ACTUAL AND ANTICIPATED REACTIONS TO ENGAGING WITH AND DISMISSING POLITICAL OPPONENTS: WHO AND WHERE THEY COME FROM, AND WHY THEY MATTER

by

Gordon Heltzel

M.A., University of British Columbia, 2019

B.A., Indiana University, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November 2023

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Actual and anticipated reactions to engaging with and dismissing political opponents: Who and where they come from, and why they matter

Submitted by Gordon Heltzel in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Psychology

Examining Committee:

Dr. Kristin Laurin, Professor, Department of Psychology, UBC
Supervisor

Dr. Steven J. Heine, Professor, Department of Psychology, UBC
Supervisory Committee Member

Dr. Frances Chen, Associate Professor, Department of Psychology, UBC
Supervisory Committee Member

Dr. Ara Norenzayan, Professor, Department of Psychology, UBC
University Examiner

Dr. Paul Quirk, Professor, Department of Political Science, UBC
University Examiner

Abstract

Growing political polarization has fueled calls for people to constructively engage with opponents and better understand their perspectives rather than dismissively avoiding or condemning them. People who heed these calls may be doing their part to benefit democracy, but what about their reputations—will their behavior elevate them or abase them in their allies' eyes? My dissertation reports ten studies answering this and related questions. Building on my MA thesis, which finds that people usually like allies who constructively engage with opponents' views, Chapter 2's Studies 1 and 2 examined why they hold this preference and when it is most likely to emerge. In Chapter 3, Studies 3 and 4 found a case when people prefer the opposite: U.S. Senators' tweets received more positive feedback when they dismissed opponents compared to engaging with them. Studies 5 and 6 (and Appendix Studies S1-S4) test various explanations for this contradictory pattern, finding that Twitter popularity represents the genuine preferences of a small group of active users with unusual attitudes, as well as inauthentic preferences expressed by everyone else. Drawing on this observation that popular opinion is not represented on (social) media, Chapter 4 considered whether people fail to realize that their allies endorse cross-party engaging. Indeed, Studies 7 and 8 find that people mistakenly think they are alone in preferring allies who engage over those who dismiss. I theorized that perceived polarization causes these misperceptions, but Studies 9 and 10 found that reducing perceived polarization does not reduce misperceptions nor encourage people to engage with opposing views. This work on one hand highlights reputational benefits of engaging with opposing views; on the other, it suggests social media distorts these benefits, and people generally fail to realize them. At the same time, it leaves open how interventions might motivate engagement with opposing views.

Lay Summary

Americans have polarized politically, prompting calls for cross-party engagement. How are people who heed this call received by peers from their political camp? In one series of studies, most Americans applauded cross-party engagement; in another, U.S. Senators' tweets modeling such behavior received less positive feedback (Likes, Retweets) because people providing feedback, unlike most, are extremists who disavow constructive cross-party engagement. Since these results suggested social media misrepresents majority preferences, additional studies tested whether people realize their peers support cross-party engaging. They do not: Most people like political allies who engage constructively with opposing views, yet think they alone feel this way and expect such behavior to provoke backlash. Further studies tried (but failed) to correct this by rectifying people's overblown perceptions of polarization. Taken together, this dissertation helps explain why Americans do not bridge divides and highlights a potential solution: raising awareness of widespread support for this behavior.

Preface

The studies reported in Chapter 2 are published in *Psychological Science*; those reported in Chapter 3 are under review at an academic journal. Gordon Heltzel is the lead author on both articles. Dr. Kristin Laurin supported analysis and contributed edits and feedback on the writing of the manuscript. Gordon Heltzel led data collection, analysis, data visualizations, and writing. The conceptualization and design of these studies was co-led by Gordon Heltzel and Dr. Kristin Laurin.

- Heltzel, G., & Laurin, K. (2021). Seek and ye shall be fine: Attitudes toward political-perspective seekers. *Psychological Science*, *32*(11), 1782-1800.
- Heltzel, G., & Laurin, K. Why Twitter sometimes rewards what most people disapprove of: The case of cross-party political relations (In prep).

Gordon Heltzel led the writing of Chapters 1, 4, and 5, as well as data collection, analysis, and visualizations of the results reported in Chapter 4. Dr. Kristin Laurin supported the writing of these chapters, data collection, analysis, and visualizations. Gordon Heltzel and Dr. Kristin Laurin both contributed to the conceptualization and design of these studies.

All studies included in this dissertation have received approval from the UBC Ethics review board (Chapter 2: H17-02530, H18-02799; Chapter 3: H18-02303; Chapter 4: H18-02733, H21-00150).

Table of Contents

List of Tables

List of Figures

Acknowledgements

To Kristin Laurin, for patiently and kindly channeling my enthusiasm for psychology to help me grow as a scholar, person, and especially as a writer. This dissertation's best parts trace to you. To my committee, for their thoughtful feedback (especially Steve Heine, for wisdom lent to my MA and PhD committees and for letting me dogsit Ponto—give him a belly rub for me!). To my labmates, for invaluable feedback on this work, brilliant ideas on important debates (is guacamole a sauce?), and, most of all, lasting friendships. To my research assistants, for making grad school meaningful and this dissertation possible. I'm grateful to have met and learned from so many brilliant, kind people.

To Audrey, the best part of grad school, for listening to ideas and complaints, helping me find the right word when a sentence stumps me, and for other things too, but there are too many right words for that grateful sentence. To my family, especially my grandma, aunt Millie and uncle Steve, and mom (who, with Jeff Probst's help, got me into social psychology), for your support from afar and sending love via packages with exotic Oreos or pictures of René and life back home. To René, you've literally consumed a book so you can surely figure out how to read this. Thanks for shepherding mom while I'm away and for excited greetings when I get home. Wiggle on. To my pals Ted Hartog and Adam Alic, whose long calls kept me laughing and grounded (Ted also gets credit for inspiring Chapter 2).

To my Heltzel family, who all passed while or before I was in grad school. Grandpa, my namesake, a high school teacher, modeled that I could teach or even be a professor. Granny, eulogized as a peacemaker, inspired this research. My dad and aunt taught me curiosity and grit, grad school must-haves. You all taught me to make the most of this short time on earth.

Dedication

To Kurt Vonnegut: Reading this from humanist heaven, you will hear through academic garble a band saw cutting galvanized tin with vocabulary as unornamental as a monkey wrench—that is, the voice of a fellow Hoosier.

Chapter 1: Introduction

1.1 Conflict between moral and democratic values

We must end this uncivil war that pits red against blue, rural versus urban, conservative versus liberal. We can do this if we open our souls instead of hardening our hearts. If we show a little tolerance and humility. If we're willing to stand in the other person's shoes just for a moment.

- American President Joe Biden, 2020 Inauguration Speech

The rise of democracies is often considered a crowning achievement of the enlightenment era, helping both citizens and their societies to flourish (Przeworski et al., 2000). One reason democratic societies have done so well is they have developed values (e.g., tolerance; rationality; compromise) and norms (e.g., civil discourse and deliberation) that help ensure protected rights for groups with minority demographics or beliefs, and that harness rational deliberation to identify optimal policies among options (Christiano, 2008; Mill, 1861).

That said, democratic values can conflict with other important values, including morality. When people see their beliefs as based on a fundamental moral truth (Goodwin & Darley, 2008; Skitka et al., 2021), they often see contrasting beliefs as lacking a valid moral basis (Haidt, 2012). As a result, people develop strong negative attitudes towards dissenting individuals or groups: They feel contempt for them (Rozin et al., 1999) and see them as immoral (Meindl et al., 2016) or even subhuman (Haslam & Loughnan, 2014). They also feel justified in avoiding and dismissing outright views that they morally disagree with (Tetlock, 2003), in intolerantly withholding the rights of moral dissenters (Skitka et al., 2015), and (sometimes) in resorting to violence against them (Rai et al., 2017). In sum, moral values can lead people to dismiss adversaries—ostracizing them, overlooking their ideas, encroaching on their rights, or actively harming them—behaviors that violate democratic values of tolerance, compromise, and rationality.

When democratic societies feature political groups steeped in moral conflict, citizens face a dilemma. On one hand, their society's democratic values compel them to tolerantly engage with—learn about and deliberate over—different political views. On the other hand, the moral values that undergird their own political beliefs compel them to at the very least avoid political groups who seem to accept or endorse immoral ideas, and perhaps to actively exclude, silence, or hurt seemingly immoral others. This value conflict, along with most literature that informs my theorizing, has emerged in contemporary America, so my dissertation focuses on this context.

Americans' actions increasingly suggest that they have resolved this conflict in favor of moral values, embodying the sort of intolerant behaviors that both are typical of moral conflict and unravel the fabric of democratic societies. Over time, they have become less likely to engage with political opponents' views (Rodriguez et al., 2017), opting instead to hear from congenial views in the news they watch (Iyengar & Hahn, 2009; Peterson & Iyengar, 2020), the people they interact with (Skitka et al., 2005), and the communities they live in (Motyl et al., 2014). Not only are Americans increasingly refusing to engage with opponents, they are more often choosing to dismiss them outright as well: They explicitly dehumanize them (Finkel et al., 2020; Iyengar et al., 2019), and occasionally take the sort of extreme actions against them that I mentioned earlier (e.g., suppressing opposing views, violence; Cole Wright et al., 2008; Crawford & Pilanski, 2014; Kalmoe & Mason, 2022; Skitka & Morgan, 2014).

In response, some voices—scholars (Ditto & Koleva, 2011), non-profit organizations (e.g., More in Common; Listen First Project), and politicians (e.g., Joe Biden, as quoted at the beginning of this paper)—have called for people to turn the tide and do the opposite: Try to constructively engage with opponents' political views by exposing themselves to their ideas and considering their merits. The people who make these calls typically justify them on the basis of

democratic values (e.g., Mutz, 2002), implying these values should take precedence over moral qualms. But the people being called to engage constructively with opponents may have other values in mind, such as protecting their reputation. They might consider whether engaging would help or harm their reputation, especially in the eyes of their political in-group, whom they most want to impress and remain connected to (Baumeister & Leary, 1995). Based on this idea, my dissertation asks three interrelated questions: (1) what explains how engaging with versus dismissing opponents' views affects one's reputation among allies, (2) how well is this reputational effect represented on social media, and (3) how accurately can people predict it?

Answering these questions in the American political context can be informative for different reasons. Theoretically, America is a democracy amid high polarization (Heltzel & Laurin, 2021), so its citizens face a conflict between their strong democratic values and their strong moral values, which lead them to object to their opponents' views. In places where polarization is weaker, or where democracy is not the norm, this conflict would be more lopsided, and the results less informative. From a practical standpoint, America plays a major role in global affairs, so citizens in other countries might pay attention to and emulate its sociopolitical dynamics. Whether the results I observe in America generalize elsewhere is a question I will discuss more thoroughly in Chapter 5, but it depends in part on whether America is more polarized than elsewhere, which is the subject of much debate (e.g., Westwood et al., 2018; Gidron et al., 2019; Wagner, 2021; Boxell et al., 2022; Fletcher et al., 2019).

1.2 Reputational effects of engaging with versus dismissing opposing perspectives

How do observers react to political allies who engage with, or dismiss, shared opponents' political perspectives? Figure 1.1 illustrates the components involved in this question using

different colored profiles to depict observers and actor as *allies*[1], contrasted with a shared opposing view. Using a variety of methods, my MA thesis consistently found evidence that engaging garners more reputational benefits: Observers strongly preferred politically allied actors who engaged with (green handshake in Figure 1.1) rather than dismissed (red blocking hand in Figure 1.1) opposing target viewpoints.

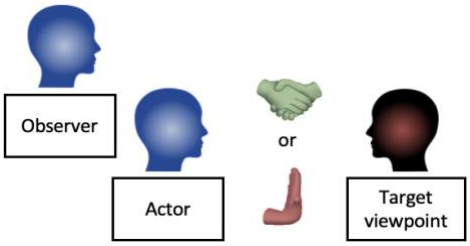


Figure 1.1. Engaging and dismissing dynamics

That work (Heltzel & Laurin, 2021, which includes my MA thesis studies and several more) spoke to related questions and in doing so, bolstered confidence that people genuinely, typically prefer engaging. For example, several studies included an allied actor who neither engaged with nor dismissed political opponents' views; relative to this control target, people *liked more* an ally who engaged with opposing views and *liked less* an ally who dismissed opposing views. In addition, that work showed that people prefer actors who engage over those who dismiss both in the abstract (e.g., when they read about them in hypothetical vignettes) as well as in real, concrete situations (e.g., when interacting with them in-person). Likewise, they disliked dismissers whom they heard explicitly advocate for this approach, as well as those whom they merely observed more passively avoiding hearing from opposing viewpoints (e.g., seeing their news consumption habits). It also showed that this preference generalizes across both U.S. and Canadian samples, with both groups showing a large preference for allies who

[1] I use this term throughout to refer to politically like-minded others: Those who support the same policy stance (i.e., pro-choice), or identify with the same broad ideological orientation (i.e., liberal) or party affiliation (i.e., Democrat).

engage with opponents' views. Finally, that work showed that people prefer engaging regardless of the reason an actor engages with opposing views (e.g., to learn or persuade) or dismisses them (e.g., out of loyalty or to avoid the emotional burden).

The above findings dovetail with evidence that people generally respect and want to be around individuals who express more willingness to hear out opposing political views (Yeomans et al., 2020), as well as evidence that people want allies to speak and behave in ways that facilitate constructive cross-party engagement. For example, people prefer allies—both citizens and leaders—who are civil and polite towards opponents (Frimer & Skitka, 2018, 2020), so much so that they disidentify with their political group after seeing allies behave uncivilly towards opponents (Druckman et al., 2019; Klar et al., 2018).

There are different ways in which people can engage with or dismiss opponents and their viewpoints. For example, engaging with opponents can mean gaining initial exposure to their viewpoints and spending time processing and evaluating them (e.g., Minson & Chen, 2020), or it can be deeper, involving searching for common ground and cooperating with opponents (e.g., Galinsky et al., 2005). Likewise, people can dismissively overlook opposing views by simply not bothering to expose themselves to or think deeply about them (e.g., Heltzel & Laurin, 2021), or they can dismiss outright opponents and their views by condemning them and saying they are not worth hearing out (e.g., Frimer & Skitka, 2018, 2020). These two forms of dismissing respectively map onto, on one hand, the disgust or contempt people feel toward immoral others and that makes them want to withdraw from and avoid them, and on the other hand the anger they feel that makes them want to expel and push away opponents (Hutcherson & Gross, 2011).

Many studies suggest that people prefer engaging, but none explore *why* people prefer it, and some recent studies and narratives would instead support the prediction that they would

prefer the opposite. I therefore developed a process model that could account both for people's overall preference, and for the forces that might push for its opposite.

1.2.1 Engagers might seem tolerant, rational, and cooperative

Observers might like actors who engage because of the socially desirable effects of this action. For example, people who engage thoughtfully with opponents' beliefs often become more tolerant of them (Mutz, 2002) and better appreciate the reasons and experiences informing them (Kubin et al., 2021; Stanley et al., 2020). Perhaps as a result, people who engage with opposing views tend to develop more rational, nuanced, and better-informed beliefs (Golman et al., 2017). And because they better understand their opponents' beliefs and values, they are better able to cooperate with them, finding common ground toward compromise (Galinsky et al., 2005). In other words, people who engage with opposing views may appear, and actually tend to become, more tolerant, rational, and cooperative, each of which garner liking and respect (Fehr & Fischbacher, 2004; Fiske et al., 2006; Ståhl et al., 2016), especially in Western democracies where tolerance, compromise, and rationality are highly regarded (Brown, 2009; Norenzayan et al., 2002).

1.2.2 Engaging may convey openness to validating and adopting opposing views

And yet, most people who have followed media coverage or empirical research on politics over the last decade (e.g., Abramowitz & Saunders, 2008; Finkel et al., 2020; Iyengar et al., 2019; Levendusky & Malhotra, 2016a) will have heard of people's rising moral disdain for opponents, and so might expect Americans to be skeptical of allies who engage with those opponents' views. For example, people usually like tolerant, cooperative individuals, but they dislike co-partisan politicians who tolerantly compromise with opponents (Ryan, 2017; Wolf et al., 2012).

Based on this moral disdain for opponents, there are at least two reasons why people might prefer allies who refuse to validate opponents' immoral views, instead demonstrating loyalty and commitment to the group. For one, people might dislike allies who seem to validate opponents' views by empathizing with or otherwise tolerating them, thereby implying there may be good reasons behind them. After all, people refrain from empathizing with individuals whom they deem to be immoral (Anderson & Cameron, 2023) and dislike anyone else who tries to do the same (Wang & Todd, 2020); since partisans increasingly feel that their opponents' views are morally illegitimate (Finkel et al., 2020; Haidt et al., 2003; Stanley et al., 2020), they might condemn engaging with those views, as this may seem to validate them. For another, whereas people who engage with opposing views may seem alarmingly open to changing their minds on fundamental moral issues, those who dismiss opposing views may seem loyally invested in their political views and willing to take a stance on contentious debates, both of which people like (Silver & Shaw, 2022; Zlatev, 2019).

1.2.3 Testing psychological mechanisms underlying the preference for engaging

Together with my prior findings that people like engaging more than dismissing, this led me to hypothesize that they might have a strong preference for tolerant, cooperative and / or rational allies, somewhat offset by a (comparatively weaker) preference for allies who refuse to legitimize opponents and seem unlikely to change their minds. These preferences should have opposing effects on people's responses to engaging. One straightforward way to test this idea is through statistical mediation through each of these variables.

A second way to test it is through its implied moderators. Consider the case where there is an especially large perceived difference between the observers' and target's perspectives; see Figure 1.2. In this case, observers may find that opposing perspective especially immoral,

alarming, or distasteful, so they might feel relatively more favorable toward allies who refuse to legitimize it and seem unlikely to change their minds—that is, they should have a weaker overall preference for engaging. For example, if an outgroup *target* has extreme views (relative to other targets; see Figure 1.2, bottom panel), an observer will perceive their own views as more at odds with this target's and dislike an allied actor who engages with (vs. dismisses) that outgroup target's views. Likewise, if an *observer* has extreme views (relative to other observers; see Figure 1.2, middle panel), they will perceive their views as more at odds with an outgroup target's and dislike an allied actor who engages with (or dismisses) that outgroup target's views. In such cases—when an actor engages with an extreme opposing target's view and/or when the actor's allied observers are themselves extreme—there should be a weaker overall preference for engaging.

