses (specifically in our serial mediation model), and included the additional groups as filler.

Results

Effect of Dissent

As in previous studies, we ran a 3x2 ANCOVA comparing opponent agreement estimates across conditions and partisanship, controlling for participants' feelings towards opponents at the start of the survey. We chose to use initial opponent feelings as a covariate, as this would indicate whether participants' feelings towards opponents were altered following the manipulation over-and-above their feelings at the outset. As predicted, we found a main effect of condition, $F(2, 541) = 12.08$, $p < .001$, $\eta^2 = .043$. LSD post-hoc tests revealed that participants believed the extreme attitude was less prevalent among their opponents when the tweet was followed by a dogpile of dissent compared to a single dissenting response, or no responses at all (see Table 18 for descriptives and post-hoc comparisons).[11] No significant interaction was found, $F(1,541) = .622$, $p = .537$, $\eta^2 = .002$, nor was there a main effect of partisanship.

[11] These results held when running the analyses without the covariate.

Table 18

Opponent Agreement Estimations by Condition for Study 5

	Single Response	Dogpile	Control	Actual Opponent Agreement
	M (SD)	M (SD)	M(SD)	
Overall	55.00_b (27.11)	45.88_a (28.09)	59.47_b (26.33)	25%

Note: Subscripts denote differences at the $p < .05$ level; post-hoc tests were run using Bonferroni corrections.

Feelings Towards Opponent Partisans and Engagement

Again, 3x2 ANCOVAs were run using opponent condition and partisanship as the predictors, and feelings towards opponents and willingness to engage as separate outcomes, with initial opponents' feelings as a covariate. There were no main effects of condition for either feelings towards opponents, $F(2, 541) = .42$, $p = .655$, $\eta^2 = .002$, or engagement, $F(2,541) = .956$, $p = .385$, $\eta^2 = .004$ (see Table 19).

Additionally, there were no significant interactions for either opponent feelings $F(2, 541) = 2.44$, $p = .088$, $\eta^2 = .009$, or engagement, $F(2,541) = .918$, $p = .400$, $\eta^2 = .003$. There was, however, a main effect of partisanship on both feelings towards opponent, $F(1,541) = 20.80$, $p < .001$, $\eta^2 = .037$, and engagement, $F(1,541) = 7.57$, $p = .006$, $\eta^2 = .014$; conservatives felt more positively toward, and more willing to engage with the opponent than liberals (see Table 20 for means).

Table 19

Participant Feelings Towards and Engagement with the Opponent

		Single Response	Dogpile	Control
		M (SD)	M (SD)	M(SD)
Overall	Opponent Feelings	3.48 (1.06)	3.42 (1.07)	3.28 (1.11)
	Engage	4.02 (.90)	3.85 (.90)	3.85 (.90)

Note: Higher "feelings" scores correspond to more positive feelings.

Table 20

Liberal and Conservative Feelings Toward, and Engagement with the Opponent

	Liberals	Conservatives
	M (SD)	M (SD)
Opponent Feelings	3.21 (1.10)	3.65 (1.00)
Engagement	3.80 (.92)	4.05 (.86)

Indirect effect of response condition on engagement. We then ran the serial mediation model outlined in previous studies (see Figure 5 for conceptual model), using our existing coded contrasts; all effects reported are partially standardized (see Table 21 for pathway coefficients). However, in addition, we used participant ratings of opponents on the Feeling Thermometer administered at the start of the survey as a covariate, as these ratings were correlated with opponent agreement estimates with the extreme tweet, $r(546) = -.142$, $p < .001$.

We find a significant indirect effect of D2 (Control vs. Dogpile) on willingness to engage through opponent estimations and opponent feelings (effect: .046, 95% CI [.024, .075], controlling for feeling thermometer scores towards opponents at the outset of the survey. Additionally, we find a significant indirect effect of D3 (Single Response vs. Dogpile) on willingness to engage through opponent estimations and opponent feelings (effect: .033, 95% CI [.013, .058]), also controlling for feeling thermometer scores. This was not the case for D1 (Control vs. Single Response).

Table 21

Pathway Coefficients for Study 5 Serial Mediation

Mediator		*a1*	*a2*	*b1*	*b2*	*d*	*c*	*c'*
Partisan Feelings	D1	-3.77	.05				.10	.06
	D2	-13.45***	.00	-.001	.35***	-.009***	-.02	-.07
	D3	-9.69***	.05				.11	-.13

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Study 5 Discussion

Again, we found that participants who saw a "dogpile" of dissent changed their estimates of opponent agreement with an extreme tweet the most, compared to just a single response, or no response at all. While in study 3 and study 4, we found an interaction between condition and partisanship, here we found that liberals and conservatives were equally affected by the dissenting responses.

Though, again, there were no main effects of condition on feelings towards, or willingness to engage with opponents. However, the serial mediation testing the indirect relationship between exposure to dissent and willingness to engage with opponents was replicated. Further, this replication held even when controlling for participants initial feelings towards opponents, before they saw any tweets or dissent. This suggests that dissent indirectly influences participants' feelings towards political opponents, over-and-above how they felt going into the survey.

General Discussion

Partisans often hold negative views of their opponents in part because they dramatically overestimate the prevalence of their opponents' most extreme, noxious, even caricatured views (Parker et al., 2023). In the current research, we sought to counter these miscalibrated views by presenting individuals with counter-stereotypical examples of opponents, specifically ones who speak out against those extreme views on Twitter. Across five studies, seeing political opponents dissent against the extreme, caricatured views often associated with their "side" reduced how prevalent participants believed those attitudes to be amongst their political opponents. Particularly, when participants saw a "dogpile" of dissent (i.e., multiple members of the opposing political party speaking up against a fellow member's extreme tweet), this accumulation of evidence decreased how prevalent participants thought those extreme attitudes were amongst opponents. Seeing dissent on social media could therefore be one potential avenue in correcting misperceptions individuals may have regarding the kind of people their political opponents are. In some cases, even a single dissenter was enough to make the opponent group appear much less like a homogenous cluster, since the dissenting responses are evidence that there are varied opinions within the group.

These findings are consistent with past work using similar methodologies in interventions meant to correct misperceptions (Lees & Cikara, 2019) and stereotypes (Brauer & Er-rafiy, 2011; FitzGerald et al., 2019), but extends these insights in unique ways. Notably, our approach focused on delivering an intervention in a context where misperceptions are likely to be formed and reinforced, by presenting a simulated (but plausible) Twitter interaction. This approach allowed us to attempt to correct misperceptions as they were occurring, as participants were seeing the dissent immediately following the initial egregious tweet (a tweet meant to make opponent-focused misperceptions salient). Instead of using statistics, or descriptions of outgroup members out-of-context, our approach employed more ecologically valid exemplars of "real" outgroup members challenging the stereotype (Kubin et al., 2021). Thus, the current package contributes to the growing pool of knowledge regarding effective approaches to ameliorating intergroup relations through the correction of misperceptions.

In our initial studies, a single dissenting response was, overall, enough to reduce the estimates of prevalence. However, in our subsequent studies comparing the effects of a single dissenting response to the effects of multiple dissenting responses, a single dissenter ceased to be effective at changing participants' perceptions of their opponents' attitudes. This might have been because these studies were all conducted in times of increasing political turmoil. Between our first study in May 2019, and the last study in February 2021, the United States experienced a number of events contributing to social upheaval and political division, including the start of the COVID-19 pandemic, protests against police brutality, and the January 6th insurrection (wherein extremist Trump supporters stormed the Capitol building), among others. Each of these historic events involved a wide range of behaviours, but the ones most likely to be receiving media attention were often from the political fringes reacting in extreme, egregious, sometimes violent ways. These spotlighted actions then perpetuated the idea that those actors (and their often-reprehensible behaviours) represented common values of their respective sides. During the first few studies, prior to many of these events, a single dissenting response might have been enough to shift perceptions of opponents, but as the social climate became increasingly hostile, it is possible that more evidence was needed to convince participants that their perceptions may have been incorrect.