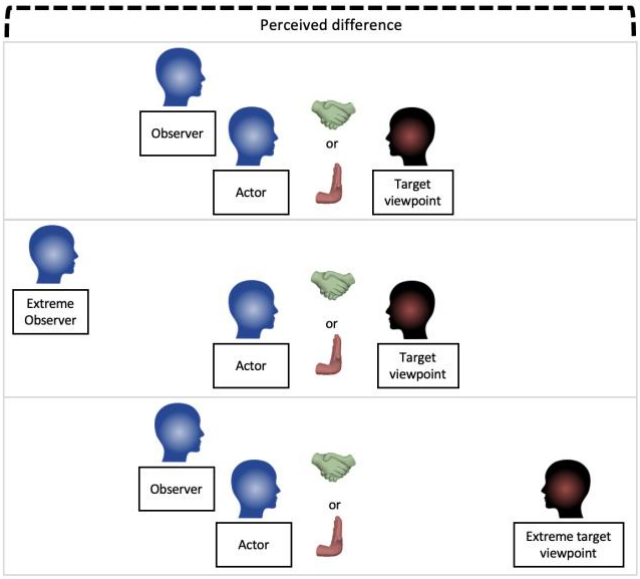


Figure 1.2. Engaging and dismissing dynamics at baseline (top) vs. with a relatively extreme observer (middle) or target viewpoint (right)

Chapter 2 in this dissertation will use both of these approaches—statistical mediation as well as moderation—to test my hypotheses about psychological mechanisms.

1.3 The potentially special case of politicians' tweets

Having established the mechanisms underlying people's general preferences, Chapter 3 studies a specific case where the (implied) preference may reverse: Social media feedback to politicians who engage with or dismiss opponents' perspectives in their posts. This context—social media—has special practical importance, as sites like Twitter (now rebranded as X) have an increasing role in political discourse: They allow politicians to easily communicate and interact with constituents. Since politicians are the prototypical members of their political groups and have uniquely large social media followings, their posts can powerfully shape citizens' perceptions of which behaviors are normal and appropriate; in turn, citizens can influence politicians by responding to their posts.

I therefore asked, which type of post—an engaging or dismissing one—generates more positive feedback on Twitter (e.g., Likes and Retweets). Politicians can safely assume that if they engage with opponents' views, they will gain favor among those opponents (Minson & Chen, 2022), but a politician's social media audience is made up primarily of allies and only very few opponents (Pew Research Center, 2022a; Wojcieszak et al., 2022). Politicians, like most people, may choose to post that which appeals to their allies (Marie & Petersen, 2023). Which post would these allied observers approve of? More broadly, are politicians rewarded more for tweeting about engaging with, or dismissing, opponents and their views? And what does this say about what most citizens want from their political leaders?

1.3.1 People may prefer their leaders' tweets that engage with, rather than dismiss, opponents

Some of the evidence reviewed above suggests people prefer positive interparty interactions in general (e.g., Frimer & Skitka, 2018), and engaging specifically (Heltzel & Laurin, 2021), so perhaps this would seamlessly transfer to online contexts. After all, the broader

evidence that people favor positive interparty interactions has been shown to apply even to politicians, and even to social media. For example, at least in principle, most Americans endorse positive relations between elected representatives: Nearly all (93%) say politicians should compromise and make laws through bipartisan efforts (Pew Research Center, 2019a; see also Harbridge & Malhotra, 2011). They also like partisan allies, including leaders, who express more respect and civility when tweeting about opponents (Frimer & Skitka, 2018, 2020). (In)civility is distinct from engaging versus dismissing—one can politely explain why another view is unworthy of attention or have rude things to say upon considering opponents' ideas in good faith—but they likely cooccur in natural communication. If people prefer positive interparty relations, they might also prefer when ally leaders tweet about engaging with, rather than dismissing, opponents' views.

1.3.2 Politicians' dismissing tweets may fare better than their engaging ones

This broader evidence notwithstanding, even if most people generally prefer positive interparty relations, social media may not reflect that majority preference and may instead reward dismissing over constructively engaging. An emerging literature has documented this kind of disconnect on posts related to engaging and dismissing; for example, although people prefer civil communication (Frimer & Skitka, 2020), tweets that are more uncivil receive more Likes and Retweets (Frimer et al., 2022); likewise, people condemn negativity and divisiveness (Rathje et al., 2022), but negative, divisive posts receive more engagement (Brady et al., 2017; Yu et al., 2023). There are at least two ways such a disconnect could arise with regards to politicians' engaging and dismissing posts.

1.3.2.1 People might genuinely prefer politicians' dismissing tweets

People might genuinely prefer different behavior by their political leaders on Twitter, compared to what they would otherwise prefer. When observers see an allied *politician* engage with opponents' views, they may worry this leader would make tangible policy concessions that an ordinary citizen could not (see also Pew Research Center, 2019a; Ryan, 2017). In addition, people really dislike *opposing politicians* (Kingzette, 2021); when leaders engage with opposing views on Twitter, they are more likely to hear from these opposing politicians (as opposed to opposing regular citizens), so their allied observers might react less favorably (e.g., Hussein & Wheeler, 2023). Moreover, *Twitter* facilitates social conflict (e.g., Crockett, 2017), and people think antagonistic, divisive content is normal there (Rathje et al., 2022). People often approve more of normative behaviors (Eriksson et al., 2015; Lerner, 1980), so to the extent dismissing tweets seem antagonistic and divisive, people may genuinely like these tweets more than engaging ones.

Although people might genuinely prefer dismissing (vs. engaging) by allies when it is communicated by leaders and / or on Twitter, this possibility seems less likely given evidence reviewed above that people (at least claim to) prefer bipartisanship *among politicians* and civil, respectful *tweets* by their leaders.

1.3.2.2 Twitter might reward dismissing even if most people do not prefer it

Alternatively, politicians' dismissing tweets might in fact receive more Likes and Retweets *despite* most allies' genuine preferences. This mismatch could occur if people use Likes and Retweets for reasons other than signaling their preferences, or if the people who provide most Likes and Retweets have atypical preferences.

1.3.2.2.1 What do Likes and Retweets signal?

On their face, Likes and Retweets seem to signal approval. The term "Like" directly conveys approval and Twitter users claim that when they Retweet a post, they typically endorse its message (Metaxas et al., 2017; Pew Research Center, 2021a). Likewise, when users' posts receive Likes and Retweets, they act as if it is a reward, posting similar content more often in the future (Brady et al., 2021). For these reasons, one might assume Likes and Retweets reflect endorsement. But to do so might be misguided: Many users discourage this prevalent assumption with statements in their bio (e.g., "retweet ≠ endorsement") suggesting that they themselves use Twitter reactions *not* as endorsement, but for other reasons: to highlight posts ironically, to raise awareness, or to bookmark for later (Pew Research Center, 2021). Moreover, Like and Retweet counts can be confounded by exposure, or how many people see the post (Frimer et al., 2022). Imagine one tweet seen by ten users with five Likes, and another seen by 1000 users with ten Likes. The former has fewer Likes in absolute terms, but a greater portion of viewers liked it—its smaller Like count may underestimate how most people (would) feel about it. If Like and Retweet counts reflect something other than users' approval, the feedback dismissing and engaging tweets receive may not align with most people's genuine preferences.

1.3.2.2.2 Who Likes and Retweets politicians' posts?

Alternatively, Likes and Retweets on politicians' posts might reflect approval—namely, the approval of whomever tends to Like and Retweet politicians' tweets. A specific subset of people will tend to have their opinions represented by the feedback to politicians' tweets: People who use Twitter, follow politicians, and use Twitter's functions to react to their posts. Few people meet all three of these criteria. Only a quarter of Americans use Twitter, and most do not post political content, such that 80% of political tweets come from only about 6% of Americans (Pew Research Center, 2021). Posting tweets is different from responding to politicians' tweets,

but other evidence suggests that only a minority of Twitter users follow more than a single political elite (i.e., Obama, Trump, Bernie Sanders; Wojcieszak et al., 2022), and even among this group, most use Twitter passively (Research Center, 2021), without interacting with the tweets they see. Thus, only a small percentage of Twitter users respond to politicians' tweets.

The minority of people who are politically active on Twitter tend to be more extreme, affectively polarized and hostile (Bor & Petersen, 2022; Kumar et al., 2023; Mukerjee et al., 2022; Pew Research Center, 2019b, 2021, 2022a). For this reason, the feedback on political leaders' tweets may reflect preferences, but the unusual preferences of a small vocal minority (e.g., Joseph et al., 2021) who, as extremists, approve less of positive interparty relations (Heltzel & Laurin, 2021; Pew Research Center, 2022a). If so, Twitter's Likes and Retweets might reward politicians for dismissing, even if most people disapprove of it. Chapter 3 reports four studies testing these ideas.

1.4 How accurately do people perceive engaging's reputational benefits?

Chapter 3 implies that popular domains of political discourse (e.g., social media) might misrepresent peoples' attitudes; Chapter 4 considers a potential implication. Most people like and respect allies who engage constructively with opposing views, but do they realize that they are not alone in feeling this way—that this is what their allies prefer as well? The answer to this question could be practically important, as actors deciding whether to engage with or dismiss opposing views are likely to consider how their actions will bode in the eyes of allied observers (e.g., Ajzen, 1991). I theorized that people may systematically misperceive their allies' preferences, such that they *underestimate* the reputational benefits they could reap from constructively engaging with opposing views.

Why would people underestimate their allies' preference for engaging? I have already argued that social media may distort what is most popular, so if people take their cues from there, this could cause them to overestimate how much their allies like dismissing. This same misperception could also arise from people perceiving high degrees of political polarization—that is, from people assuming their allies and opponents are deeply divided and have each become extreme.

Empirically, people *do* perceive a lot of polarization—even more than exists. The number of news stories about polarization have increased from 2000 to 2015 (Levendusky & Malhotra, 2016a), giving Americans the impression that their compatriots are deeply divided, united only in their desire to silence opponents (Dias et al., 2022; Han & Yzer, 2020; Yang et al., 2016). Indeed, whereas 71% of Americans perceive *very strong* conflicts between Democrats and Republicans, much fewer feel the same way about divisions of class (31%), race (19%), or age (14%); moreover, 78% of Americans think political divides are growing (Pew Research Center, 2020b). Stories about either ideological polarization (i.e., policy disagreement; Abramowitz & Saunders, 2008; Fiorina & Abrams, 2008) or affective polarization (i.e., liking ones' own party and disliking opponents; Iyengar et al., 2019; see also Jost et al., 2022), as well as salient but rare instances of interparty violence (e.g., Lacey & Herszenhorn, 2011) likely contribute to perceptions of vast divides between political groups. But these perceptions are overblown: A growing literature on false polarization highlights how people overestimate the magnitude of divides (for reviews, see Fernbach & Van Boven, 2022; Wilson et al., 2020).

If people think (and overestimate how much) their allies dislike and resort to overt hostility against opponents, they may (wrongly) assume those allies would disavow engaging with those opponents' views. That is, if an actor believes that their ingroup detests the outgroup

and/or engages in violence against them, that actor might expect ingroup observers to rebuke efforts to engage positively with that outgroup's perspectives. This dovetails with theorizing depicted in Figure 1.2, which suggests that engaging garners more backlash when observers perceive greater difference between their views and the target's. If actors intuit this dynamic, then hearing about growing divides will lead them to expect more backlash for engaging.

Similarly, people might overestimate others' support for dismissing because of narratives around "cancel culture"—a widespread social practice in which people are punished for promoting unacceptable views. This is a direct form of dismissing an opposing view in which people try to not only avoid it for themselves but actively try to remove it from public discussion so others will not hear it. But people overestimate the prevalence of cancel culture: Despite being a relatively new term, 62% of Americans are aware of cancel culture (Pew Research Center, 2022), and they overestimate how much their allies want to cancel people with other political views by a factor of 1.3 to 2.2 (Dias et al., 2023). If people overestimate canceling, they may similarly overestimate the prevalence of (and support for) tamer versions of dismissing.

In short, whether due to either the direct reactions people witness on social media, their general sense that their country is polarized, or the salience of phenomena like cancel culture, Chapter 4's first hypothesis is that despite their own preference for engaging, people will generally assume their allies do not share that preference.

1.4.1 Consequences for engaging behavior

If people (incorrectly) expect that dismissing, not engaging, is what will earn them reputational rewards within their social group, they might conform to this misperceived norm, further reinforcing it. This would be a case of pluralistic ignorance (Prentice & Miller, 1996; Van Boven, 2000), in which people privately like one thing, fail to realize that others share this

opinion, and conform to the norm that they (inaccurately) perceive, further reinforcing it among others who witness their behavior (see Ajzen, 1991; Cialdini & Goldstein, 2004).

Conditions are ripe for pluralistic ignorance, which arises from the motivation to seem like a good group member in the eyes of others (Prentice & Miller, 1996). In this case, compared to someone who engages constructively with opponents, someone who dismisses opponents may seem to be a more loyal, committed group member. Nowadays, Americans perceive large divides between parties (Pew Research Center, 2020b), so they might feel they should be loyal to their own. Pluralistic ignorance theorizing suggests these perceptions of how one *should* behave (i.e., injunctive norms; Cialdini & Goldstein, 2004) come from seeing how one's peers *do* behave (i.e., descriptive norms). To this end, people might have noticed that their group members have become more likely to both dismissively overlook opponents' views and vocally dismiss them outright—that allies have taken efforts to physically avoid hearing from opponents (Motyl et al., 2014; Skitka et al., 2005) and are increasingly willing to engage in partisan violence against them (Mernyk et al., 2021)—which would further validate the idea that a good group member dismisses rather than engages. In sum, people might see how often their allies *do* dismiss and assume these allies' dismissive behavior reflects their personal values and beliefs, rather than desires to be a good group member.

This idea—that people might engage with or dismiss opposing views in accordance with what their allies seem to prefer—aligns with claims made by foundational theories of behavior: that people's behavior follows not only from what they personally endorse, but also (and sometimes more strongly) from what they believe their peers will approve of (i.e., Ajzen, 1991). It aligns even more precisely with claims in the intergroup relations literature, that people's willingness to engage positively with outgroups is powerfully shaped by whether they believe

peers will approve of such behavior (Sherif & Sherif, 1953; Vial et al., 2019); in fact, the two correlate almost perfectly (r = .96; Crandall et al., 2002).

This connection between anticipated social approval and behavior exists in the specific intergroup domain of politics. Around 60% of university students report feeling afraid to express openness to alternative political ideas because of concerns around social pushback from peers—even though, as noted above, these concerns are overblown: Two-thirds of these peers reported that if another student said something offensive, they would respond not by criticizing or condemning them, but by "asking questions to better understand" (Zhao & Barbaro, 2023). Thus, I also hypothesize that part of why people tend to dismiss rather than engage is because they underestimate how much their allies appreciate the latter.

1.4.2 Evidence consistent with my hypotheses

No prior studies have shown that partisans misperceive their allies' liking for engaging, or that this drives them to dismiss, but two recent papers are consistent with these ideas. First, when partisans were made to believe political allies were watching, they engaged more with ingroup views and less with opponents' (Moore et al., 2021). This is in line with what one would expect if, as I hypothesize, actors assume allied observers endorse dismissing opposing views, so they act on this assumption. Second, when partisans learned norms suggesting their allies approve of tolerance and open-mindedness, they were more likely to engage with opposing and balanced viewpoints and less likely to shirk congenial views, relative to when norms suggested the opposite (Wojcieszak et al., 2020). This suggests that people do act on the assumptions they make about their allies' preferences. However, neither of those studies speak to what people assume in the absence of an experimental manipulation, so it remains unclear whether people

actually assume their allies do not prefer engaging, and whether that assumption can explain rising rates of dismissing.

With this in mind, Chapter 4 tests whether manipulating perceived polarization will change people's misperceptions and therefore influence their engaging behavior.

Chapter 2: Why and when people prefer engaging over dismissing

Chapter 2 examines the bases and boundaries of people's preference for allies who constructively engage with opposing views over allies who dismiss them. I propose that people see people who engage as having both likable traits that align with democratic values (e.g., tolerance, rationality, compromise), and unlikable traits that violate moral values (e.g., willing to validate or adopt immoral views); I further propose that the former pathway will have a stronger influence on attitudes, thus accounting for people's overall preference for engaging.

Study 1 tests this process model using statistical mediation. Study 2 and an internal meta-analysis test an implication of the model, depicted in Figure 1.2: To the degree that observers perceive a greater gap between their views and the target view, this greater disconnect should strengthen the influence of the pathway whereby people who engage seem to validate immoral and/or irrational views, and thus weaken the preference for engaging. Study 2 operationalizes the gap by manipulating the extremity of the target view; the internal meta-analysis operationalizes it as the extremity of observers' own views.

Together, these studies promise to shed light on the mechanism driving the results seen in my MA thesis and, more broadly, may help reconcile contradictory results in the literature on reactions to interparty relations by documenting that people have mixed feelings about them. Moreover, Study 2 tests a potential boundary condition on a large, robust effect, and the internal meta-analysis contributes to the debate over whether intolerance is greatest on the political right, or at both extremes of the political spectrum (e.g., Brandt et al., 2014; Jost, 2017).

Chapter 2's studies operationalize *engaging* as initial attempts to understand and think more about opposing views (without necessarily finding common ground or cooperating) and

dismissing as taking steps to overlook opposing views, avoiding exposure to and processing them (without necessarily condemning them outright).

2.1 Study 1

Participants reported their political views then read a short vignette describing two actors, both of whom were political allies to the participant (i.e., they both shared the participant's policy stance on an issue). For my manipulation, I described one actor as engaging opposing political views and the other as dismissing them. Participants next reported their perceptions of the actor, rating them on several qualities reflecting each proposed path (i.e., *tolerant, rational,* and *cooperative*; *open to changing their mind, validates illegitimate views*). They then reported their attitudes towards each actor across several measures (positivity of feelings, emotional reactions, desired social proximity).

2.1.1 Method

Study 1 was preregistered; everything below followed the *a priori* plan unless otherwise specified.

2.1.1.1 Participants.

I recruited 251 Americans from Prolific Academic. I excluded 17 who failed an attention check (a block of instruction text asking them to type a specific phrase into a text box) and one who self-reported providing low-quality data, leaving 233 participants (52% female, 1% non-binary, 47% male; age M = 33.15 years). They tended toward liberalism (M = 2.91, SD = 1.60, significantly below the scale midpoint of 4), $t(232) = -10.36$, $p < .001$.

2.1.1.2 Procedure

2.1.1.2.1 Issue selection and ally assignment

Participants read about four hotly debated issues—affirmative action, climate change, immigration, and welfare—and chose, for each issue, which of two stances they preferred. For example, they read that immigration "concerns how the government regulates the movement of individuals from foreign countries into the United States of America, especially those that intend to work and stay in the country." The two stances they could choose from were "support tougher immigration policies" and "oppose tougher immigration policies."

Participants then read a vignette describing Individual A and Individual B, who shared their beliefs about one of the four issues selected at random. For example, a participant supporting tougher immigration policies might read this:

> Both Individual A and Individual B support tougher immigration policies and believe that immigrants can have negative social and economic effects on the country. In addition, both Individual A and Individual B often hear about immigration from the perspective of those who have similar views, such as TV news anchors and authors who argue in favor of tougher immigration policies.

2.1.1.2.2 Actor manipulation

The vignette then described the actors' engaging and dismissing actions:

> However, Individual A also sometimes watches news anchors who oppose tougher immigration policies, and occasionally reads articles by authors with these views. In contrast, Individual B hardly ever watches news anchors who oppose tougher immigration policies, and rarely reads articles by authors with these views.

2.1.1.2.3 Measures

The dependent variable comprised three measures of attitudes toward each actor. The mediators measure included five attributes rated for each actor. These two sets of measures were counterbalanced.

2.1.1.2.3.1 Attitudes

Participants first completed a feeling thermometer for each actor, using a sliding scale from 0 to 100 to respond to the prompt:

> Please rate your feelings towards the individuals you read about in the vignettes using the labels provided above. Ratings above 50 mean you feel favorable and warm toward the person, with 100 being the most

positive response; ratings below 50 mean you don't feel favorable toward the person and that you don't care too much for that person, with 0 being the most negative rating.

Second, again for each actor, participants rated four statements (α = .94) beginning with the stem, "I would be happy to have [Individual A/B] as . . ." and ending with "a friend," "the teacher of my children," "governor of my state," or "President of the United States" (adapted from Skitka et al., 2005). Responses were made on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

Third, again for each actor, participants reported whether they felt two positive emotions ("proud of" and "respect for") and four negative emotions ("angry at," "disgusted at," "look down on," and "ashamed of"; all reverse scored). Responses were made on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*; α = .89).

I standardized then averaged these measures into a dependent variable (α = .85).

2.1.1.2.3.2 Mediators

Participants rated each actor on five attributes. They read the stem, "Please consider Individual [A/B]'s behavior. By behaving this way, Individual [A/B] is . . ." and then, for each attribute, responded on a scale ranging from 1 (*not at all*) to 7 (*very much*). Three of these attributes were the valued traits I thought participants would ascribe to people who engage constructively: "tolerant of dissimilar people," "cooperative," and "rational, logical" (α = .90). The other two reflected concerns with validating or adopting opponents' views: "implying anti-immigration action views could be right" and "open to changing her mind about immigration" (α = .71). Exploratory analyses for each attribute found that each accounted for an independent indirect effect in the expected direction, and all but one achieved full or marginal significance.

Finally, at the end of this and all other survey studies (i.e., all studies but 3 and 4), I asked participants if they provided honest, trustworthy data and assured them they would be compensated regardless of their answer; I excluded participants who said they did not.

2.1.2 Results

2.1.2.1 Overall preferences

First, a model predicted attitudes from actor type (engage = 1, dismiss = 0) and with a random intercept for participant. Participants preferred the actor who engaged constructively with rather than dismissed opposing political perspectives, $b = 0.44$, $SE = 0.067$, 95% CI = [0.30, 0.57], $t(232) = 6.48$, $p < .001$, $d = 0.51$ (see Figure 2.1).

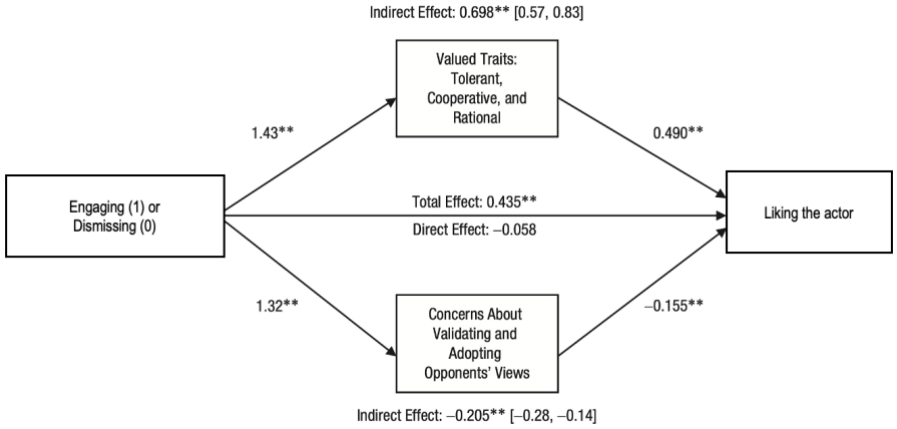


Figure 2.1. Mediation path model, Study 1

2.1.2.2 Process model

Next, I tested a parallel multilevel mediation model to determine whether the two sets of attributes (valued traits and concerns about validating and adopting opponents' views) independently accounted for participants' preference for engaging (*lavaan* Version 0.6-5; Rosseel, 2012). As expected, the two composites accounted for opposing indirect effects (see Figure 2.1). On one hand, there was a relatively large positive effect through participants' view of actors who engage as tolerant, cooperative, and rational, which explains why they preferred

them. On the other hand, there was a relatively small negative indirect effect through participants' concerns that the actor who engaged was validating and could adopt opponents' views, and these concerns suppressed their overall preference. Accounting for both mechanisms accounted for the entire total effect (i.e., no direct effect remained).

2.1.3 Discussion

Study 1 replicated the overall preference for engaging over dismissing. These findings are practically important: Engaging constructively with opponents' views offers reputational benefits; this is something advocates of this approach may wish to advertise.

Study 1's mediation analyses also supported my hypothesized process model: People see those who engage as admirably tolerant, rational, and cooperative, yet at the same time (and with a similar effect size) as alarmingly willing to validate illegitimate views and change their mind. However, the former predicted liking to a stronger degree than the latter, which is why participants preferred engaging overall. Mediation analyses cannot speak to the causal role of any mediator; in that sense, this provides merely correlational information. Study 2 built on this by manipulating a variable hypothesized to contribute to the second set of mediators relating to the violation of moral values.

2.2 Study 2

Study 2 experimentally varied the distance between observers' views and the target view via having actors engage with relatively more (versus less) extreme views. I reasoned that an average observer sees their own views as more different from extreme opposing views, so an actor who engages with those views would seem to violate shared moral principles more, leading the observer to like that actor less; see Figure 1.2.

Participants reported their political views before seeing a vignette about two allied actors, one who engages with opposing views and another who dismisses. The stance that the actors engaged with or dismissively avoided was manipulated to be either relatively less or more extreme. Participants reported their attitudes toward this actor using the same measures as in Study 1.

2.2.1 Method

Study 2 was preregistered; everything below followed the *a priori* plan unless otherwise specified.

2.2.1.1 Participants

I recruited 263 Americans through Amazon's Mechanical Turk. I excluded six who failed an attention check, 26 who failed an English-comprehension check, and six who self-reported providing low-quality data. These exclusions overlapped, leaving 231 participants (50% female, 1% nonbinary; mean age = 34.53 years). They tended toward liberalism ($M = 3.57$, $SD = 1.73$, below the midpoint of 4), $t(230) = -3.77$, $p < .001$.

2.2.1.2 Procedure

Participants first rated their stance on the same four issues as in Study 1. Next, as a within-subjects manipulation, participants read a vignette like that used in Study 1, which described two actors. Both were political allies but one was described as wanting to engage constructively with opposing views whereas the other as wanting to dismissively avoid those views.

I manipulated another factor between subjects: the extremity of the views these actors engaged with or dismissively avoided. Actors engaged with or dismissed either standard views—the same broadly described ones from Study 1—or a specific, relatively more extreme version of

those views: For example, rather than reading about actors who engage with or dismiss the views of pro-choice policy supporters in general, a pro-life participant would read about actors who engage with or dismiss the views of those who support specific, extreme pro-choice policies, such as allowing abortions far later into a pregnancy than is currently legal anywhere in the United States (participants' home country). Table 2.1 presents broad and specific, extreme stances for all four issues.

Table 2.1

Standard and extreme issue stances used in Study 2

Issue	Standard stance	More extreme stance
Abortion	Support pro-choice policies	Support extreme pro-choice policies, like allowing abortions for any reason until the 6th month of pregnancy and for specific health reasons until the 8th month
	Support pro-life policies	Support extreme pro-life policies, like prohibiting abortion completely, even in cases of rape and incest*
Universal healthcare	Support universal healthcare policies	Support extreme universal healthcare policies, like the government fully funding any medical services for U.S. citizens
	Oppose universal healthcare policies	Oppose even basic government funded healthcare programs, and support the extreme policy of leaving individual citizens entirely responsible for paying for their health care needs
Gun regulation	Support increased gun control	Support extreme increases in gun control policies, like the government seizing control of citizen's firearms
	Support gun rights	Support extreme gun rights policies, like removing all existing background checks
Immigration	Support tougher policies	Support extremely tough immigration policies, like creating walls and flight bans that completely halt immigration from several countries
	Oppose tougher policies	Oppose even basic limits on immigration, and support the extreme policy of allowing mass immigration with no restrictions

Note: Each phrase started with "Both Individual A and Individual B" and was followed by one of standard or more extreme stances shown. *This study was run in 2020 before America's 2022 overturn of Roe vs. Wade, when laws described in the pro-life stance would have been unconstitutional.

Participants completed the same three dependent measures ($\alpha = .84$) from Study 1: feeling thermometer, desired social proximity ($\alpha = .96$; 4-item scale), and emotions ($\alpha = .91$). They also completed demographic information and attention checks.

Finally, to check the extremity manipulation, participants rated their agreement with two items, both of which imported the stance (standard vs. extreme) shown in the vignette earlier (stance in [X]): "I question the moral character of people who [X]" and "It's hard to imagine someone having good, valid reasons for [X]" using a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). I averaged responses to index perceived illegitimacy ($\alpha = .89$).

2.2.2 Results

2.2.2.1 Manipulation check

As expected, participants rated the more extreme stances as more illegitimate ($M = 5.29$, $SD = 1.38$) than standard ones ($M = 4.09$, $SD = 1.70$), $t(439) = 8.34$, $p < .001$.

2.2.2.2 Role of viewpoint extremity

A multilevel linear model predicting attitudes from action (engaging = 1, dismissing = 0), view extremity (extreme = 1, standard = 0), their interaction, and random intercepts for participant and vignette issue revealed a significant interaction, $b = -0.40$, $SE = 0.155$, 95% CI = $[-0.71, -0.10]$, $t(456) = -2.62$, $p = .009$. Simple slopes tests revealed that for standard opposing views, participants preferred engaging over dismissing as in Study 1, $b = 0.67$, $SE = 0.100$, 95% CI = [0.48, 0.87], $t(230) = 6.73$, $p < .001$, $d = 0.88$; for extreme opposing views, this preference was still present, but with less than half the size, $b = 0.27$, $SE = 0.118$, 95% CI = [0.04, 0.50], $t(228) = 2.28$, $p = .023$, $d = 0.30$; see Figure 2.2[2].

[2] Throughout this manuscript, I show means above their respective bars and standard deviations in brackets below, and use error bars to show standard error.

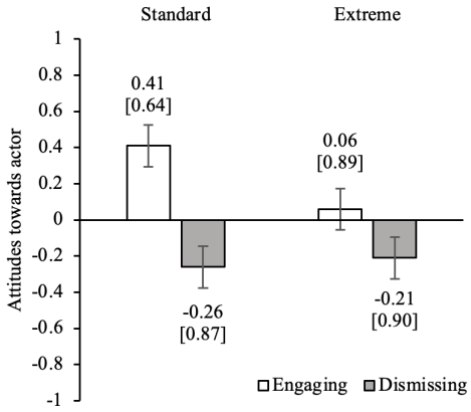


Figure 2.2: Graph of viewpoint extremity by target action, Study 2

Viewed differently, participants disliked dismissing regardless of whether it was directed at standard or extreme views, $b = 0.05$, $SE = 0.116$, 95% CI = [−0.17, 0.28], $t(229) = 0.47$, $p = .641$, $d = 0.06$. They liked engaging less, however, when directed towards more extreme views, $b = -0.35$, $SE = 0.102$, 95% CI = [−0.55, −0.15], $t(229) = 3.44$, $p = .001$, $d = -0.45$. This is consistent with my theorizing: Observers (participants) liked engaging less when directed towards more different—here, extreme—views.

2.2.3 Discussion

In Study 2, people showed a weaker preference for an actor who engaged constructively with (vs. dismissed) opposing views when the view in question was more extreme (relative to a standard, broadly defined view). This offers complimentary evidence for part of my process model: If engaging raises concerns about validating potentially illegitimate views, engaging with *extreme* views would exacerbate this concern and generate backlash.

For people choosing which views to engage with, these findings are practically important: engaging usually has reputational benefits, but not always. To reap these benefits, people should aim to engage with views that are considered less extreme. And yet, even when actors engaged with extreme views, people still preferred them over a similar actor who

dismissed these views. This attests to how robust this preference is and highlights the social costs of dismissing opposing views, extreme or otherwise.

2.3 Internal Meta-analysis: Testing moderation by participant's ideological extremity

Using similar logic about the distance between observers' views and the target view, my process model also predicts *extremist observers* would feel less favorable towards an ally actor who engages with shared opponents' views. Building on the nine-study internal meta-analysis reported in my MA thesis, I conducted an updated internal meta-analysis to test this prediction across 15 studies included in Heltzel and Laurin (2021); one additional study used a binary dependent variable and was therefore incompatible with the continuous dependent variables from these 15. All studies measured ideological orientation, allowing me to test whether people who identify at the extremes of this measure differ in their preference for engaging.

2.3.1 Method

I included all data testing reactions to allies who engage versus dismiss, with a sample of 4231 participants from both within- and between-subjects studies. The dependent variable was typically the attitude composite described in Study 1, but two studies used only a feeling thermometer, and another study used only the thermometer and emotion measures. In all cases I standardized individual measures, averaging them together when there were multiple.

I measured participants' ideological orientation in every study using the same item, but with three different response scales. Most studies used a scale ranging from 1 (*extremely liberal*) to 4 (*moderate*) to 7 (*extremely conservative*), but two used 6-point scales with no *moderate* midpoint, and one study used a 10-point scale ranging from 1 (*left*) to 10 (*right*). To address this inconsistency, I standardized ideological orientation scores within each study. Following prior

work (Toner et al., 2013; Harris et al., 2021), to index extremity, I squared the standardized ideological orientation value.

2.3.2 Results

To test for moderation, I ran a multilevel model predicting attitudes from action (engaging coded 1, dismissing coded 0), standardized ideological orientation, and their interaction, as well as from actor type, extremity (i.e., squared standardized ideological orientation), and their interaction, and including a random intercept nesting participant within study. There was a significant interaction with extremity, b = -0.18, SE = 0.018, $t(7015)$, p < .001 (but not with ideological orientation, b = 0.01, SE = 0.020, p = .718). Simple slopes suggest that the most moderate participants (score of 0) preferred engaging, b = 0.88, p < .001, and so did highly extreme participants (+2 SDs above the average level of extremity), b = 0.31, p < .001.

2.3.3 Discussion

Extremists—operationalized here as identifying more strongly with liberal or conservative ideologies—show a markedly weaker preference for engaging. This finding parallels Study 2's: In both cases, the preference for engaging weakens as the distance between observers' views and the target view grows. Even extremists, though, preferred engaging over dismissing, which again attests to the robustness of this preference. Confidence in this result is boosted by my large meta-analytic sample comprised of studies that used varying methodologies: Results cannot be due to artifacts of individual samples or methods.

This finding extends theories that suggest intolerance is greatest among extremists (Brandt et al., 2014; Crawford & Pilanski, 2014), suggesting that extremists are not just more intolerant of political opponents (Woitzel & Koch, 2022), but also of allies who tolerantly engaged with those opponents' views. In contrast, I find no support for the other side in the

debate over who is most intolerant (e.g., Jost, 2017): Compared to liberals, conservatives were no less tolerant of their allies engaging constructively with opponents' views. That said, allies who engage with opponents are but one of many potential targets of intolerance; my data cannot speak to whether conservatives or extremists are more intolerant of other important targets, nor if they are intolerant in different ways (e.g., more violent), which are issues at the core of this broader debate (see Badaan & Jost, 2020).

2.4 General discussion

After my MA thesis found that people prefer allies who engage with opponents' views over those who dismiss them, Chapter 2's studies documented *why* people typically prefer engaging. Study 1 found that people like individuals who engage because they seemingly possess attributes aligned with democratic values, and that to a lesser degree they dislike them for violating moral values—for legitimizing opponents' views and seeming open to changing their minds. Studies 2 and the internal meta-analysis expand on this latter point, suggesting that the concerns around legitimizing immoral, irrational views get stronger when the distance grows between the participant's and target's views.

Chapter 2 illustrates how useful it can be to use multiple complementary methods to test theorizing around psychological mechanism (Vancouver & Carlson, 2014; see also Jacoby & Sassenberg, 2011): Study 1 used statistical mediation, which in some sense offers the most straightforward test of process, but which cannot fully demonstrate causality: Study 1 did not demonstrate that perceptions of democratic-aligned attributes, or of violated moral values, causally determined participants' liking. Conversely, Study 2's interpretation required some assumptions, but also demonstrated that concerns around legitimizing immoral, irrational views—which I can likely safely assume are higher when actors engage with more extreme

viewpoints—cause people to prefer engaging less. The converging evidence from these two approaches together offers more confidence in the conceptual model than could either alone.

The present results are even more surprising given that many people see their opponents' views as not only immoral but also *irrational*, and thus might find allies who engage with those views to similarly be less rational. Rational values can compel people to dismiss and disregard ill-founded or blatantly false viewpoints. For example, people may find it pointless or offensive to consider the factually incorrect perspectives of flat earthers or holocaust deniers. These fringe beliefs aside, many partisans think their opponents are unintelligent (Iyengar et al., 2019) and corrupted by fake news and misinformation (Traberg & van der Linden, 2022). One might expect these partisans to rebuke allies who engage with opponents' seemingly irrational views, but the present results suggest otherwise: Observers rated those who engaged as *more* rational. Thus, even though prior work suggests that partisans nowadays see opponents as irrational, my results suggest they nonetheless see people who engage with those opponents' views as rational and likable.