Although our central goal was to reduce prevalence over-estimates, our secondary goal was to alter patterns of cross-party animosity. We were not able to change feelings towards opponents directly (i.e., people receiving the dogpile intervention did not report greater liking of their opponents relative to people receiving the single dissenter intervention, or no intervention). The lack of direct effects of our intervention on opponent liking may be because of how entrenched animosity has become among political opponents, particularly as these extreme, morally reprehensible exemplars have become so salient. However, we *were* able to establish a consistent indirect relationship between exposure to dissent and willingness to engage with opponents through reduced prevalence estimates and positive opponent-targeted feelings. Across all studies, we successfully replicated a serial mediation model which demonstrated that seeing a dogpile of dissent (as compared to a single dissenting response, or no response at all) decreased the perceived prevalence of opponents' extreme attitudes, and that this decrease predicted more positive feelings towards opponents. Positive feelings towards opponents subsequently predicted greater willingness to engage with them. While we did not find consistent total effects across these models, some research has suggested that mediation analyses with these characteristics (significant indirect effects with no total effects) are still theoretically significant, and should be explored (O'Rourke & MacKinnon, 2014, 2018; Wohl et al., 2019). We cannot claim that the relationships explored in the model are causal due to its cross-sectional nature (Spencer et al., 2005; Maxwell & Cole, 2007), so these findings should be interpretated with caution; researchers often caution that cross-sectional mediation models are biased, and that a series of well-designed experiments is a superior method of demonstrating causal chains. However, our model replicated across all studies, offering tentative evidence consistent with the proposed causal model. Further, we can conclude that exposure to opponent dissent from extreme views does (causally) reduce prevalence over-estimates, especially for liberals, and when multiple dissenters are observed.

Of course, because the manipulations do not have a total effect on liking, it is possible that prevalence ratings are not shifting patterns of liking, but rather it is the pre-existing patterns of liking that are affecting prevalence ratings. This possibility (that variation in liking at the outset are the main

determinant of later links between prevalence estimates and feelings toward opponents) cannot account for the effects of the manipulation on prevalence ratings, but it is an alternative explanation warranting a more cautious interpretation of the mediation model. Thus, in the final study, we also included a measure of opponent liking at the beginning of the study, as well as the existing measure that followed the manipulation. This allowed us to control for baseline liking in all analyses. The indirect effects of dissent on liking through prevalence estimates held, even holding constant initial opponent dislike. These analyses helped to rule out the possibility that the link from prevalence estimates to liking was in the opposite direction - even though initial liking did indeed predict higher prevalence estimates, experimentally altered prevalence estimates predicted liking even controlling for initial liking.

Finally, there was a consistent and unpredicted main effect of partisanship on feelings towards opponents across a number of studies. Specifically, liberals felt more negatively towards conservatives, compared to conservatives' feelings towards liberals. This may have been because of differences in the types of issues we asked liberals and conservatives about (as they differed between partisans). It may also be, in part, due to the current cultural moment, wherein increasingly extreme Republican policies are being passed, and are subsequently receiving a lot of media attention (e.g., overturning Roe v. Wade). Some past research points to differences in perceived lack of care for certain issues as contributing to greater dehumanization of opponents (Kubin et al., 2022); specifically, those supporting physical distancing during COVID-19 (largely liberals) believed that their opponents did not care about the life-or-death nature of the threat, and morally condemned them for it.

Limitations and Future Directions

The current research has a number of limitations, some of which are due to features of the design that offered both strengths and weaknesses. First, we aimed to capture people's reactions to extreme, caricatured tweets. There is no validated database of such content, and the "hot" issues are often a moving target in the social media "culture war" discourse. We sought to capture these issues by creating tweets that were based on the kinds of discussions we had been witnessing on social media. This means that they relied on our observations and were not generated systematically; while we did conduct one pilot test to

assess the extremity of several potential tweets, not all topics were tested. Despite the limitations, this approach allowed us to create ecologically valid materials, since they were meant to reflect the kinds of attitudes a participant could conceivably run into on social media. However, this meant that we gave up some experimental control in how parallel the attitudes could be between liberals and conservatives (e.g., are the liberal and conservative tweets equally extreme or outrage-inducing?). We did try to address this in later studies by deliberately selecting attitudes that were based on the same social issue (i.e., police conduct), but we recognize that this does not guarantee that the tweets were perceived as equally extreme. Save for study 1 (wherein conservative agreement with the extreme tweet was unexpectedly high), the majority of liberal and conservative participants *did not* agree with the extreme tweet from their respective side; this might suggest that the tweets were, indeed, being perceived by ingroup members as extreme and unrepresentative of the group as a whole.

Secondly, while the Twitter interactions (between an extreme opponent and dissenting opponent(s) were meant to mirror an *in-vivo* social media interaction, they deviated from an actual social media interaction in a few key ways. For example, the dissenting responses, both in the single-response and dogpile conditions, were consistently civil and polite; this was done to ensure that the content of the dissent, and not the tone of the response, was the salient variable being manipulated. This was also done as the responders were meant to reflect a "good role model" of what online dissent *could* look like. However, online interactions often involve incivility, particularly if the topic is political in nature (Sun et al., 2021). Further, the individuals most likely to be commenting publicly on social media are the ones using toxic language, encouraging further toxicity from those who choose to engage with them (Kim et al., 2021). Thus, follow up studies should also examine the influence of response tone, not only on how participants perceive their opponents (i.e., would uncivil dissent be effective at correcting misperceptions?), but also on its influence on participants' desire to engage in the interaction themselves. One can imagine how an uncivil online environment might further stifle those whose dissent would be beneficial.

Additionally, respondents in our fabricated Twitter interactions explicitly identified themselves as liberal or conservative. This was done to ensure participants associated the dissent with the opponent group. However, social media users very rarely do this, leaving readers to use context clues to determine what "side" responders are on; they may be wary of backlash from other members of the group they identify with (Otten & Gordijn, 2014; Packer, 2007). Future studies should determine how explicit versus non-identification of political orientation changes how participants classify responders (i.e., does dissent with no identification lead readers to assume that they are not part of the opponent group?), as well as how it shapes group perceptions.

Lastly, while our work has established that seeing online dissent can have very real benefits regarding the correction of misperceptions, past research has identified the difficulties and social costs of choosing to dissent against one's group (see Packer, 2007 for review). Group members who deviate from the established norms are rejected more often by their group than members who choose to instead conform (Tata et al., 1996), and they are subsequently evaluated much more negatively by fellow ingroup members than an outgroup member behaving in the same way (Abrams et al., 2000). Dissent has a cost, and this cost likely inhibits individuals from choosing to dissent, particularly because research suggests that, under normal circumstances, 90% of users are considered "lurkers" (website users who do not engage or contribute) (Nielsen, 2006). Thus, though dissent has its benefits, this work documents the potential of a process that is all too infrequent in the real world. Future work should endeavor to establish the conditions under which social media users choose to speak up, engage, and publicly dissent against views that are harmful to their groups. Further, it should establish the conditions wherein dissent online is costly (i.e., incites ingroup rejection, derision, or in-fighting), versus when it can be clearly framed as helpful to group health (i.e., improving external perceptions of group values), and thus worth doing.

Conclusion

We are at a point in history where the reduction of inter-party hostility, and the encouragement of intergroup cooperation is of the utmost importance. False polarization, that is, the misperception that our political opponents endorse egregious attitudes, contributes to a cycle of affective polarization, pushing

each side of the political spectrum farther and farther apart. The current research points to at least one way that these false perceptions can be combatted, and how this means of correction, at least indirectly, has the power to ameliorate cross-party animosity. The choice to dissent against an ingroups' stereotyped views is an impactful one, and may encourage others to do the same, potentially giving casual readers a new, improved perspective on groups they may have otherwise continued to despise.

References

Abramowitz, A. I., & Webster, S. W. (2018). Negative partisanship: Why Americans dislike parties but behave like rabid partisans. *Political Psychology, 39*, 119-135. doi:10.1111/pops.12479

Abrams, D., Marques, J. M., Bown, N., & Henson, M. (2000). Pro-norm and anti-norm deviance within and between groups. *Journal of Personality and Social Psychology, 78*(5), 906–912. https://doi.org/10.1037/0022-3514.78.5.906

Ahler, D. J. (2018). The group theory of parties: Identity Politics, party stereotypes, and polarization in the 21st Century. *The Forum, 16*(1), 3–22. https://doi.org/10.1515/for-2018-0002

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics, 80*(3), 964-981. doi:10.1086/697253

Bail, C. (2021). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing.* Princeton, USA: Princeton University Press.

Berry, J. M., & Sobieraj, S. (2013). *The Outrage Industry: Political Opinion Media and the New Incivility*. Oxford: Oxford University Press.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313-7318. doi:10.1073/pnas.1618923114

Brauer, M., & Er-rafiy, A. (2011). Increasing perceived variability reduces prejudice and discrimination. *Journal of Experimental Social Psychology, 47*(5), 871–881. https://doi.org/10.1016/j.jesp.2011.03.003

Brock, D., & Rabin-Havt, A. (2012). *The Fox Effect: How Roger Ailes Turned a Network into a Propaganda Machine.* New York: Anchor Books.

Chambers, J. R., Baron, R. S., & Inman, M. L. (2006). Misperceptions in intergroup conflict: Disagreeing about what we disagree about. *Psychological Science, 17*(1), 38-45. doi:10.1111/j.1467-9280.2005.01662.x

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science, 1*(3), 98-101. doi:10.1111/1467-8721.ep10768783

Columb, C., & Plant, E. A. (2011). Revisiting the Obama effect: Exposure to Obama reduces implicit prejudice. *Journal of Experimental Social Psychology, 47*(2), 499–501. https://doi.org/10.1016/j.jesp.2010.11.012

Crawford, J., Modri, S., & Motyl, M. (2013). Bleeding-heart liberals and hard-hearted conservatives: Subtle political dehumanization through differential attributions of human nature and human uniqueness traits. *Journal of Social and Political Psychology, 1*(1), 86-104. doi:10.5964/jspp.v1i1.184

Crockett M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*, 769–771.

De Wit, J. B., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? an assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology, 27*(1), 110–115. https://doi.org/10.1037/0278-6133.27.1.110

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology, 7*(29). https://doi.org/10.1186/s40359-019-0299-7

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gal, D., & Rucker, D. D. (2011). Answering the unasked question: Response substitution in consumer surveys. *Journal of Marketing Research*, *48*(1), 185–195. https://doi.org/10.1509/jmkr.48.1.185

Gaisbauer, F., Pournaki, A., Banisch, S., & Olbrich, E. (2021). Ideological differences in engagement in public debate on Twitter. *PLOS ONE, 16*(3), e0249241. https://doi.org/10.1371/journal.pone.0249241

Haslam, S. A., Oakes, P. J., McGarty, C., Turner, J. C., & Onorato, R. S. (1995). Contextual changes in the prototypicality of extreme and moderate outgroup members. *European Journal of Social Psychology*, *25*(5), 509–530. https://doi.org/10.1002/ejsp.2420250504

Hendricks, V. F., & Vestergaard, M. (2019). *Reality lost: Markets of attention, misinformation and manipulation.* Springer International Publishing.

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

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Kingzette, J. (2020). Who do you loathe? feelings toward politicians vs. ordinary people in the opposing party. *Journal of Experimental Political Science*, *8*(1), 75–84. https://doi.org/10.1017/xps.2020.9

Klein, E. (2020). *Why We're Polarized.* New York: Avid Reader Press.

Kogut, T., & Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, *18*(3), 157–167. https://doi.org/10.1002/bdm.492

Kubin, E., Kachanoff, F. J., & Gray, K. (2022). Threat rejection fuels political dehumanization. *Social Psychological and Personality Science*, *14*(5), 487–500. https://doi.org/10.1177/19485506211068922

Kubin, E., Puryear, C., Schein, C., & Gray, K. (2021). Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, *118*(6), e2008389118. https://doi.org/10.1073/pnas.2008389118

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*, 565–579.

Landry, A. P., Schooler, J. W., Willer, R., & Seli, P. (2022). Reducing explicit blatant dehumanization by correcting exaggerated meta-perceptions. *Social Psychological and Personality Science*, 1–12. https://doi.org/10.1177/19485506221099146

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

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*, 20200143. https://doi.org/10.1098/rstb.2020.0143

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly, 80*(S1), 378-391. doi:10.1093/poq/nfv045

Manis, M., Shedler, J., Jonides, J., & Nelson, T. E. (1993). Availability heuristic in judgments of set size and frequency of occurrence. *Journal of Personality and Social Psychology, 65*(3), 448–457. doi: 10.1037/0022-3514.65.3.448

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23–44. https://doi.org/10.1037/1082-989x.12.1.23

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences, 119*(16), e2116851119. https://doi.org/10.1073/pnas.2116851119

Moore-Berg, S. L., Hameiri, B., & Bruneau, E. G. (2021). Empathy, dehumanization, and misperceptions: A media intervention humanizes migrants and increases empathy for their plight but only if misinformation about migrants is also corrected. *Social Psychological and Personality Science, 13*(2), 645–655. https://doi.org/10.1177/19485506211012793

Nielsen, J. (2006, October 6). *Participation inequality: The 90-9-1 rule for social features.* Nielsen Norman Group. Retrieved February 10, 2023, from https://www.nngroup.com/articles/participation-inequality/#:~:text=Summary%3A%20In%20most%20online%20communities,for%20almost%20all%20the%20action.

O'Rourke, H. P., & MacKinnon, D. P. (2014). When the test of mediation is more powerful than the test of the total effect. *Behavior Research Methods, 47*(2), 424–442. https://doi.org/10.3758/s13428-014-0481-z

O'Rourke, H. P., & MacKinnon, D. P. (2018). Reasons for testing mediation in the absence of an intervention effect: A research imperative in prevention and intervention research. *Journal of Studies on Alcohol and Drugs, 79*(2), 171–181. https://doi.org/10.15288/jsad.2018.79.171

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin, 112*(3), 536–552. https://doi.org/10.1037/0033-2909.112.3.536

Otten, S., & Gordijn, E. H. (2014). Was it one of us? how people cope with misconduct by fellow ingroup members. *Social and Personality Psychology Compass, 8*(4), 165–177. https://doi.org/10.1111/spc3.12098

Packer, D. J. (2007). On being both with US and against US: A normative conflict model of dissent in social groups. *Personality and Social Psychology Review, 12*(1), 50–72. https://doi.org/10.1177/1088868307309606

Parker, V.A., Feinberg, M., Tullett, A., Lefebvre, P., & Wilson, A.E. (2023). The Ties that Blind: Misperceptions of the Opponent Fringe and the Miscalibration of Political Contempt. *Manuscript in prep. Preprint on Psyarxiv:* https://psyarxiv.com/cr23g/

Pew Research Center. (2014, June 12). *Political polarization in the American public.* Pew Research Center. Retrieved February 20, 2023, from https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2010). Membership status and subjective group dynamics: Who triggers the Black Sheep Effect? *Journal of Personality and Social Psychology, 99*(1), 107–119. https://doi.org/10.1037/a0018187

Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science, 29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066x.54.2.93

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*(6), 845–851. https://doi.org/10.1037/0022-3514.89.6.845

Sun, Q., Wojcieszak, M., & Davidson, S. (2021). Over-time trends in incivility on social media: Evidence from political, non-political, and mixed sub-reddits over eleven years. *Frontiers in Political Science*, *3*. https://doi.org/10.3389/fpos.2021.741605

Tata, J., Anthony, T., Lin, H., Newman, B., Tang, S., Millson, M., et al. (1996). Proportionate group size and rejection of the deviate: A meta analytic integration. *Journal of Social Behavior and Personality*, *11*, 739-752.