2.4.1 Unanswered questions

Chapter 2 provided insight into boundary conditions of people's preference for engaging, but other boundary conditions may exist. For instance, across all my studies examining observers' responses to engaging vs. dismissing, people tended to prefer actors more who engage with opposing views on a specific policy debate (as in Studies 1 and 2), rather than engaging with the views of an opponent, broadly construed (see Heltzel & Laurin, 2021). People have exaggerated, negative ideas of who their political opponents are and what views they hold (Ahler & Sood, 2018). If a person engages with opponents without specifying which views they are trying to better understand, observers might assume the worst: Liberals may interpret allies

engaging with conservative views as sympathizing with white supremacists. Likewise, conservatives may interpret allies engaging with liberal views as sympathizing with flag-burning communists.

Of course, if a person *is* trying to engage with more extreme opponents' views, specifying this probably would not help. Indeed, a recent pre-print (Hussein & Wheeler, 2023) finds opposite effects to those reported in my MA thesis and in Chapter 2 here: People disliked allied actors who engaged with opposing political elites' views (e.g., Donald Trump). Compared to laypersons, political elites are exemplars of their respective sides, so they might either seem or genuinely be more extreme and unlikable (Hetherington, 2001; Kingzette, 2021). As a result, engaging with their views might elicit less liking. Chapter 3 builds on this idea by examining targets who are political elites, and who are typically engaging with or dismissing the views of other elites.

Finally, Chapter 2 used particular operationalizations of engaging and dismissing. The engaging actors made *initial* attempts to connect with opponents, aiming to understand and think more about their views but not necessarily trying to actually cooperate or compromise. And the dismissing actors overlooked opponents' views, without necessarily condemning them outright.

These operationalizations may correspond well to common behaviors by ordinary people. Laypersons have countless opportunities to hear about opponents' views (and healthy democracies rely on them to do so), but even when they find common ground, they rarely have any real opportunity to do something about it by cooperating with opponents on actual policy proposals. Likewise, evidence speaks to the increasing frequency with which people choose to avoiding exposing themselves to and thinking about uncongenial views (e.g., Rodriguez et al., 2017), while other work suggests that going so far as to condemn opponents is relatively

uncommon (e.g., Kumar et al., 2023). Thus, Chapter 2—by studying initial attempts at engaging with versus dismissively overlooking opponents' views—speaks more to how engaging and dismissing would affect laypersons' reputations in common situations.

Nevertheless, Chapter 2 examined relatively mild forms of these behaviors. Compared to someone who takes initial steps to engage with opposing views, someone who engages deeply with those views might seem more cooperative but also more willing to validate opposing views. Likewise, someone who dismissively overlooks opponents' views may seem willfully ignorant and intolerant, whereas someone who dismisses opponents outright might seem to be taking an informed stance and may avoid these negative attributions by justifying their condemnation. That said, people seemed to strongly prefer allies who engage over those who dismiss in Chapter 2, so perhaps the same preference would emerge even with these more intense versions of engaging and dismissing. Because Chapter 3 operationalized engaging and dismissing in these broader ways, I could test whether results replicate the general patterns seen in Chapter 2.

Chapter 3: Preferences for dismissing on Twitter

Chapter 3 aimed to study reputational effects of engaging and dismissing behaviors in a novel context that has become increasingly important for political discourse in recent years: Social media. In particular, Chapter 3 focuses on real-world responses to U.S. Senators' engaging and dismissing tweets.

Recent literature suggests negativity toward political opponents gets more rewards on social media, even though most people explicitly report preferring positive interparty interactions (e.g., see Rathje et al., 2022). I expected to find the same dynamic at play for tweets that engage with vs. dismiss opposing political perspectives, whereby the latter would get more rewards on social media despite most people preferring the former. In Studies 3 and 4, coders identified U.S. Senators' tweets that modeled engaging with or dismissing opponents and their views. Indeed, dismissing tweets received more positive feedback (Likes and Retweets) than engaging ones, so I tested different hypotheses to explain why Twitter rewards that which most people do not prefer. Preregistered Studies S1-S4 (reported in full in the Appendix) found no evidence that these unique preferences owe to the context of social media (e.g., Crockett, 2017), nor my focus on politicians: People did not genuinely prefer dismissing in the context of *politicians*' communications or the medium of *Twitter*. Preregistered Study 5 therefore tested whether Likes and Retweets do not mean what they intuitively seem to (e.g., Frimer et al., 2022), and whether the minority of users responsible for most social media rewards have unusual preferences (e.g., Bor & Petersen, 2022). Finally, Study 6 examined perceptions of engaging and dismissing tweets, further illuminating the disconnect between Twitter and general preferences.

Online popularity is worth attending to in its own right (e.g., Warzel, 2020): Citizens observe and learn political norms from politicians (Zaller, 1992), so politicians' posts, the

feedback they receive, and their coverage on news media (McGregor, 2019) could cause their many followers to replicate, on Twitter or in offline behavior, that which generates rewards online. Nonetheless, it is important to know whether social media feedback is a mirror, reflecting true population preferences, or a funhouse mirror, presenting a warped and misleading picture. Even should Twitter collapse, my findings will bear implications for archival analyses of its data, as well as for new platforms that might replace it as hubs for political discourse.

Chapter 3 operationalizes engaging and dismissing differently from Chapter 2, this time studying deeper attempts to engage and outright dismissal. It is unclear whether Chapter 2's conclusions would apply given Chapter 3's different operationalizations, so I directly test whether they replicate here. In some sense, the case of politicians' tweets is ideal for studying deeper engaging and dismissing outright. Practically, politicians—in their role as lawmakers—are among the only people who can engage deeply by meaningfully enacting compromise with opposing views. Theoretically, their role requires that they advance their party's and constituents' policy goals, yet they often need to compromise to do so, so they could be a potential boundary condition on the preference observed with layperson actors. Likewise, I suggested earlier that people rarely dismiss opposing views outright, but Twitter is a context where such negativity across political lines is more common; thus, whereas Chapter 2 studied the reputational benefits of engaging and dismissing in a specific context (offline) and with a particular actor (laypersons), Chapter 3 studies these benefits for a different type of actor and in a new context, both of which suit its different operationalizations of engaging and dismissing.

Data, analysis code, and preregistered methods, materials, and analysis plans for Study 2 and Appendix Studies S1-S4, are here. I plan to submit these studies to *Science Advances*.

3.1 Studies 3-4

Studies 3 and 4 hand-coded real tweets by U.S. Senators as either engaging with or dismissing opponents' views, or neither of those, then compared the Likes and Retweets each type of tweet received. Study 3 provided initial results; Study 4 replicated them using a different sample of tweets. Both used multiple analytic strategies to ensure robustness.

3.1.1 Method

3.1.1.1 Sample of tweets: Study 3

On December 4, 2018, I collected the text, timestamp, and Like and Retweet counts from as many U.S. Senators' posts as possible, using the Tweepy API through Python (tweepy.readthedocs.io/; limited to 3240 posts per Senator). This yielded data from 246314 original tweets (excluding retweets but including quote tweets[3], per Frimer et al., 2022). I focused on U.S. politicians' tweets because my work builds on literatures about primarily U.S. polarization and political dynamics, and because Twitter pervades U.S. politics: Americans see Twitter as the go-to platform to discuss politics (Mukejee et al., 2022), perhaps because all U.S. lawmakers use Twitter and post there more than other platforms (Pew Research Center, 2020a, 2020b). Moreover, I focused specifically on Senators because they have more notoriety and political power (versus Congressional representatives) and are numerous enough that I can generalize across many targets (versus Presidents). The Chapter 3 General Discussion considers whether and when my findings might extend to other countries, types of political leaders, and social media platforms.

[3] When people react to a retweeted post, this is typically in reference to the original post itself; indeed, Like and Retweets metrics are assigned to the original tweet (they are not recorded as being unique to the retweet). When people react to a quote tweet, this is typically in reference to the commentary the quote-tweeter added, not to the original post; indeed, Like and Retweet metrics are assigned to the quote tweet itself. Thus, in my dataset, reactions to quote-tweeted posts, but not retweeted posts, count as feedback intended for the engaging or dismissing Senator.

If engaging and dismissing tweets could be coded automatically with supervised machine learning algorithms or word dictionaries, I could have analyzed the entire corpus of a quarter million tweets. But identifying whether a Senator engages with or dismisses an *opponent's* views depends not only on the presence of engaging or dismissing language, but also on the target of that language and that target's relationship to the Senator. That is, proper categorization must account for the Senator's and target's beliefs and the relation between them, and for changes in the meaning of terms (e.g., "@POTUS" in this corpus could refer to Obama or Trump [or, in Study 4, Biden] depending on the tweet's timestamp). I therefore used human coders, accepting the cost of a smaller sample for the greater precision and accuracy of coding, compared to available automated methods.

Accordingly, I aimed to select a subset of tweets that was small enough to be hand-coded by my RA team, but that had a relatively high likelihood of reflecting engaging or dismissing to maximize the sample of relevant tweets. As a first step, I screened for tweets that contained at least one of the following character strings: *perspective, empath, view, understand, angle, mindset, belief, learn, consider, think about, listen, reach out, discuss, meet with, met with, hear, debate, other side, avoid, ignore, stance,* and *opponent*.[4] On their face, more of the words captured by these strings relate to engaging than dismissing. But preliminary examination had indicated that many dismissing messages used these same words with a negative modifier (e.g., *refuse* to listen; *cannot* understand; should *not* empathize). Nonetheless, the list of strings was experimenter-made and non-exhaustive, so the set of 32333 tweets that remained after I screened for this list certainly did not contain every single instance of engaging or dismissing in the

[4] The character string *view* often returned tweets promoting Senators' interviews, which were seldom relevant to engaging or dismissing. The string *hear* returned tweets about heart (heartwarming; having heart) and nominee hearings, which were also seldom relevant. Hence, I excluded tweets with the strings *interview*, *heart* and *hearing*.

overall corpus. I have no reason to believe this introduced bias into the *kinds* of engaging and dismissing tweets my coders identified, but I do not recommend any interpretations based on the *frequency* with tweets in each category emerged.

I used a random number generator to select 5000 of these screened tweets for hand-coding. This *N* came from the first author's preliminary coding of 100 screened tweets, which identified five relevant to engaging or dismissing. From this, I estimated that 5000 tweets would generate about 250 tweets for key analyses; assuming similar numbers of engaging and dismissing tweets, this would provide over 95% power to detect medium-size effects (d = .50).

3.1.1.2 *Sample of tweets: Study 4*

On May 20, 2021, I used the rtweet package (Kearney, 2019) to collect 244123 tweets by all sitting U.S. Senators, again excluding retweets but including quote tweets. Compared to Study 3, this initial corpus included tweets from the intervening 2.5 years and reflected changes in sitting Senators from 2018 to 2021.

I made three changes to how I selected tweets to code. First, since Study 3 had identified relatively few dismissing tweets (see below), I added more character strings that were, without a negative modifier, relevant to dismissing. The final list of strings was: *listen, common ground, more in common, view, belie*, both sides, other side, perspective, understand, ignore, disregard, avoid, agree, wrong, lie_, hear_, consider, empath, discussion with, up for debate, be debated, has no place, cancel*, and *shut down* (* allows for alternative endings; _ represents a space).

Second, I automated an additional step in the coding process: To be coded as engaging with or dismissing *opponents*, a tweet necessarily had to refer to an opponent of its author. Study 3's coders identified these references manually as described below; in Study 4, I used specific terms to screen for tweets that contained character strings referring to the Senator's ideological

or partisan opponents, broadly construed (for Democratic Senators: *republican*, *GOP*, or *conservative*; for Republican Senators: *democrat*, *liberal*, *libs*, and *dems*; for both party's Senators, *opponents*). These character strings did not include more specific opponents (e.g., individual opposing Senators, ideologically opposed lobby groups), but in using this method I discarded all tweets that did not reference an opponent at all, giving retained tweets a higher chance of reflecting either engaging or dismissing.

Third, because after this two-step screening procedure only 1135 tweets remained, I did not need to randomly select a fraction of them; the RA team simply coded the entire subset.

These three changes, along with the different timespan of tweets I collected, virtually eliminated overlap between the two studies' samples. Nonetheless, Study 4's final sample contained 20 tweets that were also present in Study 3's sample. To maximize power, the analyses presented here preserve these overlapping tweets for both studies; the Appendix reports virtually identical results for both studies when excluding these tweets.

3.1.1.3 Coding procedure: Study 3

I trained seven research assistants to code each tweet as engaging or dismissing, or neither of those. Coders were Canadian undergraduate students with exposure to U.S. politics but no partisan ties nor strong feelings about individual Senators. They had access to a datafile listing each tweet's text, timestamp, and the authoring Senator's name and party affiliation; they gleaned additional context online when necessary. They were blind to Like and Retweet counts and information about research questions or hypotheses.

As noted above, coders first checked if the tweet referred to an opponent (i.e., a person or group the Senator disagreed with). If it did, they coded it as engaging if its author described good-faith or bipartisan contact with opponents that they had had, wanted or intended to have, or

believed should happen. They coded opponent-referring tweets as dismissing if its author rejected an opponent or their views outright, or implied that others should do the same (e.g., condemning the opponents or their views as factually or morally wrong). They coded all other tweets as neither. See Figure 3.1 for examples from both Study 3 and 4.

Figure 3.1. Examples of tweets coded as engaging (top) or dismissing (middle) opponents and their views, or as neither of these (bottom). Tweets on the left (right) appear in Study 3's (4's) sample.

3.1.1.4 Coding procedure: Study 4

Study 4 used stricter criteria to identify relevant tweets. Some tweets coded as engaging in Study 3 technically fit the definition even though their central goal was something else. For example, some tweets technically called for bipartisanship while their central goal appeared to be a favorable comparison between the Senator or their party and opponents (see Figure 3.2, left);

others whose central goal was to announce a public appearance could convey a desire to hear from citizens in attendance, who presumably might include opponents (e.g., see Figure 3.2, right). In Study 4, coders searched specifically for tweets whose central message was engaging or dismissing, like those in Figure 3.1.

Figure 3.2. Study 3 tweets that only loosely or secondarily reflected engaging

Research assistants worked until two had independently coded each tweet. For Study 3 (4) they agreed 76% (81%) of the time, Gwet's AC_1 = .72 (.78), indicating substantial agreement (Gwet, 2008); I resolved disagreements. Table 3.1 lists the number of tweets in each category, overall and by Democratic and Republican Senators. All Study 4 tweets referred to opponents, but most were still coded as neither. In these tweets Senators might, for example, encourage opponents to engage with, or accuse them of dismissing, their own party's views; see Figure 3.2. Sensitivity analyses for Study 3 (4) show that the sample size for key contrasts (engaging vs. dismissing tweets) provided 80% power to detect effects of at least d = .38 (.47).

Table 3.1

Counts and characteristics for tweets in each category

	Category								
	Engaging			Dismissing			Neither		
Study	Total	Democrat	Republican	Total	Democrat	Republican	Total	Democrat	Republican
3	198	108	90	76	65	11	4726	2339	2387
4	53	28	25	118	65	53	964	485	479

To validate the coding, an independent sample of 536 participants (age M = 41.3; 49% women, 49% men, 2% nonbinary, agender, or genderqueer) rated the tweets coded as engaging or dismissing in Study 4 (see Appendix). They used a bipolar scale anchored by the terms "Mostly positive engagement with opponents' views" (-2) to "Mostly dismissive of opponents' views" (+2). This naïve sample placed 100% of tweets in the same category as coders did—all means for dismissing tweets were above 0; all means for engaging were below 0. Thus, laypeople perceived the same clear distinction coders did between the two categories of tweets.

3.1.1.5 Measures: Studies 3 and 4

The dependent measures were Like and Retweet counts. Both were highly skewed so I log transformed them (after adding .01 to all observations, since values of 0 cannot be log transformed); the Appendix reports similar results from analyses of raw counts. I analyzed data with and without covariates that prior work (and the present data) generally finds to be correlated with Likes and Retweets: The Senator's party affiliation (I treated Independents Bernie Sanders and Angus King Jr. as Democrats, with whom they caucus), ideology (based on their voting record; see Lewis et al., 2021), and publicly listed gender and race, the tweet's length in characters, whether it included media (e.g., a URL, picture, or video), and the year it was posted.

In Chapter 1, I noted that Like and Retweet counts may be confounded with exposure (see Frimer et al., 2022). Study 5 looks more closely at whether they nonetheless align with positive attitudes, but in the meantime, Studies 3 and 4 account for exposure by controlling for Senators' follower counts. Others have argued that Retweets are an important source of exposure, and that therefore researchers should analyze the number of Likes adjusted for the number of Retweets (Frimer et al., 2022). But this strategy may not be appropriate for tweets by U.S. Senators, who have far more followers (e.g., a median near 200000 in Study 4) than the

typical person who might Retweet their posts (average Twitter users have 100 followers or fewer, per Pew Research Center, 2021). For Senators, then, followers will account for the lion's share of the exposure most of their posts get; each Retweet by a typical user will increase this exposure by only .05% or less. In other words, Senators' follower counts are likely a better proxy for their tweets' exposure than those tweets' Retweet counts; analyses therefore account for exposure by controlling for the former.

3.1.2 Results

I conducted the same analyses and obtained the same key results for both datasets. For each study's raw Like and Retweet counts for tweets in each category (values rounded to the nearest whole number), see Figure 3.3.

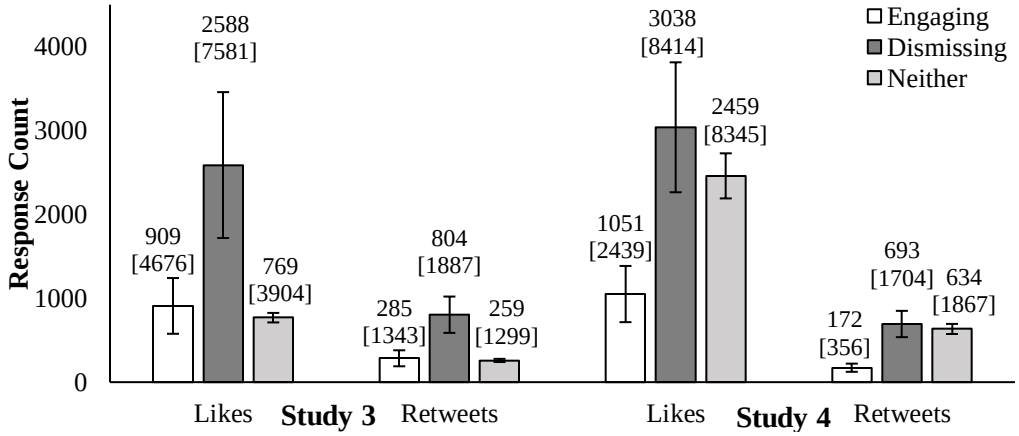


Figure 3.3. Raw like and retweet counts for Senators' tweets modeling engaging, dismissing, or neither

3.1.2.1 Which received more positive feedback, engaging or dismissing tweets?

My first set of analyses set aside *neither* tweets, and compared positive feedback received by Senators' engaging versus dismissing tweets. Multilevel models predicted log-transformed Likes and Retweets from tweet category (dismissing = 0; engaging = 1), and a random intercept for Senator; see Table 3.2. In all models, with medium to large effect sizes, Senators' dismissing tweets received more positive feedback than their engaging tweets.

Table 3.2

Results of comparisons between engaging, dismissing tweets in Studies 3 and 4

Study	DV	Covariates	*b (SE)*	95% CI	*t*	*p*	*d*
3	Likes	No	-1.04 (0.23)	-1.50, -0.59	-4.53	< .001	0.80
		Yes	-0.86 (0.19)	-1.24, -0.47	-4.42	< .001	0.75
	Retweets	No	-1.23 (0.21)	-1.63, -0.82	-5.94	< .001	1.04
		Yes	-1.12 (0.19)	-1.50, -0.75	-5.87	< .001	0.99
4	Likes	No	-0.59 (0.25)	-1.08, -0.09	-2.32	.022	0.55
		Yes	-0.59 (0.27)	-1.12, -0.06	-2.19	.030	0.55
	Retweets	No	-0.90 (0.25)	-1.38, -0.41	-3.66	< .001	0.85
		Yes	-0.95 (0.26)	-1.47, -0.43	-3.62	< .001	0.90

Note. Likes and Retweets are log-transformed. With (without) covariates, the models for Study 3 had 261 (270) *df*; those for Study 4 had 129 (167).

3.1.2.2 How do these compare to Senators' other tweets?

A second set of models included all tweets, and dummy coded the category variable with *neither* as the reference category; see Table 3.3. Reactions to *neither* tweets differed across studies. In Study 3, these tweets received relatively less positive feedback (similar to engaging tweets), but in Study 4, they received relatively more (similar to dismissing tweets). I attribute this difference to the tweet selection procedure: In Study 4 but not 3, all tweets mentioned Senators' opponents, which tends to generate more likes and retweets (Rathje et al., 2021). Tweets that mention opponents while engaging with their views are an exception to this rule.

Table 3.3

Comparing engagement with Senators' neither vs. engaging and neither vs. dismissing tweets

____ vs. neither	DV	Covariates	Study	*b (SE)*	95% CI	*t*	*p*	*d*
Engaging	Likes	No	3	0.11 (0.10)	-0.09, 0.31	1.09	.278	0.08
			4	-0.67 (0.20)	-1.07, -0.27	-3.30	< .001	-0.50
		Yes	3	0.02 (0.09)	-0.15, 0.19	0.21	.837	0.02
			4	-0.72 (0.21)	-1.14, -0.31	-3.41	< .001	-0.55
	Retweets	No	3	0.01 (0.09)	-0.17, 0.19	0.13	.894	0.01
			4	-1.01 (0.20)	-1.40, -0.62	-5.09	< .001	-0.76
		Yes	3	-0.06 (0.08)	-0.23, 0.10	-0.73	.467	-0.05
			4	-1.08 (0.21)	-1.50, -0.66	-5.10	< .001	-0.82
Dismissing	Likes	No	3	0.89 (0.17)	0.57, 1.22	5.40	< .001	0.64
			4	0.15 (0.14)	-0.12, 0.43	1.08	.279	0.11
		Yes	3	0.64 (0.14)	0.37, 0.92	4.61	< .001	0.54
			4	0.14 (0.15)	-0.15, 0.43	0.95	.344	0.11
	Retweets	No	3	0.96 (0.15)	0.68, 1.25	6.66	< .001	0.79
			4	0.14 (0.14)	-0.13, 0.41	1.02	.307	0.11
		Yes	3	0.77 (0.13)	0.50, 1.03	5.71	< .001	0.67
			4	0.12 (0.15)	-0.17, 0.41	0.84	.402	0.09

Note. Likes and Retweets are log-transformed. With (and without) covariates, the models for Study 3 had 4976 (4995) *df*; those for Study 4 had 994 (1130).

3.1.3 Discussion

Across two independent samples, U.S. Senators' tweets received more Likes and Retweets when they dismissed, rather than engaged constructively with, opposing views. That is, Senators received more positive reinforcement for dismissing opponents and their views than for cross-party rapprochement and dialogue. Exploratory analyses (see Appendix) found that Senators' dismissing tweets received more positive feedback even after accounting for linguistic markers that prior research has shown attract this feedback (e.g., uncivil, negative, or outraged language).

The popularity of dismissing tweets presents a puzzle: Prior work (along with Chapter 2) suggest people generally dislike allies who dismissively avoid opponents (Heltzel & Laurin, 2021), or who otherwise perpetuate interparty negativity (e.g., Frimer et al., 2018), albeit they did so using different operationalizations of engaging and dismissing than Chapter 3. Why does Twitter provide positive feedback to Senators' tweets that dismiss outright opposing views? This pattern is particularly surprising since Senators have at least some outpartisan followers, who would obviously prefer Senators' engaging tweets. If anything, this would have suppressed the effect I found. For it to have emerged so robustly, Senators' allies on Twitter must have Liked and Retweeted dismissing tweets to an even more striking degree.

3.2 Studies S1-S4: A summary

One explanation for this puzzling discrepancy (and for some other instances of the broader trend whereby online feedback clashes with popular opinion) is that, whereas most people prefer ally *citizens* who engage with rather than dismiss *offline*, their genuine preference flips when it comes to *politicians*, or in the context of *Twitter*. Preregistered Studies S1 and S2 cast doubt on this possibility: Participants responded to virtually identical statements either by

Senators or ordinary citizens, on Twitter or offline; in neither study did they prefer dismissing by politicians on Twitter. In fact, in most in analyses, Senators and tweets elicited a stronger relative preference for *engaging* (see Table 3.4, and appendix for details). These two studies thus provide no evidence that people genuinely prefer dismissing by Senators on Twitter.

Table 3.4

Summary of sample, method, and results, Studies S1-S4

	N[a]	Population	Stimulus source	Measure	Overall preference	Moderation
S1	659 / 657 (1027)	Prolific Americans, even *n* of D / R	Experimenter generated	Feelings of warmth toward person who engaged / dismissed	Engaging	Greater preference for engaging by politicians (vs. citizens), on Twitter (vs. offline; S1 only)
S2	200 / 199 (468)		Six tweets from 4 (minor tweaks)		Null	
S3	200 / 198 (1967)	Prolific Americans, even *n* of D / R / I	All tweets from 4 (unmodified)		Null	Greater preference for dismissing among those who react frequently to politicians' tweets (vs. everyone else) and extremists (vs. moderates)
S4	200 / 200 (1998)				Null	

[a] This column reports the number of participants before / after exclusions (described in the Appendix), and in brackets the number of observations analyzed.

Participants in Studies S3 and S4 responded to the entire set of Study 4 tweets, and again showed no preference for dismissing.[5] Together, these four studies challenge the idea that most people genuinely prefer when their politicians tweet about dismissing rather than constructively engaging with opponents. Study 5 therefore tested two reasons why Twitter feedback might nonetheless reward Senators' dismissing tweets.

3.3 Study 5

First, the bulk of Likes and Retweets might come from a vocal minority with unusual preferences. They may be more extreme, hostile, and opposed to positive interparty interactions compared to the larger majority (Bor & Petersen, 2022; Joseph et al., 2021; Pew Research

[5] Findings from all four studies contrast with the popularity of dismissing tweets using Twitter metrics, but three of the four also contrast with prior work showing a general preference for interparty relations (Frimer & Skitka, 2018, 2020; Heltzel & Laurin, 2021). The Appendix reports additional data and analyses that explain why Study S1 alone replicated this prior work: Compared to Study S1's experimenter-generated tweets, Studies S2-S4's real dismissing tweets made their authors come across as less intolerant, uncooperative and irrational—that is, more appealing.

Center, 2020b, 2022a), and might thus have different preferences. Indeed, exploratory analyses in Studies S3-S4 found that only a minority (7%) of crowdsourced participants reported frequently reacting to politicians' tweets, and that this minority *did* prefer Senators' dismissing tweets to engaging ones; see Appendix. To confirm and extend this pattern, Study 5 solicited responses to Study 4's tweets from a large sample of frequent reactors, alongside other crowdsourced participants who, despite their demographic peculiarities (they are more likely to be younger, liberal- and democrat-leaning, and women; Berinsky et al., 2012; Prolific Team, 2022), typically respond similarly to nationally representative samples on political measures (Mullinix et al., 2015).

Second, Likes and Retweets may not reflect preferences at all (e.g., Frimer et al., 2022). To test this idea, each of Study 5's participants reacted to tweets using one of three measures: Feelings toward the Senator who posted the tweet (one of the measures used in Study 1 and 2), approval of the tweet itself, and intentions to Like and Retweet it. This feature of Study 5 also allowed for the possibility that Appendix Studies S1-S4's results did not present the full picture: That people genuinely do approve most of dismissing *tweets*, but that this does not translate into feelings toward tweets' *authors*, as measured in those studies.

3.3.1 Method

Study 5 was preregistered; everything below followed the *a priori* plan unless otherwise specified.

3.3.1.1 Participants

I recruited Americans on Prolific Academic, screening for equal thirds self-identified Democrats, Republicans, and Independents / other. My initial preregistration anticipated 160 participants to represent the majority who rarely or never reacted to politicians' tweets, and 80 of

the more unusual participants who reported frequently doing so. Analyses of this initial sample supported my preregistered hypotheses (see Appendix) but were critically confounded: The smaller subsample of 80 frequent reactors had not rated every tweet on each of the three measures, so analyses comparing subsamples and measures were also comparing responses to slightly different sets of tweets. Thus, I posted an amended preregistration specifying I would continue recruiting participants until frequent reactors had provided at least two ratings per measure per tweet. The final sample included 323 participants (156 frequent reactors; age M = 42.7, 52% men, 46% women, 1% non-binary, 3 missing data) who provided 4571 observations (this excludes two additional participants who self-reported providing low-quality data).

3.3.1.2 Procedure

The survey automatically ejected anyone who failed an English comprehension check (as in Study 2). Participants then reported demographics, including party affiliation: "Which of the following best describes your political party affiliation?" (1 = Strongly Democrat, 2 = Somewhat Democrat, 3 = Lean Democrat, 4 = Lean Republican, 5 = Somewhat Republican, 6 = Strongly Republican; the sample leaned to the Democratic side of the scale midpoint of 3.5; M = 3.25, SD = 1.87, $t(322) = -2.42$, $p = .016$). I assigned participants who chose 1, 2 or 3 (4, 5 or 6) to see tweets by Democratic (Republican) Senators.

Participants also reported whether they had "ever liked or retweeted something by a U.S. Senator or house representative." Frequent reactors were those who selected *yes, frequently* (n = 156); I screened 2128 participants to obtain this subsample. This 7% hit rate is similar to what I had observed in Studies S3 and S4, and similar to the figure I noted above for the percentage of the overall population responsible for the bulk of political activity on Twitter (Pew Research Center, 2019b). As preregistered, I recruited participants who selected any other response (*yes,*

once or twice, $n = 50$; or *no, never*, $n = 117$) into the study until I reached the target ($n = 160$) for that group, then diverted any further such participants to other surveys.

The survey presented 10 tweets in random order, randomly selected from the engaging and dismissing tweets authored by Senators from the participant's own party, and present in Study 4's set of 171 (Study 4's tweets rather than 3's because of the former's stricter coding scheme). As per the preregistration amendment, frequent reactors recruited following that amendment saw tweets drawn from the set that, due to chance, had not yet received two ratings by frequent reactors; they also saw 16 tweets instead of 10 so I could achieve the target number of ratings before exhausting the pool of available participants. All participants saw half engaging and half dismissing tweets.

I measured preferences using three different measures, varying between participants which measure they completed (rather than having the same participants complete all three measures, which could have artificially inflated the concordance between measures). One random third of participants reported their feelings of warmth toward the Senator who authored each tweet using the same feeling thermometer item as in Study 1 and 2 (this was the measure that produced null results in Studies S2-S4).

A second third of participants reported their approval of the tweet itself ("To what extent do you approve of this tweet's message?"; 1 = Strongly disapprove; 5 = Strongly approve).

A final third reported their intent to provide positive feedback on Twitter ("The following questions refer to Twitter's 'like' and 'retweet' features: Would you have given this tweet a "like" at that time it was posted? [1 = yes, 0 = no] Would you have retweeted this tweet at that time it was posted [1 = yes, 0 = no]). I averaged these items into a single intention measure (they correlated at $r = .61$ and produced similar patterns when analyzed separately).

To make scores comparable across the three response measures' different scales, I standardized all of them. Study 5 also measured various individual differences that I thought might distinguish frequent reactors from everyone else. The preregistration highlighted one measure, partisan extremity (quadratic distance, as described in Chapter 2; the Appendix reports identical results when indexing extremity differently). I also measured affective polarization (warmth towards inparty [Democrats or Republicans] minus warmth toward outparty, measured with the same 0-100 scale described above; Iyengar et al., 2019), desire for group status ("I want [inparty] to have more power and status than [outparty]"; 1 = Strongly disagree to 7 = Strongly agree), and endorsement of compromise ("[Inparty] politicians should be willing to compromise with [outparty] politicians"; 1 = Strongly disagree to 7 = Strongly agree). These variables are all linked with moralized political attitudes (Garrett & Bankert, 2020; Finkel et al., 2020; Ryan, 2017; Ward & Tavits, 2019); I reasoned they might help explain why frequent reactors might genuinely prefer their politicians to dismiss, rather than engage with, opponents' views.

3.3.2 Results

I first analyzed data across all participants, testing relative preferences across all response measures. I then followed up with models that included interactions between response measure and all key variables, asking whether, for example, people's relative preferences differed if they reported their intentions to Like or Retweet compared to their approval of tweets.

3.3.2.1 Do frequent reactors respond differently than everyone else to engaging versus dismissing?

A multilevel model predicted responses from tweet type (dismissing = 0, engaging = 1), participant group (frequent reactor = 0, everyone else = 1), their interaction, and random intercepts for tweet and participant; see Figure 3.4. The key participant group × tweet type

interaction was significant (b = 0.24 (0.05), t(4552) = 5.24, p < .001): Frequent reactors responded more positively to dismissing tweets than engaging (b = -0.24 (0.03), t(4552) = 7.12, p < .001), replicating the feedback these tweets actually received on Twitter. Everyone else showed no preference (b = -0.00 (0.04), t(4552) = -0.07, p = .946), replicating Studies S2-S4. In other words, Twitter rewards what *some* people prefer, but not necessarily what *most* people prefer, presumably because Likes and Retweets specifically come from a small subgroup of people with unusual preferences.

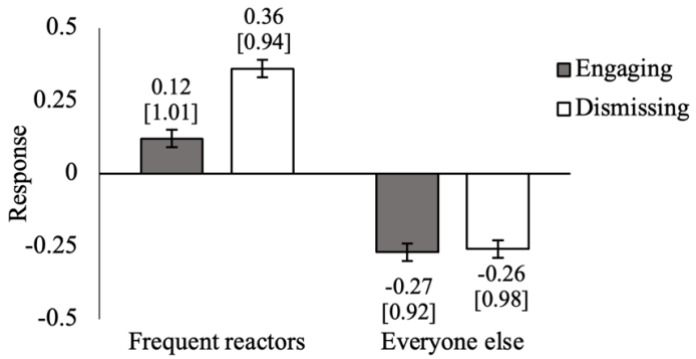


Figure 3.4. Participants' responses to engaging and dismissing as a function of reaction frequency.

3.3.2.2 How do (intended) Likes and Retweets correspond to more traditional preference measures?

The analyses above collapsed across three measures, masking potential differences in how people respond when asked directly for their feelings toward a tweet's author, about their judgment of the tweet itself, or about their behavioral intentions. To unmask these differences, I added a variable for response type (dummy coded with approval of tweet as the reference group) and its interactions to the model described above; see Table 3.5 and Figure 3.5. Significant three-way interactions suggested that the key finding—the participant group × tweet type interaction—differed depending on how people reported their preferences.

Table 3.5

Full model including moderation by measure, Study 5

Effect	*b (SE)*	95% CI	*t*	*p*

Intercept	0.24 (0.09)	0.07, 0.41	2.78	.006
Tweet type	-0.27 (0.06)	-0.39, -0.15	-4.34	< .001
Participant group	-0.49 (0.12)	-0.73, -0.26	-4.09	< .001
Feeling toward Senator (vs. approval of tweet)	0.04 (0.12)	-0.21, 0.28	0.29	.770
Intent to Like and Retweet (vs. approval of tweet)	0.19 (0.12)	-0.04, 0.42	1.62	.105
Tweet type × participant group	-0.55 (0.09)	0.38, 0.73	-6.24	< .001
Tweet type × feeling toward Senator (vs. approval of tweet)	0.005 (0.09)	-0.17, 0.18	0.05	.958
Tweet type × intent to Like and Retweet (vs. approval of tweet)	0.06 (0.07)	-0.09, 0.20	0.75	.455
Participant group × feeling toward Senator (vs. approval of tweet)	0.11 (0.17)	-0.23, 0.44	0.61	.543
Participant group × intent to Like and Retweet (vs. approval of tweet)	-0.28 (0.17)	-0.61, 0.04	-1.71	.087
3-way interaction (feeling vs. approval)	-0.37 (0.13)	-0.62, -0.12	-2.88	.004
3-way interaction (intent vs. approval)	-0.45 (0.11)	-0.67, -0.24	-4.15	< .001

Note. The model had 4544 *df*. Positive simple slope coefficients reflect a preference for engaging in that condition.