Tversky, A., & Kahnema, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology, 45*(5), 961–977. https://doi.org/10.1037/0022-3514.45.5.961

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049–1062. https://doi.org/10.1037/0022-3514.67.6.1049

Webster, D. M., & Kruglanski, A. W. (1997). Cognitive and social consequences of the need for cognitive closure. *European Review of Social Psychology*, *8*(1), 133–173. https://doi.org/10.1080/14792779643000100

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States. *Perspectives on Psychological Science, 10*(2), 145-158. doi:10.1177/1745691615569849

Wohl, M. J. A., Tabri, N., Hollingshead, S. J., Dupuis, D. R., & Caouette, J. (2019). Empathetic collective angst predicts perpetrator group members' support for the empowerment of the victimized group.

Journal of Personality and Social Psychology, *117*(6), 1083–1104.

https://doi.org/10.1037/pspi0000176

Appendix A

Study 1, 2, and 3 Materials

Note that the key materials for the first three studies were identical, and were also used in studies 4 and 5. Materials for studies 4 and 5 include only the new questions that have not been outlined here.

Definition of Liberal and Conservative

Before you get started, we would just like to take a moment to go over some definitions of words you may encounter during this survey.

When we use the word **liberal**, we are referring to individuals who typically vote Democrat, the Green Party, or who don't necessarily vote, but hold liberal perspectives and opinions.

When we use the word **conservative**, we are referring to individuals who typically vote Republican, have supported the Tea Party, or who don't necessarily vote, but hold conservative perspectives and opinions. People who typically vote Libertarian may align with some conservative views but not others.

With this in mind, please hit the arrow button to continue.

Political Opinion

1. If I were to support a political party in the USA, I would support…

- The Republican Party
- The Democratic Party
- The Libertarian Party
- The Green Party
- Other: _____

2. Using the following slider bar, please indicate (by sliding the dot) the point that you believe best represents your overall political orientation.

100% Liberal (0)	100% Conservative (100)

3. On average, on most societal topics, are you:

- More conservative
- More liberal
- Both equally

Percentage Estimates

1. Please estimate what percentage of <u>liberals</u> **at least somewhat agree** with the opinion expressed in this Tweet?

0	10	20	30	40	50	60	70	80	90	100

2. Please estimate what percentage of <u>conservatives</u> **at least somewhat agree** with the opinion expressed in this Tweet?

0	10	20	30	40	50	60	70	80	90	100

3. How much do you personally agree with the opinion expressed in this Tweet?

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

Liking, Trust, and Engagement

1. How much do you like liberals in general?

Not At All (0)	Very Much (100)

2. How much do you trust liberals in general?

Not At All (0)	Very Much (100)

3. How much do you like conservatives in general?

Not At All (0)	Very Much (100)

4. How much do you trust conservatives in general?

Not At All (0)	Very Much (100)

5. Please respond to the following hypothetical scenarios as though they were happening in your life. If the circumstances reflect your actual life situation, respond based on what you would do. However if circumstances are different from your real life, please respond hypothetically regarding what you would do if you were actually in that situation.

a. Imagine, hypothetically, that you have a chance to engage in political discussion on some of the topics listed previously. How willing would you be to **have a political discussion** with a committed member of the opposing political party (a liberal if you're a conservative, and a conservative if you're a liberal)?

Extremely Unwilling	Unwilling	Somewhat Unwilling	Somewhat Willing	Willing	Extremely Willing
1	2	3	4	5	6

b. Imagine, hypothetically, that you have a young adult son or daughter. How negatively or positively would you feel **if your child got engaged** to a committed member of the opposing political party (a liberal if you're a conservative, and a conservative if you're a liberal)?

Extremely Negatively	Negatively	Somewhat Negatively	Somewhat Positively	Positively	Extremely Positively
1	2	3	4	5	6

c. Imagine, hypothetically, that you were interested in meeting potential dating partners. How willing would you be to **go out on a date** with a committed member of the opposing political party (a liberal if you're a conservative, and a conservative if you're a liberal)?

Extremely Unwilling	Unwilling	Somewhat Unwilling	Somewhat Willing	Willing	Extremely Willing
1	2	3	4	5	6

d. How willing would you be to **shake hands** with a committed member of the opposing political party (a liberal if you're a conservative, and a conservative if you're a liberal)?

Extremely Unwilling	Unwilling	Somewhat Unwilling	Somewhat Willing	Willing	Extremely Willing
1	2	3	4	5	6

e. How willing would you be to **share a taxi** with a committed member of the opposing political party (a liberal if you're a conservative, and a conservative if you're a liberal)?

Extremely Unwilling	Unwilling	Somewhat Unwilling	Somewhat Willing	Willing	Extremely Willing
1	2	3	4	5	6

Demographics

1. Please indicate your age: _____

2. Please indicate your gender:
    a. Male

b. Female
c. Other (please specify): _____

3. Race:
    a. White/ Caucasian
    b. Black/ African American
    c. Asian
    d. Hispanic/ Latino
    e. Native Hawaiian/ Pacific Islander
    f. Aboriginal
    g. American Indian/ Alaska Native
    h. Other (please specify): _____

Study 4 Materials

Feelings Towards Political Elites

For the following questions, please think about your feelings towards **political elites**; that is, people who hold positions in the American government. This can include the president, governors, senators, etc.

1. Please rate how much you agree or disagree with each of the following statements regarding **Democratic political leaders.**

Are Democratic political leaders...

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

a. Worthy of respect
b. Likeable
c. Worthy of hatred
d. A threat to democracy
e. Trustworthy

2. Please rate how much you agree or disagree with each of the following statements regarding **Republican political leaders.**

Are Republican political leaders...

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

f. Worthy of respect
g. Likeable
h. Worthy of hatred
i. A threat to democracy

j. Trustworthy

Feelings Towards Regular Citizens

For the following questions, please think about your feelings towards **regular citizens**; that is, people who identify as being liberal or conservative, but are not part of the American government, and don't hold positions of political power.

1. Please rate how much you agree or disagree with each of the following statements regarding **liberals in general.**

Are liberals in general...

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

a. Worthy of respect
b. Likeable
c. Worthy of hatred
d. A threat to democracy
e. Trustworthy

2. Please rate how much you agree or disagree with each of the following statements regarding **conservatives in general.**

Are conservatives in general...

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

a. Worthy of respect
b. Likeable
c. Worthy of hatred
d. A threat to democracy
e. Trustworthy

Study 5 Materials

Feeling Thermometer

We'd like to get your feelings toward a number of groups in the U.S. on a “feeling thermometer.” A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the group at 50 degrees if you don’t feel particularly positive or negative toward the group.