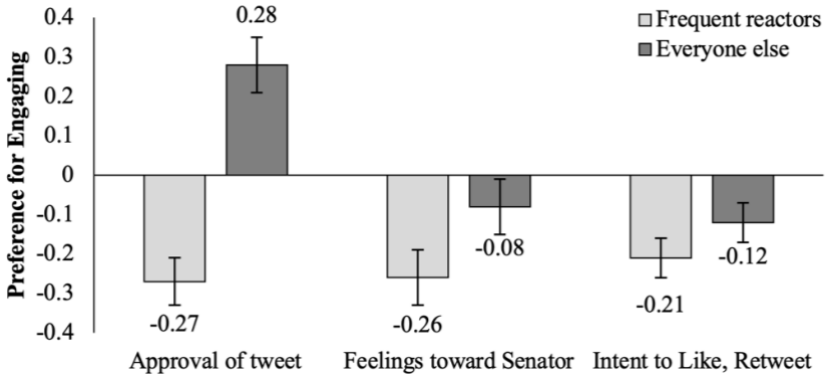


Figure 3.5. Participants' preference for engaging over dismissing as a function of reaction frequency, measure

To unpack this, Table 3.6 reports the simple slopes representing participants' relative preferences for engaging versus dismissing, separately for each preference measure and each participant group. Table 3.6 also shows the two sets of simple two-way interactions that compare these relative preferences. The final column shows participant group × tweet type interactions testing whether frequent reactors show different preferences than everyone else, separately for each measure. The bottom three rows show response type × tweet type interactions testing whether the different measures elicit different preferences, separately for each participant group.

Table 3.6

Comparing preferences for engaging (1) over dismissing (0), Study 5

	Frequent reactors	Everyone else	Frequent reactors vs. everyone else
Approval of tweet	*b* = -0.27, *t* = -4.34***	*b* = 0.28, *t* = 4.30***	*b* = -0.55, *t* = -6.24***
Feeling toward Senator	*b* = -0.26, *t* = -4.02***	*b* = -0.08, *t* = -1.20	*b* = -0.19, *t* = -2.02*
Intention to Like / Retweet	*b* = -0.21, *t* = -4.69***	*b* = -0.12, *t* = -2.35*	*b* = -0.10, *t* = -1.52
Approval vs. feelings	*b* = .004, *t* = .05	*b* = .36, *t* = 3.96***	
Approval vs. intentions	*b* = .06, *t* = .75	*b* = .40, *t* = 4.98***	
Feelings vs. intentions	*b* = .05, *t* = .66	*b* = .04, *t* = 0.45	

Note. The model had 4544 *df*. Positive simple slope coefficients reflect a preference for engaging in that condition.

These analyses support several conclusions. First, frequent reactors showed the same consistent preference across all measures: They felt more warmly toward Senators who posted dismissing tweets, they approved more of those dismissing tweets, and they intended to Like and Retweet them more. This replicates the actual feedback the Tweets received, indicating that Likes and Retweets *do* index preferences—at least those of frequent reactors.

Second, everyone else—the infrequent reactors—used these measures differently. When asked directly about their personal feelings or their approval, they report different preferences than do frequent reactors. Similar to Studies S2-S4's samples, they reported equally warm feelings toward Senators who posted engaging versus dismissing tweets; similar to prior work documenting preferences for positive interparty contact, they approved more of engaging versus dismissing tweets (this echoes Chapter 2's findings, despite its studies using different, broader conceptualizations). These two findings further confirm that frequent reactors' preferences differ from everyone else's.

Third, the two participant groups did *not* differ in what they said they would more likely Like or Retweet: Both groups were more inclined to reward dismissing tweets with positive public feedback, meaning that the group representing the majority were more willing to Like and Retweet dismissing tweets despite being *less* likely to approve of these tweets. My preregistration had not anticipated this, as it focused on the prediction around frequent versus infrequent reactors, but this may be an additional reason why Twitter feedback does not reflect the mostly silent majority's preferences: Even when this group of people *does* provide feedback on Twitter, they may not do so in line with what they actually like.

3.3.2.3 What accounts for frequent reactors' unique preferences?

I next tested whether extremists' preferences matched frequent reactors': A multilevel model predicted responses from party affiliation (midpoint-centered), party extremity (squared value of participant's midpoint-centered partisan affiliation), and their interactions with tweet type (dismissing = 0, engaging = 1). The key party extremity × tweet type interaction was significant, -0.07 (0.01), $t(4550) = -7.36$, $p < .001$; see Figure 3.6 (gray band around lines indicates standard error). The most extreme participants—like frequent reactors—preferred Senators' dismissing tweets over their engaging ones ($b = -0.30$ (0.04), $t(3422) = 8.61$, $p < .001$). The most moderate participants instead preferred Senators' engaging tweets, ($b = 0.10$ (0.04), $t(3751) = 2.51$, $p = .012$). Put differently, extremists responded more positively than moderates to engaging tweets ($b = .04$ (0.01), $t(4550) = 2.78$, $p = .005$), but even more so to dismissing tweets ($b = 0.11$ (0.01), $t(4550) = 7.59$, $p < .001$).[6] The same extremity × tweet type interaction emerged consistently in each of the Appendix studies. As well, the Appendix reports preregistered analyses that include interactions with response measure; similar to above, the key party extremity × tweet type interaction was significant for feelings of warmth and for approval, but not for intentions to Like and Retweet.

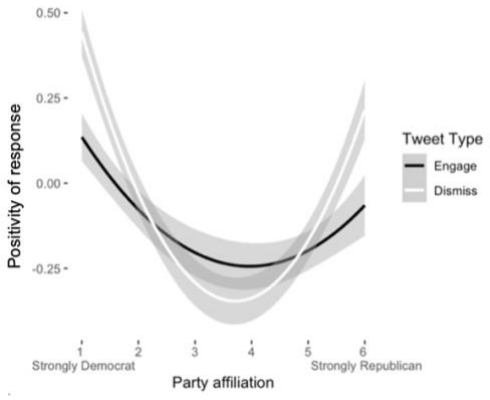


Figure 3.6. Responses to Senators' engaging and dismissing tweets as a function of partisan extremity.

[6] Participants responded only to tweets by inparty Senators, so one might have expected extremists to respond more *negatively* than moderates to engaging tweets. Exploratory analyses (see Appendix) explain why they did not: Extremists are more affectively polarized, which accounted (marginally) for their positivity toward engaging tweets. That is, extremists like their own party more, and so like all tweets authored by inparty Senators. Their preference for dismissing tweets then amplifies this (or, put differently, their aversion to engaging is too weak to eliminate it).

Having observed in these preregistered analyses that extremists and frequent reactors had the same unusual preferences, I ran exploratory analyses (that the preregistration anticipated but in no great detail) testing whether extremism and three other individual difference measures (affective polarization, desire for status, and endorsement of compromise) account for frequent reactors' distinctiveness. Preliminary exploratory analyses informed the mediation model I tested: Predicting participant group from all four individual differences simultaneously (along with midpoint-centered affiliation), extremism had the only significant association (Table 3.6, first numerical column). Predicting instead participant's preference for engaging over dismissing (i.e., the difference score between their average responses to each tweet type), the pattern was reversed: extremity had no significant association whereas the three other measures did (Table 3.6, second numerical column).

Table 3.7

How individual differences relate to frequent reactors' preferences, Study 5

Variable	Link to participant group (df = 317)	Link to preference for engaging (df = 312)	Indirect effect (IV → extremity → ____ → DV)	Indirect effect (IV → ____ → DV)
Extremity	-0.30, z = -5.24***	-0.01, t = -0.78	n/a	0.01, z = 0.36
Affiliation	0.04, z = 0.57	-0.00, t = -0.01	n/a	*n/a*
Compromise	0.09, z = 1.16	0.17, t = 6.76***	0.04, z = 2.84**	0.04, z = 1.30
Desire for status	-0.13, z = -1.55	-0.07, t = -2.53*	0.03, z = 2.31*	0.02, z = 1.25
Affective polarization	0.005, z = 1.01	-0.004, t = -2.58*	0.05, z = 2.59*	-0.004, z = -0.34

Note: All indirect effects in this table are controlling for each other; that is, the third numerical column reports indirect effects through single mediators *after* accounting for the serial mediations.

Along with the theoretical reasoning I described above, these two observations suggested a serial mediation model where the IV was participant group (frequent reactors = 0; everyone else = 1), the DV was within-participant preference for engaging, extremity was the first mediator, the other measures were parallel downstream mediators, and midpoint-centered party affiliation was a covariate. All three serial pathways showed significant indirect effects (Table 3.6, third numerical column), with no indirect effects outside of those serial pathways (Table 3.6,

fourth numerical column). My interpretation of these findings is not causal—in any case, statistical mediation models of cross-sectional data cannot provide evidence of causality. Instead, I take these findings to mean a) that frequent reactors are more politically extreme than everyone else, b) that their political extremity is linked with them opposing compromise, strongly preferring their own side, and wanting their party to gain status, and c) that the shared variance described in b) above is also shared with relative preferences for dismissing.

3.3.3 Discussion

Dismissing tweets' online popularity fails to track most people's self-reported preferences, for two reasons. First, the minority who interacts most with politicians' tweets reports unusual preferences: Unlike everyone else, they genuinely like dismissing tweets, and distribute Likes and Retweets accordingly. Second, the larger majority (intend to) use Likes and Retweets in an unintuitive way: They approve more of engaging tweets but are more liable to Like or Retweet dismissing ones. I had not anticipated this second observation, but it seems like a textbook case of pluralistic ignorance (Prentice & Miller, 1996): The majority's private views differ from those expressed by the loudest voices, so the majority publicly conforms to what it perceives as a widespread norm (e.g., Van Boven, 2000).

Study 5 also helped reconcile the general pattern I observed in the Appendix studies (a null effect) with prior findings suggesting people prefer positive interparty relations: People do prefer engaging tweets themselves, compared to their dismissing counterparts, but this effect on the kind of communication people prefer does not appear to extend to more global feelings toward the communicators. Perhaps it is not strong enough to overcome other associations people have with their party's Senators.

3.4 Study 6

Study 5 also explored how frequent reactors' unusual preferences might relate to their personal characteristics (e.g., extremism). Study 6 explored characteristics of tweets themselves, asking what dismissing (engaging) tweets convey that uniquely attracts (repels) frequent reactors.

3.4.1 Method

This study was not preregistered. Exact materials are available here.

3.4.1.1 Participants

I recruited 323 participants from Prolific Academic such that roughly equal parts reported identifying as part of the Democratic or Republican parties or as Independents / other-affiliated. Following the same criteria as in Study 5, 55 participants who failed an English comprehension check at the start of the study were ejected immediately and I excluded six who self-reported providing low-quality data, leaving 261 (age M = 41.1; 47% women, 48% men, 2% nonbinary or agender; n = 6 did not report gender).

3.4.1.2 Procedure

The procedure was identical to Study 5 with one exception: Rather than rate their preferences for each of ten tweets, participants used separate 5-point scales (1 = Not at all to 5 = Very much so) to rate how much the inparty Senator who authored the tweet seemed: tolerant of people with different opinions; cooperative; rational, logical; to think [outparty]'s views are reasonable; open to changing their political views[7]. These five attributes mirror those from Study 1, so I expected engaging tweets to score higher than dismissing ones.

3.4.2 Results

[7] Due to a coding error, a minority of participants saw this item phrased in the reverse ("unwilling to change their views about the topic of this tweet"); I reverse-scored these participants' responses.

Before proceeding with main analyses, I examined whether the five attributes did in fact distinguish between engaging and dismissing tweets. I tested five multilevel models, each predicting one of the five attributes from tweet type (dismissing = 0; engaging = 1) and random intercepts for participant and tweet; see Table 3.7 and Figure 3.7. As expected, results replicated those of Study 1: Engaging (compared to dismissing) tweets made the Senator who posted them seem more tolerant, cooperative and rational, but also more legitimizing of opponents' views and willing to change their mind.

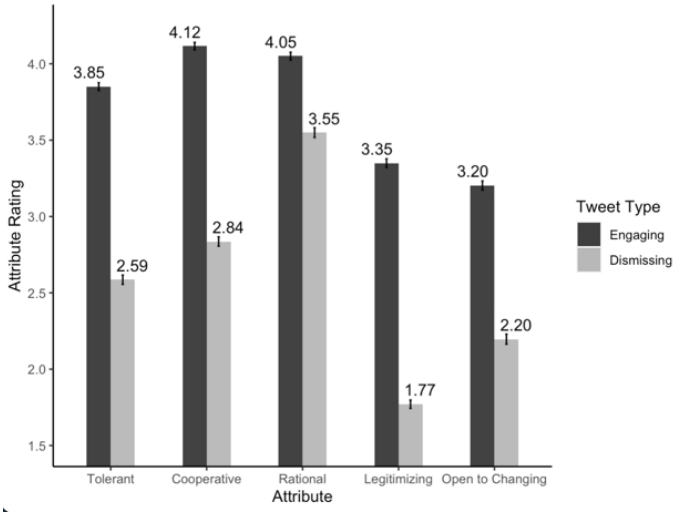


Figure 3.7. Attribute ratings for engaging and dismissing tweets

Table 3.8

Differences in attribute perceptions, engaging vs. dismissing tweets

Attribute	*b* (*SE*)	*t*	*df*	*p*	*d*
Tolerant	1.26 (0.04)	28.32	2470	< .001	1.66
Cooperative	1.28 (0.04)	30.00	2463	< .001	1.68
Rational	0.50 (0.04)	11.84	2457	< .001	0.66
Legitimizing	1.58 (0.05)	33.54	2468	< .001	2.08
Open to changing	1.00 (0.06)	17.36	2470	< .001	1.32

Note. Positive coefficients reflect Senators conveying this attribute more in their engaging tweets.

I then turned to the central question: How did these five attributes (coded in Study 6) relate to preferences of frequent reactors versus of everyone else (reported in Study 5)? I merged the datasets from the two studies into a single dataset containing a row for each individual Study 5 participant's rating of each tweet they saw (4571 total observations), with five additional

columns representing the tweet in question's average rating on each attribute from Study 6. I report analyses that preserve power by using all available datapoints collapsing across Study 5's three (standardized) preference measures; analyses using each measure separately yielded identical results.

Because the tweet types were very different on each attribute, there was a danger that analyses could yield illusory effects: Significant coefficients for a particular attribute might be driven by shared variance with other differences between engaging and dismissing tweets. To avoid this and increase my confidence in ascribing effects to the specific attribute in question, analyses controlled for whether the tweet engaged or dismissed. That is, I predicted (standardized) preferences from one of the five coded attributes (mean centered), participant group (frequent reactors = 0; everyone else = 1), their interaction, tweet type (dismissing = 0, engaging = 1) and random intercepts for tweet and participant (see Table 3.8 for key findings).

Table 3.9

Trait perceptions predicting responses to tweets

Trait	*Effect*	*b (SE)*	*t*	*p*
Tolerant	Interaction with participant group	0.16 (0.03)	5.07	< .001
	Frequent reactors simple slope	0.07 (0.08)	0.87	.384
	Everyone else simple slope	0.23 (0.08)	3.02	.003
Cooperative	Interaction with participant group	0.18 (0.03)	5.50	< .001
	Frequent reactors simple slope	0.11 (0.07)	1.55	.121
	Everyone else simple slope	0.29 (0.07)	3.93	< .001
Rational	Interaction with participant group	0.18 (0.06)	3.13	.002
	Frequent reactors simple slope	0.12 (0.07)	1.65	.099
	Everyone else simple slope	0.30 (0.07)	4.13	< .001
Legitimizing	Interaction with participant group	0.15 (0.03)	5.63	< .001
	Frequent reactors simple slope	-0.17 (0.07)	-2.37	.018
	Everyone else simple slope	-0.02 (0.07)	-0.26	.798
Open to changing	Interaction with participant group	0.17 (0.04)	4.67	<.001
	Frequent reactors simple slope	-0.02 (0.06)	-0.27	.790
	Everyone else simple slope	0.15 (0.06)	2.60	.009

Note: All models had *df* 4551

Significant interactions suggested that frequent reactors responded differently than everyone else to all five attributes. On one hand, frequent reactors disliked tweets that legitimized opponents' views, but their preferences did not relate to any other attribute. Everyone

else showed a complementary pattern: They did not appear to care about whether tweets legitimized opponents' beliefs, but preferred tweets that made the Senator seem tolerant, cooperative, rational, and (surprisingly, given results from Study 1) open to changing their mind about politics. These findings further illustrate the unusual preferences of the minority of people responsible for distributing the majority of feedback to politicians on Twitter. Unlike everyone else, they may dislike efforts to legitimize their opponents' views—a strategy popular among people who feel their political beliefs are based on fundamental moral truths (Goodwin, 2018; Schwalbe et al., 2020).

3.4.3 Discussion

Study 6 further validated the coding from Study 4, confirming that Senators who dismissed opponents unsurprisingly came across as more intolerant, uncooperative and irrational, and less willing to change their minds and legitimize opponents' views. More importantly, findings complemented mediation results from Study 5, by exploring how these characteristics of tweets, rather than tweet perceivers, might explain frequent reactors' unusual preferences: These individuals were uniquely sensitive to how much a tweet legitimized views they opposed, whereas most people seemed more attentive to other characteristics like the how much the tweet made the Senator seem cooperative and rational.

Similar to Study 1, Study 6 found that Senators' tweets received more positive responses from the majority of people to the extent that the Senator seems tolerant, cooperative, and rational. Unlike Study 1, the majority of people were not sensitive to whether the Senator seemed to be legitimizing opponents' views, and in fact *preferred* tweets where the Senator seemed open to changing their mind. That said, Study 1 did find that people generally cared more about the three democratic values than about the two potential violations of moral values. Study 6 is

broadly consistent with that conclusion, in that the largest predictors of the majority's preferences were, as in Study 1, tolerance, cooperation and rationality.

That said, Study 6 is limited in that I did not directly measure frequent reactors' perceptions of the tweets. Results therefore speak only to how the majority of participants (not frequent reactors) perceive the Senators who wrote these tweets, and how their perceptions link to ratings in Study 5. For frequent reactors, these results suggest that they dislike tweets that most people find to be legitimizing, but I cannot know if this is because they perceive these tweets to be more legitimizing or if they dislike legitimizing tweets more (relative to everyone else). Future research can address this by recruiting frequent reactors and measuring their perceptions.

3.5 General Discussion

On Twitter, Senators' tweets receive more rewards—Likes and Retweets—when they advocate for dismissing, rather than engaging with, opposing political views. But most people do not prefer authors of dismissing tweets; in fact, they report approving more of engaging tweets. This disconnect has two explanations. First, Twitter users who most often interact with Senators' tweets have unusual preferences: Unlike most people, they prefer dismissing tweets over engaging ones. This preference was associated with their being more politically extreme (and, in turn, less supportive of compromise, more affectively polarized, and more concerned with their party's status), and with more strongly disliking inparty Senators who seem to legitimize opposing views. Taken together, these variables highlight moralization's potential role in these effects: Frequent reactors' opposition to compromise suggests they may refuse to accept trade-offs on their moral values (Tetlock, 2003); their affective polarization and dislike for tweets that legitimize opponents' views implies contempt for people with different moral values (Finkel et

al., 2020; Rozin et al., 1999). Overall, frequent reactors' political views are seemingly both more extreme and strongly moralized, which may explain their unusual preferences.