Very Cold		Somewhat Cold			Neutral		Somewhat Warm		Very Warm	
0	10	20	30	40	50	60	70	80	90	100

1. How do you feel towards **liberals**?
2. How do you feel towards **conservatives**?
3. How do you feel toward **people who earn $250,000 per year or more**?
4. How do you feel toward **members of the LGTBQ community**?
5. How do you feel toward **Evangelical Christians**?
6. How do you feel toward **union members**?

Need for Cognitive Closure

Please read the following statements, and rate the degree to which you believe they apply to you.

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6

1. I don't like situations that are uncertain.
2. I dislike questions which could be answered in many different ways.
3. I feel uncomfortable when I don't understand the reason why an event occurred in my life.
4. I feel irritated when one person disagrees with what everyone else in a group believes.
5. I don't like to go into a situation without knowing what I can expect from it.
6. I don't like to be with people who are capable of unexpected actions.
7. I dislike it when a person's statement could mean many different things.
8. I do not usually consult many different opinions before forming my own view.
9. I dislike unpredictable situations.

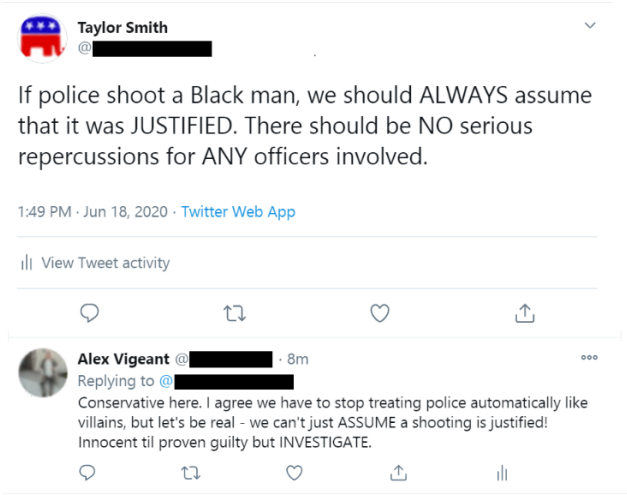
Appendix B
Conservative and Liberal Tweets for Study 3, Study 4, and Study 5

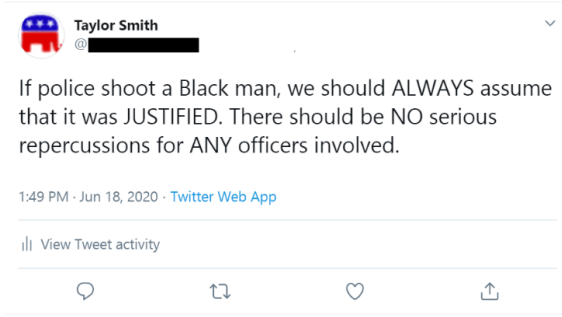
Conservative Tweets for Study 3:

Liberal Tweets for Study 3:

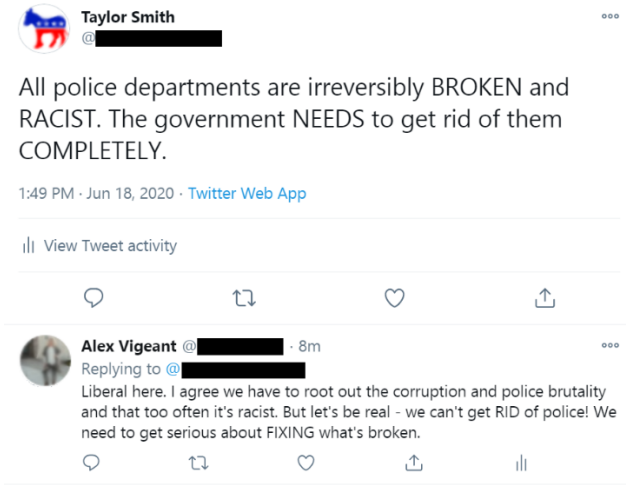
Liberal Tweets for Study 4:

Conservative Tweets for Study 5:





Liberal Tweets for Study 5:



Supplemental Materials

The Power of Dissent: Mitigating False Polarization and Cross-Party Dislike in Online Interactions

These supplemental materials include additional analyses that are referenced in the main text, but that were tangential to the manuscripts' primary findings.

Changes in Perceived Own-Party Agreement with Extreme Tweet

Across all studies in the package, our hypotheses were focused primarily on perceptions of the opponent. However, when designing the studies, we did not restrict participants to *only* seeing Tweets from the opponent (except in study 3); this was so that we could still obtain participants' personal agreement with the Tweets coming from their "side", allowing us to determine whether false polarization was occurring for the Tweeted attitudes.

Table 1, Table 2, and Table 3 display participants' *own-party* agreement estimations between conditions, with subscripts denoting differences; that is, for participants shown extreme Tweets from *their own* party, this is the proportion of *their own party* they believe agree with those extreme Tweets.

We examined differences in these estimates using a one-way ANOVA comparing estimates across the three conditions, followed by an exploration of possible interactions between condition and partisanship, as was done in the main manuscript.

In study 1, there were significant differences between conditions, $F(2, 277) = 7.38$, $p < .001$, $\eta^2 = .051$; it appeared that seeing their own party agree with the Tweet *increased* participants estimates of own-party agreement (the dissent and control conditions did not differ from one another). While there was no interaction between condition and partisanship, $F(2, 274) = .08$, $p = .925$, $\eta^2 = .001$, there was a main effect of partisanship, $F(1, 274) = 94.43$, $p < .001$, $\eta^2 = .256$. Liberals had much lower estimates of own-party agreement than conservatives, though as mentioned in the main text, conservatives had an unexpectedly high degree of agreement with their side's extreme tweet (about illegal immigrants) (see Table 4 for all means).

In study 2, there were no differences in agreement estimations between conditions, $F(2, 352) = .86$, $p = .426$, $\eta^2 = .005$. Again, there was no interaction between condition and partisanship, $F(2, 349) = 2.31$, $p = .101$, $\eta^2 = .013$, however there was a main effect of partisanship, such that conservatives had lower estimates of own-party agreement than liberals, $F(1, 349) = 7.21$, $p = .008$, $\eta^2 = .020$.

In study 4, there were differences between conditions, $F(2, 365) = 3.34$, $p = .037$, $\eta^2 = .018$; seeing a dogpile of dissent lowered participants' own-party agreement estimates, compared to their estimates in the control or single response conditions. There was no interaction between condition and partisanship, $F(2, 362) = 1.46$, $p = .235$, $\eta^2 = .008$, but there was, again, a main effect of partisanship, such that liberals had overall lower estimates of agreement for their own side compared to conservatives, $F(1, 362) = 20.67$, $p < .001$, $\eta^2 = .054$.

In study 5, there were no differences in agreement estimations between conditions, $F(2, 554) = 2.92$, $p = .055$, $\eta^2 = .010$. There was a main effect of partisanship, such that conservatives had overall lower estimates of agreement for their own side compared to liberals, $F(1, 551) = 12.99$, $p < .001$, $\eta^2 = .023$.

Table 1

Own Party Agreement Estimates by Condition for Study 1 and Study 2

		Agree	Disagree	Control	Actual Own-Party Agreement
		M (SD)	M (SD)	M(SD)	
Overall	S1	64.69_b (25.24)	50.04_a (27.45)	56.79_a (28.36)	52%
	S2	42.89 (28.08)	39.40 (26.84)	43.84 (28.44)	32%

Note: Subscripts denote differences at the $p < .05$ level; LSD post-hoc tests were used.