Second, although the majority who rarely react to Senators' tweets tend to approve more of engaging, they (intend to) use Twitter's response functions in a way that contradicts that preference: Like frequent reactors, they report greater willingness to Like and Retweet dismissing tweets. Even as a group, these people may not contribute much to the feedback on Senators' tweets. However, when they do, they may amplify the distorted image Twitter reflects of most people's preferences.

3.5.1 Implications

3.5.1.1 Reconciling contradictions in the literature

These findings join a chorus of others underscoring a contradiction between what is popular online and what people reportedly prefer. Recent literature documents how social media rewards that which people do not approve of (Rathje et al., 2022). For example, Chapter 2 found that people prefer engaging yet Chapter 3 found dismissing is rewarded more on Twitter, which could reflect a Twitter paradox (or it could have owed to the different operationalizations of these constructs across chapters). Likewise, lab studies by Frimer and colleagues have shown that most people want allies—both citizens and politicians—to treat opponents with civility (Frimer & Skitka, 2018, 2020), yet politicians' uncivil tweets earn more Likes and Retweets on Twitter (Frimer et al., 2022). More broadly, people recognize that negativity, outrage, and divisive content generates more engagement (Brady et al., 2017; Rathje et al., 2021; Schöne et al., 2021), yet they believe that it should not (Rathje et al., 2022).

Studies 5 and 6 reconcile two apparently different explanations for this contrast: That Likes and Retweets do not reflect endorsement (e.g., Frimer et al., 2022), and that the most

active Twitter users are a unique group with unusual preferences (e.g., Bor & Petersen, 2022). I find support for both explanations. The minority of active users do have unusual preferences—they genuinely prefer dismissing. In contrast, most people (intend to) Like and Retweet content they do not necessarily approve of. This explanation was put forth by Frimer and colleagues, who suggested that Likes and Retweets do not indicate endorsement but did not test this by comparing people's attitudes with their Twitter behavior.

These findings also speak to a third explanation that has not been formally tested: that social media platforms' algorithms promote such negativity because it increases user engagement (see Rathje et al., 2022). Academic psychologists cannot directly access social media platforms' algorithms to study them, but the findings above reveal that negativity spreads further on social media, and journalists' reports corroborate the idea that algorithms are to blame (e.g., Roose et al., 2020). That said, algorithms primarily act by changing what people see in their feed—engagement itself still must come from user behaviors. Rather, if algorithms promote engagement through negativity, that must be because (some) users are willing to engage with negative content. The present work identifies the users most responsible and in doing so, provides better support for a comprehensive explanation for a puzzling pattern in recent literature and in real social media behavior.

3.5.1.2 Extending and adding to theories of who prefers engaging vs. dismissing, and when they prefer it

These findings dovetail with the internal meta-analysis results reported in Chapter 2, further supporting theorizing that intolerance emerges at ideological extremes (Crawford & Pilanski, 2014). Moreover, extremists were especially likely to be frequent reactors. This fits with findings suggesting people with the most extreme political views are most politically active

(e.g., more likely to vote, protest, post about a candidate on social media; Pew Research Center, 2022a), and most often attempt to influence politicians through both social media and traditional methods (e.g., writing letters, calling; Birkhead & Hershey, 2019).

The present results also highlight novel factors that shape whether and how much people prefer engaging over dismissing. In particular, it matters what measure people use to indicate their preference, and how engaging and dismissing are conveyed. In Study 5, I observed different patterns of preferences when people reported their approval of a tweet compared to their feelings toward the authoring Senator: Most people approved more of engaging tweets than dismissing ones but felt similarly warm towards Senators who wrote engaging vs. dismissing tweets. Senators are public figures so people might have already had preexisting opinions toward them, which might have played a bigger role than people's feelings toward the content of the tweet itself; if so, measuring approval of the message better indicates people's attitudes towards engaging and dismissing.

I also observed different preferences across studies that used the same feeling thermometer measure: people felt similarly warm towards Senators who engaged vs. dismissed in Study 5 (as well as Appendix Studies S2-S4) yet felt warmer towards Senators who engaged in Studies 1 and 2 (and in Appendix Study S1). These different effects neatly align with the different stimuli sources of each study, such that experimenter-generated stimuli produced a preference for engaging whereas stimuli sourced from Senators' real tweets produced no preference. Exploratory analyses mentioned in Study 3 and 4's method suggest that people perceive these real tweets as conveying their intended concepts, and Study 6 shows that Senators conveyed similar attributes as in Study 1. It therefore seems unlikely that these divergent effects reflect a failure to capture intended constructs. That said, as noted in Footnote 5, Senators

seemed more tolerant, cooperative, and rational in their real dismissing tweets compared to in experimenter-generated tweets, which likely suppressed preferences for engaging. This finding suggests that in the real world, Senators might convey engaging vs. dismissing differently than these concepts are conveyed in conceptually precise, experimenter-generated stimuli.

3.5.2 Unanswered questions

Chapter 3 suggested that patterns on Twitter are attributable to frequent reactors and attempted to paint a psychological portrait of this unique group, but the causal relationship between these users' psychology and behaviors remains unclear. Using mediation, Study 5 showed that frequent reactors' preference for dismissing is explained by their extremism and, in turn, their tendency for moralized politics (i.e., opposing compromise, favoring the inparty, and wanting inparty status, all of which are associated with moralized political beliefs; Garrett & Bankert, 2020; Finkel et al., 2020; Ryan, 2017; Ward & Tavits, 2019). Complementing this, Study 6 used cross-sectional data to show that frequent reactors were unique in that their responses to a Senators' tweet was related more to how much that Senator legitimized opposing views. These data sources do not establish causality, but prior work speaks to possible causal routes.

For one, extremists may flock to Twitter: They tend to engage more with politicians by all available means (e.g., town halls; calling or writing to lawmakers; Birkhead & Hershey, 2019), but compared to traditional means, Twitter makes it especially easy for people with extreme or otherwise unpopular views to directly engage with politicians' messages (Zhuravskaya et al., 2020). Or perhaps actively joining political discourse on Twitter makes people more extreme: Social media use exacerbates political divisions (Bail et al., 2018), making negativity seem acceptable and opponents seem intolerable (Tucker et al., 2018; Wilson et al.,

2020). Future work could test this directly. More broadly, a variety of studies have showed bidirectional links between moral values, political engagement, and extremity (Hatemi et al., 2019; Parker & Janoff-Bulman, 2013; Skitka & Bauman, 2008; Van Zomeren et al., 2011; for a review, see Guan et al., forthcoming).

I also did not establish why frequent reactors express their preferences on Twitter, which would help determine whether extremists behave this way across contexts or only on expressive platforms with certain features. Future work could make use of exciting new technological advancements, such as tools to create a social media environment for participants. This would allow researchers to vary features of social media platform (e.g., whether content is moderated or censored; whether users' behaviors such as Likes and Retweets are public or private), and this data could be complemented by examining how participants' responses on the platform map onto their offline behaviors.

These results describe and explain Twitter dynamics in contemporary America's polarized political landscape, but would they replicate in other platforms or offline? These results should replicate best on platforms that people select into to talk politics such as Facebook (Kim et al., 2021) or politicized subreddits (Sun et al., 2021) compared to elsewhere (e.g., Instagram; Pinterest; r/rarepuppers). Twitter has a unique reputation for political content and discourse (Mukerjee et al., 2022), perhaps because political leaders post here more than other platforms (Pew Research Center, 2020a), thereby attracting people who want to engage with such content (i.e., political activists and extremists). These results might also replicate offline: People who laud dismissing may also look to voice their opinions at in-person town hall meetings and dinner parties (see Bor & Petersen, 2022), but their online behaviors are visible to more people and thus may have more influence.

Moreover, the present results speak to U.S. political and social media dynamics, but would they generalize elsewhere? Countries differ in which social media platforms they use as well as their access to technology more generally. Still, some evidence suggests these dynamics would generalize: In other places, hostility on social media also comes from a small subset of the population—in Denmark (on Facebook, Bor & Petersen, 2021) and in Slovenia (on Twitter, Evkoski et al., 2022). But elsewhere the opposite is true (e.g., in Italy on Youtube; Cinelli et al., 2021). I speculate that politicians' dismissing behavior may fare better in places (cultures, nations) where politically engaged citizens imbue their political beliefs and identities with moral significance (Finkel et al., 2020), to the point that democratically sanctioned compromise feels like a betrayal of sacred values (Guan et al., 2023). When politically engaged citizens feel this way, popular venues for political discourse (like Twitter in America) will overrepresent their antagonistic opinions and admiration for leaders who dismiss rather than engage. These results may also replicate better in other countries with moralized political divides (e.g., Spain; Viciana et al., 2019), with only two viable parties, or without political institutions geared toward building consensus (e.g., proportional representation); in all cases due to increased political conflict (Gidron et al., 2020; Lijphart, 2010). Nonetheless, within other countries, effects should tend to replicate in contexts and on platforms where people self-select into political discourse.

3.5.3 Potential influence of how engaging, dismissing were operationalized

Contrasting Chapter 2, Chapter 3 used different operationalizations of engaging and dismissing, which may have shaped how people responded to these tweets. The present data do not allow me to directly compare how these different operationalizations affected attitudes. That said, I observed similar patterns in both cases: Most people approved more of Senators' tweets that engaged (deeply) with opposing views over those that dismissed them (outright), which

echoes the same preference for (initially) engaging with opponents' views over dismissing (by overlooking) them, as seen in Chapter 2. This is consistent with the idea that people typically prefer engaging over dismissing, regardless of how it is operationalized. That said, it seems likely that people perceive these actions somewhat differently. Anecdotally, I noticed that Senators—when tweeting their outright dismissal of opponents' views—often provided justifications for why the opponents' view was not worth considering. In doing so, Senators can imply that they already know the opponents' views and have good reasons to think those views are not worth tolerating or cooperating with. Comparing this strategy to dismissively overlooking opponents' views, Senators who vocally dismiss opponents outright and with justification may seem less willfully irrational and intolerant. Of course, this is merely speculative—I leave it to future research to study how perceptions of engaging and dismissing depend on how it is instantiated, and how this shapes people's preferences for these behaviors.

Chapter 4: (Mis)perceiving the preference for engaging

Upon seeing either the positive feedback that dismissing receives on Twitter, polarization more broadly emphasized there and in other media formats, or the seeming prevalence of cancel culture, people might infer that their allies endorse dismissing—that is, they may fail to realize how much their allies prefer engaging. Moreover, they might conform to this perceived preference, even though it contradicts what they themselves approve of most. Chapter 4 considers this idea and its implications.

Studies 7 and 8 first test whether people accurately perceive their allies' preference for engaging; I predict that they do not. If correct, this raises questions of what causes these misperceptions and how to correct them, as they might inhibit engaging behavior. Based on the idea that perceived polarization is a culprit, Studies 9 and 10 test whether reducing it disinhibits engaging via reducing misperceptions about engaging's popularity.

Chapter 4 is among the first attempts to apply pluralistic ignorance theorizing to the political realm (see Van Boven, 2000); doing so promises not only theoretical insight but also implied routes for potential interventions aimed at increasing engaging behavior by reducing misperceptions.

Chapter 4 also returned to operationalizing the key constructs as initial engaging and dismissively overlooking, similar to Chapter 2. This was because I conducted Chapter 4's studies shortly after and in response to Chapter 2's, without having yet studied how different operationalizations influence those responses. I leave it to future work to test Chapter 4's studies with other operationalizations.

4.1 Study 7

Participants watched a text message conversation unfold between two political allies. I

manipulated between subjects whether one of these allies, the actor, engaged with or dismissively avoided shared opponents' views. I also manipulated between subjects whether participants imagined themselves to be the other texter, or merely a third-party witness to the conversation. At the end of the video, the actor tried to make social plans with the other texter. The participant then reported the likelihood that the other texter (i.e., themselves or a fellow ally) would agree to the plans. I predicted participants themselves would report being more likely to agree to plans with someone who engages with opposing views, but that they would estimate that a fellow ally would not be so eager to connect.

4.1.1 Method

4.1.1.1 Participants.

Study 7 was preregistered; everything below followed the *a priori* plan unless otherwise specified. Table 4.1 reports sample characteristics from Studies 7 and 8, which both sampled Americans from Amazon's Mechanical Turk.

Table 4.1

Sample characteristics, Studies 7 and 8

	N	Exclusions due to				
Study	recruited	Failed checks	Technical issues	Final N	Gender	Age
7	606	13	7	586	47% male, 52% female, 1% non-binary	37.0
8	709	46	16	667	50% male, 50% female	38.1

4.1.1.2 Procedure.

Participants first reported their political beliefs, allowing me to later show them a text exchange between two texters who shared these beliefs. They read descriptions of four hotly debated issues—abortion, affirmative action, gun regulations, and immigration—and for each, chose which of two positions they preferred. The survey randomly selected one of these issues to be the subject of the texting conversation.

Participants then watched an experimenter-generated video of a text message conversation between two individuals; see Figure 4.1. They either imagined themselves to be one of the texters, or to merely be witnesses to the exchange. The conversation depicted the texters first establishing that they were political allies, both sharing the participant's stance on the randomly selected issue (e.g., the left [right] image depicts the conversation seen by pro-choice participants whose assigned issue was abortion and whose assigned actor engaged [dismissed]). Then, one texter—the actor, always named Jordan—sent text messages about how they had made efforts to either engage constructively with or dismiss opposing views. Jordan then invited the other texter—the participant, or another texter named Alex—to upcoming social plans.

Figure 4.1 Images from text message conversation used as Study 7 stimuli

The video ended before the reactor responded. Participants reported the chances that the reactor would agree to the social plans: *"As you saw at the end of the video, Jordan asked [you / Alex] to hang out. What are the chances that [you / Alex] would agree to hang out with Jordan?"* using a 201-point sliding scale anchored at one end with *Definitely WILL NOT hang*

out with Jordan, at the midpoint with *Not sure either way*, and at the other end with *Definitely WILL hang out with Jordan*.

4.1.2 Results

A linear model predicted participants' responses from actor behavior (engage = 1, dismiss = 0), reactor identity (self = 1, perceived ally = 0), and their interaction. The predicted interaction was significant, $b = 21.94$, $SE = 6.66$, $t(582) = 3.30$, $p = .001$; see Figure 4.1 and Table 4.2 for simple slopes. Replicating Chapter 2, participants themselves preferred the actor who engaged constructively with, rather than dismissed, opposing political perspectives. Consistent with misperceptions, participants assumed their ally would hold no such preference. In particular, people overestimated how much their allies would like dismissing.

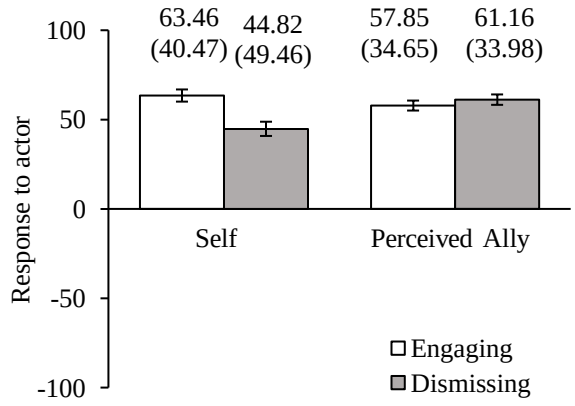


Figure 4.2. Own vs. perceived ally's attitudes toward engaging versus dismissing actors, Study 7

Table 4.2

Results of simple slopes tests, Study 7.

Simple slope effect	*b (SE)*	95% CI	*t*	*p*	*d*
Own response to engaging vs. dismissing	18.64 (4.69)	8.25, 29.03	3.53	< .001	0.41
Perceived ally response to engaging vs. dismissing	-3.30 (4.73)	-11.24, 4.63	-0.70	.486	-0.10
Own vs. perceived ally response to engaging	5.61 (4.78)	-3.20, 14.41	1.17	.241	0.15
Own vs. perceived ally response to dismissing	-16.34 (4.64)	-25.98, -6.69	-3.52	< .001	-0.38

Note. *b* refers to the beta for the difference between the two targets. *SE* represents standard error. There were 582 *df*.

4.1.3 Discussion

Study 7 revealed the hypothesized misperception. Participants themselves preferred political allies who engaged with, rather than dismissed, opposing perspectives but assumed they

were unique in this regard, such that their allies would have no preference. Study 8 aimed to replicate this initial finding, and also add a control condition where the ally neither engaged with nor dismissed opposing views.

4.2 Study 8

In addition to the extra control condition, Study 8 built on Study 7 by studying responses to engaging and dismissing across a different type of political divide: broad ideologies (i.e., liberal vs. conservative) instead of specific policy stances (as in Study 7). Study 8 was preregistered; everything below followed the *a priori* plan unless otherwise specified. Table 4.1 above reports the sample characteristics for Study 8, also drawn from Americans on Mechanical Turk. Finally, Study 7's video had one texter, Alex, reveal their political beliefs by voluntarily disclosing them outright over text, which may be counter-normative and convey that Alex is especially interested in and devoted to political topics; as a result, participants might have assumed Alex was more extreme than they were, which could explain Study 7's effects. Study 8 therefore more carefully matched Alex's extremity to the participant's own.

4.2.1 Method.

Study 8 followed a similar method as Study 7, with four differences. First, rather than reporting their policy stances, participants reported their ideological orientation measured on a six-point scale (1 = extremely liberal, 2 = liberal, 3 = slightly liberal; mirrored for conservatives from 4-6). I showed participants an exchange between two texters who shared their ideology *and* extremity (e.g., those who identified at 1 = extremely liberal, 2 = liberal, or 3 = slightly liberal saw an exchange between texters who, in the conversation, self-described as *pretty liberal*, *liberal*, or *leaning liberal*, respectively).

Second, in addition to the engaging and dismissing conditions, Study 8 included a control condition where the actor wanted to learn more about something outside of the political domain (an inventor's background and process for inventing); see Figure 4.3 (the left (middle / right) image depicts the depicts the conversation as seen by liberal participants assigned to the engaging (dismissing / control) condition).