Table 2

Own Party Agreement Estimates by Condition for Study 4

	Single Response	Dogpile	Control	Actual Own-Party Agreement
	M (SD)	M (SD)	M(SD)	
Overall	30.01_b (25.88)	22.99_a (23.44)	30.28_b (25.50)	14%

Note: Subscripts denote differences at the $p < .05$ level; LSD post-hoc tests were used.

Table 3

Own Party Agreement Estimates by Condition for Study 5

	Single Response	Dogpile	Control	Actual Own-Party Agreement
	M (SD)	M (SD)	M(SD)	
Overall	40.31 (27.81)	33.79 (28.63)	39.58 (27.66)	25%

Note: Subscripts denote differences at the $p < .05$ level; LSD post-hoc tests were used.

Table 4

Descriptives for Overall, Liberal, and Conservatives Estimates of Own-Party Agreement and Actual Agreement

		Estimates of Own-Party Agreement	Actual Own-Party Agreement
		M (SD)	
S1	Liberals	44.86 (26.58)	32%
	Conservatives	72.60 (20.04)	77%
	Overall	57.54 (27.51)	52%
S2	Liberals	45.30 (26.36)	41%
	Conservatives	38.60 (28.57)	21%
	Overall	42.39 (27.50)	32%
S4	Liberals	22.99 (22.67)	10%
	Conservatives	34.47 (26.88)	18%
	Overall	27.79 (25.13)	14%
S5	Liberals	41.80 (26.57)	31%
	Conservatives	33.06 (29.34)	17%
	Overall	38.00 (28.12)	25%

Moderated Serial Mediation Models

In study 3 and study 4, we found significant interactions between condition and partisanship on participants' estimates of opponent agreement with the extreme tweet. These were the only studies where this interaction came out as significant, and, in study 4, the interaction appeared to be primarily an artifact of methodological decisions (i.e., we created a tweet that conservatives did not think many liberals agreed with, regardless of condition). As such, we ran moderated mediations for both study 3 and study 4 (see Figure 1 for conceptual model), wherein we tested whether partisanship was moderating the relationship between condition and agreement estimates, and further, whether the overall moderated mediation was significant. Path coefficients are reported in Table 5, and all indices of moderated mediation, and indirect effects are reported in Table 6 (for study 3), and Table 7 and 8 (for study 4).

Overall, we found that the effect of the dogpile was stronger for liberals than for conservatives in study 3 when compared with the control condition; that is, liberals in the dogpile condition (compared to the control condition) lowered their estimates of opponent agreement with the extreme tweet significantly more than conservatives. Further, for both liberals and conservatives in the dogpile condition (compared to the control condition), these lowered estimates correlated with more positive feelings towards the opponent, which, in turn, was related to greater willingness to engage with them. In study 4, conservatives' estimates of opponent agreement were unaffected by the condition, while liberals (specifically in the dogpile condition) were.

However, these results should all be interpreted with caution, as the condition/ partisanship interaction was not consistent across studies, effect sizes were quite small, and the indices of moderated mediation for the tested models were not consistently significant.

Figure 1

Moderated Mediation Model for Study 3 and Study 4

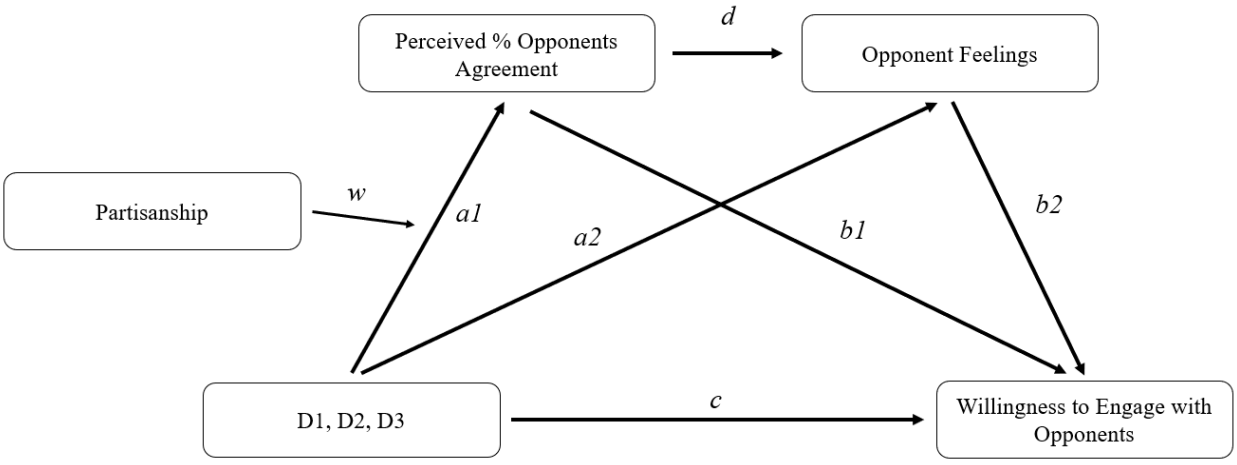


Table 5

Pathway Coefficients for Moderated Mediation Models in Study 3 and Study 4

	Mediator		*w*	*a1*	*a2*	*b1*	*b2*	*d*	*c*
S3	Liking + Trust	D1		-6.60	-2.03	-.003*	.02***	-.13***	-.02
		D2		-1.17	-.61				.10

		D3		5.43	1.42			.12
		Party	.98					
		D1xParty	-.98					
		D2xParty	-11.47*					
		D3xParty	-10.50*					
S4	Partisan Feelings	D1		-11.75	-.15			-.03
		D2		18.49	-.31*	-.004*	.39***	-.01***	-.11
		D3		30.24***	-.15			-.11
		Party	23.28***					
		D1xParty	5.89					
		D2xParty	-18.92**					
		D3xParty	-24.81***					
	Partisan Threat	D1		-11.75	.08			-.04
		D2		18.49	.25	-.006***	-.18***	.01***	-.18
		D3		30.24***	.17			-.12
		Party	23.28***					
		D1xParty	5.89					
		D2xParty	-18.92**					
		D3xParty	-24.81***					

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 6

Indirect Effects and Indices of Moderated Mediation for Study 3

				95% CI	
		Party	Effect	Lower	Upper
Condition – Estimate – Engage					
	D1	Conservative	.022	-.002	.062
		Liberal	**.025**	.004	.056
			Index of Moderated Mediation: .003, 95% CI [-.030, .037]		
	D2	Conservative	**.037**	.005	.084
		Liberal	**.071**	.015	.135
			Index of Moderated Mediation: **.034**, 95% CI [.002, .083]		
	D3	Conservative	.015	-.011	.053
		Liberal	**.050**	.008	.088
			Index of Moderated Mediation: .031, 95% CI [-.002, .080]		
Condition – Like/ Trust – Engage					
	D1		-.032	-.101	.039

	D2		-.010	-.083	.068
	D3		.022	-.051	.094
Condition – Estimate – Like/ Trust - Engage					
	D1	Conservative	.015	-.001	.039
		Liberal	**.017**	.004	.035
		Index of Moderated Mediation: .002, 95% CI [-.020, .023]			
	D2	Conservative	**.025**	.005	.054
		Liberal	**.048**	.018	.083
		Index of Moderated Mediation: **.023**, 95% CI [.002, .052]			
	D3	Conservative	.010	-.008	.034
		Liberal	**.031**	.011	.057
		Index of Moderated Mediation: .021, 95% CI [-.001, .050]			

Note: Bolded effects are significant at the $p < .05$ level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 7

Indirect Effects and Indices of Moderated Mediation (through Opponent Feelings) for Study 4

			95% CI	
	Party	Effect	Lower	Upper
Condition – Estimate – Engage				
D1	Conservative	.021	-.017	.078
	Liberal	-.000	-.033	.034
	Index of Moderated Mediation: -.021, 95% CI [-.086, .028]			
D2	Conservative	.002	-.044	.048
	Liberal	**.070**	.008	.148
	Index of Moderated Mediation: **.068**, 95% CI [.003, .161]			
D3	Conservative	-.020	-.075	.020
	Liberal	**.070**	.008	.148
	Index of Moderated Mediation: **.090**, 95% CI [.008, .202]			
Condition – Partisan Feelings – Engage				
D1		-.060	-.155	.035
D2		**-.127**	-.232	-.029
D3		-.067	-.167	.028
Condition – Estimate – Partisans Feelings - Engage				
D1	Conservative	.026	-.021	.081

	Liberal	-.000	-.039	.037
	Index of Moderated Mediation: -.026, 95% CI [-.096, .030]			
D2	Conservative	.002	-.047	.056
	Liberal	**.085**	.041	.140
	Index of Moderated Mediation: **.083**, 95% CI [.020, .158]			
D3	Conservative	-.023	-.079	.022
	Liberal	**.085**	.042	.141
	Index of Moderated Mediation: **.109**, 95% CI [.045, .190]			

Note: Bolded effects are significant at the $p < .05$ level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Table 8

Indirect Effects and Indices of Moderated Mediation (through Opponent Threat) for Study 4

			95% CI	
	Party	Effect	Lower	Upper
Condition – Estimate – Engage				
D1	Conservative	.034	-.026	.112
	Liberal	-.000	-.049	.051
	Index of Moderated Mediation: -.035, 95% CI [-.130, .040]			
D2	Conservative	.003	-.063	.075
	Liberal	**.113**	.041	.204
	Index of Moderated Mediation: **.111**, 95% CI [.003, .161]			
D3	Conservative	-.032	-.104	.030
	Liberal	**.113**	.044	.200
	Index of Moderated Mediation: **.145**, 95% CI [.047, .275]			
Condition – Partisan Feelings – Engage				
D1		-.020	-.090	.050
D2		-.062	-.136	.002
D3		-.043	-.117	.026
Condition – Estimate – Partisans Feelings - Engage				
D1	Conservative	.013	-.010	.042
	Liberal	-.000	-.020	.017
	Index of Moderated Mediation: -.013, 95% CI [-.051, .015]			
D2	Conservative	.001	-.024	.028
	Liberal	**.042**	.017	.076

		Index of Moderated Mediation: **.041**, 95% CI [.010, .083]		
D3	Conservative	-.012	-.039	.011
	Liberal	**.042**	.016	.078
		Index of Moderated Mediation: **.053**, 95% CI [.018, .077]		

Note: Bolded effects are significant at the $p < .05$ level. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Moderating Effects of Need for Cognitive Closure (NCC)

Across all studies, we had been unable to shift participants' liking or trust of their opponent, nor willingness to engage with them, following the tweets (i.e., no consistent effects of condition) despite successfully shifting their perceptions of attitude prevalence. This may have been because, for some participants, the shift in perceptions of attitude prevalence was not enough to change their overall views of the opponent; seeing one instance of dissent may not erase their opinion that opponents are monstrous, and thus, worthy of dislike and distrust. We speculated that perhaps some of participants' hesitation to change their opponent-targeted feelings may have been because of a more enduring trait. Specifically, we chose to include the Need for Cognitive Closure (NCC) scale (Webster & Kruglanski, 1994) in the final study to explore their relation to both participant agreement estimates, and their subsequent feelings towards opponents. Having a high NCC is related to a greater preference for evidence of prototypicality, as well as a greater reliance on previous knowledge when making judgements (Webster & Kruglansky, 1997). This might suggest that those high in NCC are going to be more resistant to updating their views of and beliefs about opponents even in the face of evidence.

We included the following measure of NCC into study 5:

Need for Cognitive Closure. We used nine items from the need for cognitive closure scale (Webster & Kruglanski, 1994); two from the ambiguity avoidance subscale, two from the need for predictability subscale, and two from the closed-mindedness subscale ($\alpha = .86$). These subscales seemed most applicable to the suspected experience of participants when faced with opponent dissent. Compared to the two remaining subscales (need for order and decisiveness), the three subscales we assessed gauge most directly an individual's desire to obtain accurate and predictable information on their opponent. Participants rated their agreement with each item on a 6-point scale.

We used PROCESS to test whether liberal or conservative participants with a high need for cognitive closure were less convinced by the dissenting opponents (see Figure 2 for conceptual model), and included initial feelings towards opponents as a covariate. Our condition variable was multicategorical; D1 represented Control vs. Single Response, D2 represented Control vs. Dogpile, and D3 represented Single Response vs. Dogpile.

The interactions between NCC and political orientation did significantly moderate the relationship between condition contrasts and opponent agreement estimations, $F(2, 535) = 4.30$, $p = .014$. When looking at the contrasts at each level of NCC between liberals and conservatives (see Table 9, and Figure 3), differing patterns emerge. For liberals, regardless of their level of NCC, being in the dogpile condition (compared to the control or single response) lowered their opponent agreement estimates. For conservatives, it appears that the single response condition (compared to the control) lowered their estimates when they were low in NCC, but those high in NCC needed the extra evidence provided by the dogpile to lower their estimates.

Figure 2
Conceptual Moderation Model for Study 5

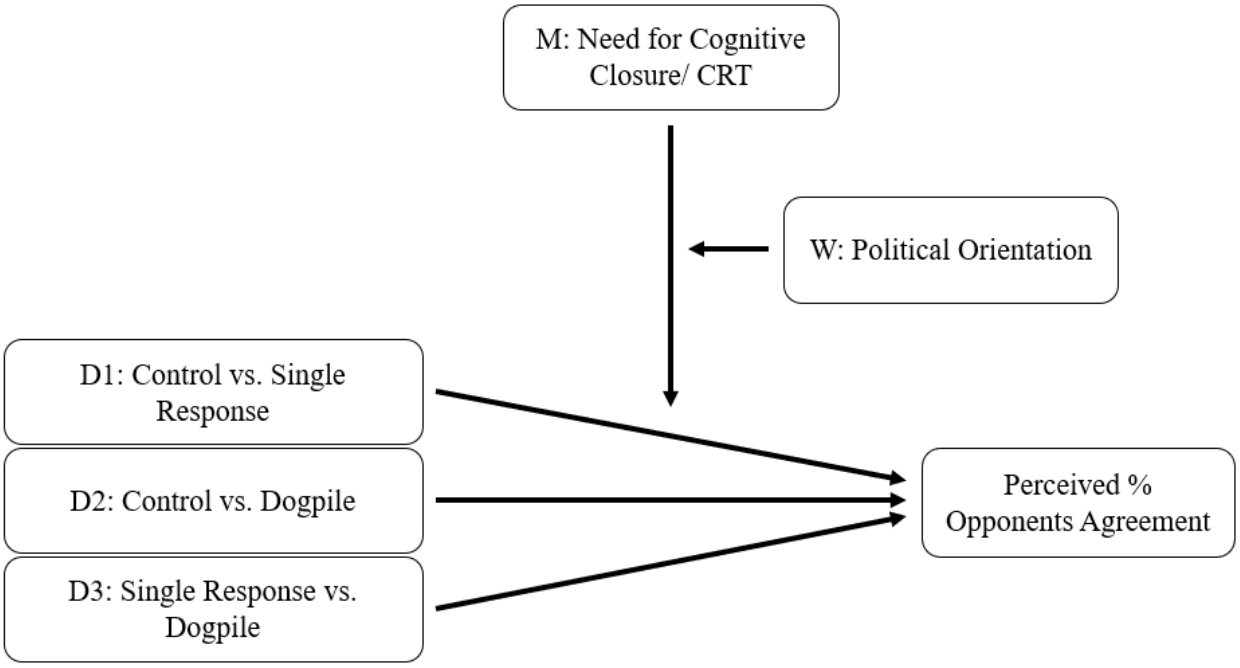


Table 9
Conditional Effects at Each Level of NCC

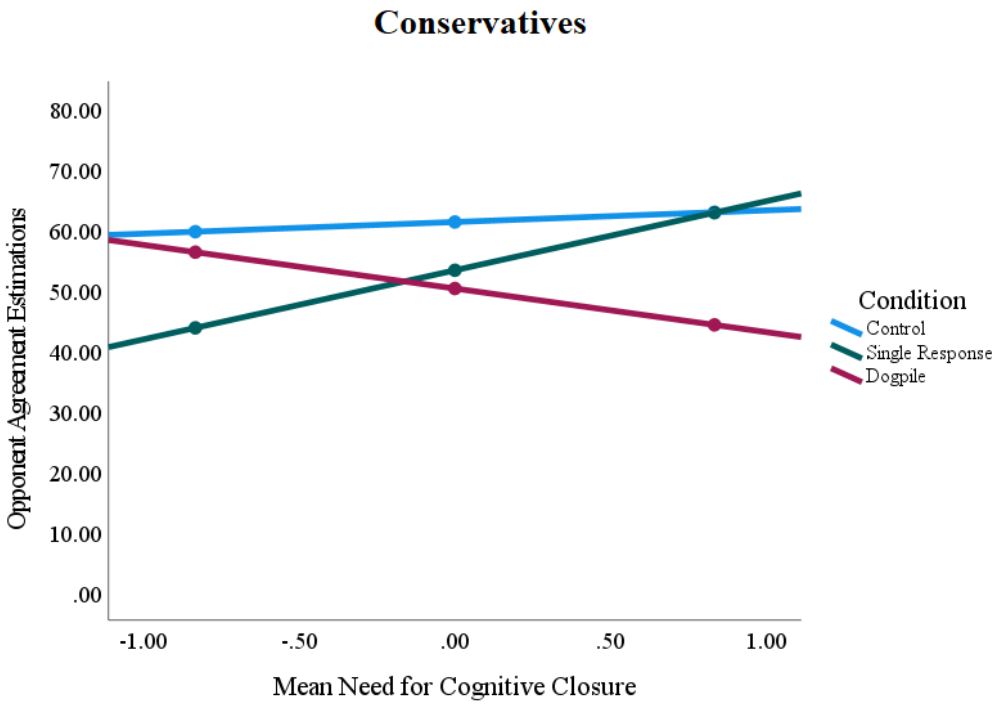
					95% CI	
		Effect	*t*	*p*	Lower	Upper
	Low NCC					
Conservative	D1	-16.45	-2.13	.033	-31.562	-1.342
	D2	-2.75	-.35	.722	-17.935	12.436
	D3	13.70	1.92	.056	-.342	27.746
Liberal	D1	-4.13	-.90	.367	-13.129	4.860
	D2	-17.83	-3.70	<.001	-27.284	-8.371
	D3	-13.69	-2.86	.004	-23.088	-4.299
	Average NCC					
Conservative	D1	-8.125	-1.80	.073	-17.017	.768
	D2	-11.02	-2.46	.014	-19.796	-2.248
	D3	-2.898	-.68	.500	-11.330	5.535
Liberal	D1	-1.35	-.36	.718	-8.698	5.996
	D2	-14.38	-3.89	<.001	-21.636	-7.115
	D3	-13.02	-3.45	<.001	-20.432	-5.616
	High NCC					
Conservative	D1	.20	.03	.974	-11.952	12.358
	D2	-19.29	-3.14	.002	-31.352	-7.237

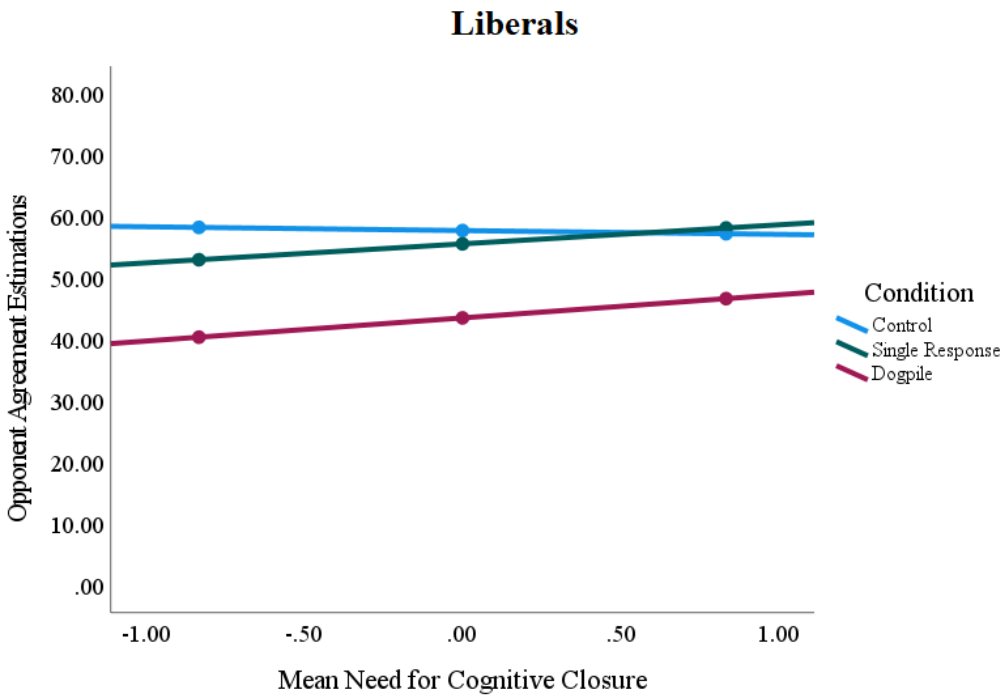
	D3	-19.50	-3.25	<.001	-30.563	-8.432
	D1	1.43	.26	.795	-9.410	12.275
Liberal	D2	-10.92	-2.11	.035	-21.087	-.758
	D3	-12.35	-2.20	.028	-23.364	-1.346

Note: NCC was centered; "low" refers to 1 SD below mean, and "high" refers to 1 SD above mean. Bolded effects are significant. D1 represents the Control vs. Single Response contrast, D2 represents the Control vs. Dogpile contrast, and D3 represents the Single Response vs. Dogpile contrast.

Figure 3

Graph of Conditional Effects at Each Level of NCC for Conservatives and Liberals





We also examined correlations between our key variables of interest, NCC, and political orientation (using a continuous measure of political orientation) (see Table 10). There were no correlations between NCC and agreement estimates. However, those high in NCC reported liking the opponent less, and being less willing to engage with them. Lastly, the more conservative participants were, the greater their need for cognitive closure.

Table 10

Correlations Between Key Variables, and NCC

	1	2	3	4	5
1. Opponent Agreement					
2. NCC	.06				
3. Partisan Feelings	-.31***	-.07			
4. Engagement	-.17***	-.14***	.56***		
5. Political Orientation	.01	.19***	.22***	.18***	

Note: *** indicates p <.001, ** indicates p < .01, * indicates p < .05.