Third, the specifics of the conversation between the two texters were different; it ended with the actor inviting the reactor to go to a bookstore.

Figure 4.3. Images from text message conversation used as Study 8 stimuli

Fourth, the response scale participants used to predict the other texter's response was condensed, having 101 rather than 201 points.

4.2.2 Results.

First, I predicted participants' responses from actor behavior (dummy coded with control condition as reference), reactor (self = 1, perceived ally = 0), and their interaction; see Table 4.3 and Figure 4.4. An initial preregistered model excluded the control condition for a more direct comparison with Study 7 and replicated its effects (see Appendix). Significant interactions for the contrast with dismissing but not engaging led me to examine the simple slopes; see Table 4.3. Participants reported that they would dislike dismissing, compared to engaging and compared to control actors. This replicates and extends my prior findings. They predicted that their fellow ingroup members would feel differently: That they would dislike *engaging*, compared to dismissing and marginally compared to a control ally. This difference from the own response condition confirms the existence and direction of misperceptions (though in Study 7 participants expected no preference between engaging and dismissing, rather than an outright dislike of dismissing).

I also separately examined participants' own feelings versus their predictions about each actor. Participants thought they would like both a control actor and one who engaged more than an ally would, but that they and their allies would feel similarly about dismissing. That is, people generally like their allies more than they think their allies like each other, with dismissive allies being a special exception to this rule, because people themselves dislike dismissing while expecting their allies not to.

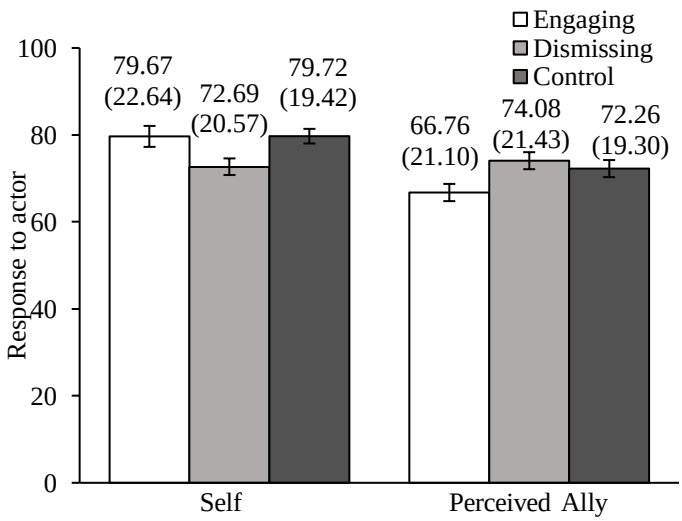


Figure 4.4. Own vs. perceived ally's attitudes toward engaging, dismissing, and control actors, Study 8

Table 4.3

Results of simple slopes tests, Study 8.

Effect	*b (SE)*	*t*	*p*
Interaction, engaging vs. dismissing	14.60 (4.00)	3.65	< .001
Interaction, control vs. dismissing	-8.85 (3.86)	2.29	.022
Interaction, control vs. engaging	5.75 (4.05)	1.42	.156
Simple slope: Own response to engaging vs. dismissing	7.28 (3.03)	2.40	.017
Simple slope: Perceived ally response to engaging vs. dismissing	-7.32 (2.80)	-2.62	.009
Simple slope: Own response to control vs. dismissing	-7.03 (2.62)	-2.68	.008
Simple slope: Perceived ally response to control vs. dismissing	1.82 (2.83)	0.64	.520
Simple slope: Own response to control vs. engaging	0.25 (2.85)	0.09	.930
Simple slope: Perceived ally response to control vs. engaging	-5.50 (2.87)	-1.91	.056
Simple slope: Own vs. perceived ally response to control	7.46 (2.77)	2.69	.007
Simple slope: Own vs. perceived ally response to engaging	13.20 (3.05)	4.33	< .001
Simple slope: Own vs. perceived ally response to dismissing	-1.39 (2.78)	-0.50	.617

Note. *b* refers to the beta for the difference between the two targets. *SE* represents standard error. There were 660 *df.* Results involving the comparison between engaging vs. dismissing come from the same model but with dismissing coded as the reference condition.

4.2.3 Discussion

Study 8 mostly replicated Study 7. In both studies, participants saw their own preferences for engaging as different from their allies', saying they themselves would dislike dismissing, compared to engaging (and control actors, which were only included in Study 8). One slight difference was that in Study 7 they assumed their allies would hold no preference between engaging and dismissing, in Study 8 they assumed their allies would dislike engaging.

Study 8's control condition permits some additional observations. First, people's estimates of how much their allies would like dismissing were accurate in an absolute sense: There was no difference between how much people thought their allies would like dismissing and how much the entire group reported liking dismissing. But they were inaccurate in a relative sense: People thought their allies would like dismissing better than engaging, and just as much as control targets, even though the entire group reported liking dismissing the least.

Prior work has suggested that people worry that engaging with opposing views might damage their relationships with the opponents (Frimer et al., 2017). My findings hint that people also worry about this damaging their relationships with allies: They think their allies feel neutrally (Study 7) or perhaps even negatively (Study 8) toward others who engage with shared opponents' views. In other words, it may be that people engage less with outgroup views in the presence of allies (as in Moore et al., 2021) because they expect those allies to disapprove of this type of behavior.

More broadly, Studies 7 and 8 are consistent with the sort of misperceptions that could fuel pluralistic ignorance (Prentice & Miller, 1996). Most people privately endorse engaging with opponents' views more than dismissing them yet misconstrue social norms as dictating the opposite. This may explain why in recent decades, people have become less and less likely to engage with opposing views (Rodriguez et al., 2017): Evidence suggests people refrain from engaging with opposing views when allies are watching (Moore et al., 2021), suggesting they are publicly conforming to a norm they privately disagree with. This could in turn reinforce the norm for the whole community.

4.3 Study 9

If people's misperceptions of what their allies want stem from perceived polarization, reducing perceived polarization should correct them. Moreover, if what people think their allies want drives their behavior, then reducing perceived polarization, by correcting the misperception, should disinhibit them from engaging with opponents' perspectives. Based on this reasoning, my primary goal in Study 9 was to test a potential intervention to increase engaging behavior: Reducing perceived polarization. If the intervention succeeded, I planned to test the role of my proposed mechanism, (corrected) perceptions of allies' preferences.

In a single-factor, between-subjects design, I randomly assigned participants to watch an evidence-based video about polarization in America either being high, in absolute terms (*more polarization* condition) or being low, relative to what most Americans think (*less polarization* condition). I then measured participants' engagement with opposing views, predicting that those induced to perceive less polarization would be more willing to engage.

4.3.1 Method

4.3.1.1 Participants.

Study 9 was not preregistered; it was a pilot study with a large sample and, if it produced promising results (which it did not), I intended to preregister and replicate it. I recruited 729 Americans from Prolific Academic, primarily partisans (with Democrats and Republicans respectively making up ~40% of the sample) since they would presumably be more prone to engaging with congenial views than unaffiliated people. The remainder of the sample (~20%) were independents and other-affiliated persons, for generalizability. From this initial sample, the survey ejected fourteen people who failed an English comprehension check (the same as in prior studies); I further excluded 63 who failed an attention check and two who self-reported providing low-quality data, leaving 650 (age M = 43.57, 49% female, 50% male, 1% nonbinary).

Sensitivity analyses revealed that the sample size for comparing between two cells provided 80% power to detect small effects (d = .20).

4.3.1.2 Procedure

Participants signed up for a two-part study examining how people form impressions about others. In reality, there was no second part to the study; this was a cover story included to convince participants that they would later exchange views and interact with another participant.

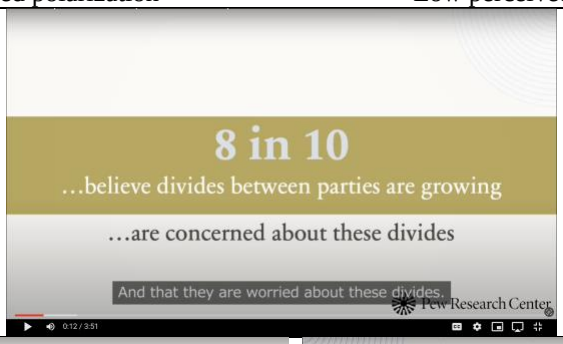
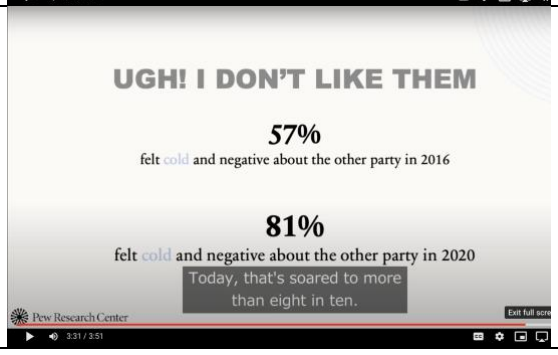
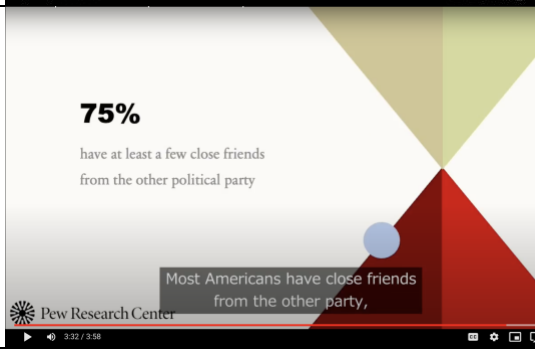
Participants completed an English comprehension check (as in Study 2) then reported demographics, including age, gender, and ethnicity, and ideological orientation using a 6-point scale from 1 = *Strongly Liberal* to 6 = *Strongly Conservative*.

4.3.1.2.1 Manipulation

Next, participants viewed a 4-minute video to manipulate perceived polarization. The video ostensibly provided background information and reviewed recent political research, so participants could then report their opinions of the political climate. Participants watched one of two videos (randomly assigned between-subjects) that showed how levels of polarization among Americans are either high (in absolute terms), or low (in relative terms compared to what most Americans think). I used videos rather than articles to make the manipulation strong and engaging, and created the videos using best practices for interventions when possible (i.e., social proof; Walton & Wilson, 2020), to improve their chances at changing participants' presumably well-established pre-existing beliefs about how polarized Americans are.

Each video had four parts; for examples, see Figure 4.5 (for exact videos, see this link). Part 1 was the same in both conditions. It provided statistics about how Americans perceive vast divides, suggested that this perception comes from the media, and posed the question, Are media depictions accurate? In Parts 2 and 3, the videos depicted ideological and affective polarization

respectively as high (e.g., real partisans disagreeing, public opinion polls showing animosity between parties) or low (e.g., real partisans agreeing, public opinion polls showing cross-party friendships and a desire for respect and tolerance); evidence was drawn from real sources (high polarization: Cut, 2015a-c; low polarization: Voelkel et al., 2023, "Misperception film" submission). In Part 4, the videos concluded that the media either accurately reflects huge political divides in America (in the high condition) or exaggerates them (in the low condition).

	High perceived polarization	Low perceived polarization
Part 1 (identical between conditions)		
Part 2		
Part 3		

Part 4

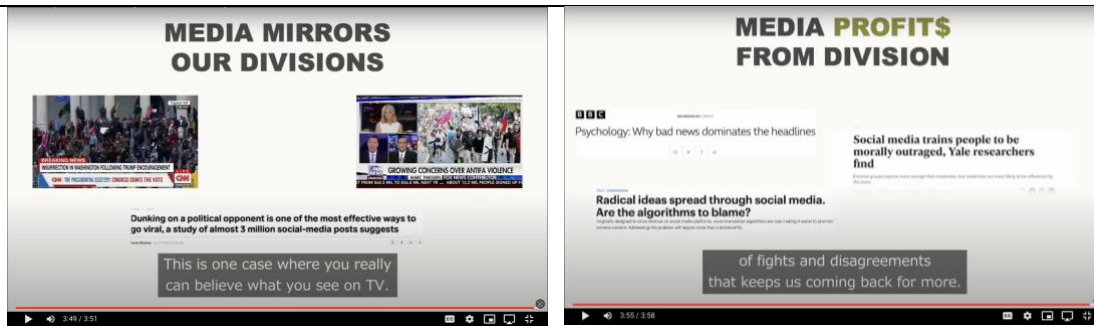


Figure 4.5. Key points in videos manipulating perceived polarization, Study 9

After watching the video, participants wrote a few sentences about their impressions of it. Participants rarely said they did not believe the videos' key point, but rates differed slightly across video conditions: Relatively more people expressed disbelief (11%) or skepticism (6%) in the low polarization video's conclusion, compared to 5% of participants expressing disbelief in the high video. This makes sense given that people likely already perceive high polarization, so it may be harder to change their mind on this topic. That said, 17% of participants who saw the high video said it highlighted *more* agreement than they expected—this was despite the video portraying vast gaps in policy support (as seen in Figure 4.5's part 2 left panel). If participants expected larger gaps than that, this speaks to how much they overestimate polarization. Study 9's discussion considers the implications of how the video was perceived.

4.3.1.2.2 Dependent measure

My key measure of engaging with opponents' views was participants' willingness to read other participants' statements explaining their preferred policy stances. To make this measure plausible, I first asked participants to write such a statement, which would ostensibly go into the set that subsequent participants could choose from. Participants reported their position on six political issues (abortion, affirmative action, climate change regulations, gun regulations, immigration, and universal healthcare), then chose one of these to explain in a short paragraph.

Participants then learned that, in a follow-up survey, they would read one similar statement written by a previous participant. I first explained that they could have some degree of choice in what they read:

> **You will read a paragraph written by another participant**, like the one you just wrote. Usually, participants say that they want to choose what they read. At the same time, for our research, we need to make sure every participant's paragraph is read by at least one other participant.
>
> So, as a compromise, **you can choose which kinds of paragraphs you are - and are not - willing to read.** You'll make this choice on the next page.

I then instructed them to choose at least six of the 12 possible types of statements—one for each of two possible positions on each of the six issues, e.g., "about abortion from an author who is pro-life":

> Please **select boxes for at least 6 topics that you are willing to read** during the upcoming follow-up survey. If you are **unwilling to read** a particular topic, you can **leave the box unselected**.
>
> You are required to choose at least 6 topics, but **you can choose more than 6** - as many as 12 - if you are open to reading them.

I used participants' selections, along with their self-reported positions, to compute two engaging scores. First, *absolute engaging* was the number of counter-attitudinal statements that a participant selected out of the six that were available. Participants could be willing to read anywhere between zero and all six available counter-attitudinal statements. This is a straightforward index of participants' willingness to hear from opponents, but it does not account for their general interest in reading statements. For example, two participants might both select three statements by opposing authors, but the first might also have selected three statements by congenial authors (for a total of six), while the second selected all six congenial statements (for a total of nine). I therefore also computed *relative engaging* by expressing the number of opposing statements participants selected as a proportion of the total number of statements they selected (ranging from six to 12). For example, the same two example participants would have relative scores of 0.50 and 0.33, respectively.

This dependent measure task is novel but similar to those used in prior work on selective exposure (e.g., Dorison et al., 2019; Frimer et al., 2017; Wojcieszak et al., 2020). However, those prior measures ask participants to choose between hearing congenial or opposing views on a single issue. My measure both incorporates generalizability across issues and disentangles the *desire to not* engage with opposing views with the *desire to* engage with congenial ones: Whereas a single choice between congenial and opposing views does not permit participants to hear from both, my measure allowed participants to choose up to all 12 options, meaning they were not forced to choose congenial views at the expense of opposing ones. I also told them they would only ultimately read one statement, making it easy for them to indicate a willingness to read as many as they wanted without incurring costs of having to actually do more work.

This measure assesses private, not public, engaging behavior, even though my reasoning should apply more to the latter: If people are conforming to their perceptions of their allies' preferences, presumably they would do so particularly when in the presence of those allies (Fehr & Fischbacher, 2004). Indeed, my initial design included a second, public measure following this first private one: The first 409 participants learned, after privately reporting their statement choices, that an ally would see their choices and could invite them to a lucrative group task in a follow-up study. These participants then had the chance to make new statement choices, now knowing that allies could see and react to their choices. But analyses on this first set of participants suggested this public measure was even less sensitive to the video manipulations than the private one, so I eliminated it for the remaining 320 participants. Private engaging is still practically important, and likely still follows from peer approval, as people often internalize norms from their peer group and follow them privately (Kelman, 1958; Ajzen, 1991); indeed, at

least one prior study manipulating perceived ally support for engaging shows that this influences private behaviors (e.g., Wojcieszak et al., 2020).

4.3.1.2.3 Additional measures

After making their choices, a manipulation check assessed perceptions of polarization: Participants saw the prompt "In your opinion, how much do Americans..." with two competing statements, "...disagree with each other about politics?" and "...dislike people with whom they disagree about politics?" (1 = Not at all; 10 = Very much). I averaged responses to both stems (r = .63, p < .001).

I also assessed two variables that should correlate with participants' engaging behavior, to help validate my key novel measure: *perceived illegitimacy* and *political homogeneity*. To measure *perceived illegitimacy* of outgroup views, similar to Study 2's manipulation check of that construct, participants reported their agreement with the statements, "They are bad, immoral" and "They are irrational, illogical" (1 = Strongly disagree, 5 = Strongly agree) with reference to people they disagree with politically; I averaged responses to the two items (r = .72, p < .001). To measure *political homogeneity* of one's social network, participants answered the question, "Of the people you regularly interact with, what portion share your political beliefs?" (1 = Practically none, 5 = Practically all).

4.3.2 Results

4.3.2.1 Manipulation check and DV validation

First, I used a linear model to predict manipulation check ratings from polarization video condition (low = 1, high = 0). Indeed, participants who saw the low polarization video perceived less dislike and disagreement between partisans, b = -1.56, SE = 0.14, t(648) = -11.21, p < .001, d = 0.88, compared to participants who saw the high polarization video. Incidentally, people who

saw the low polarization video also rated views that they disagree with as less illegitimate (i.e., less immoral and irrational), b = -0.26, SE = 0.08, t(648) = -3.23, p = .001, d = 0.25. This conceptually replicates findings from the strengthening democracy challenge, from which the low polarization video drew one of its clips (Volkel et al., 2023). In combination with Study 2, this might also suggest that the low polarization video would make people favor allies who engaged with opponents' perspectives even more, though my central question here is how it might influence people's own behavior.

Next, I tested whether my measure of engaging behavior revealed an overall preference for reading congenial information, as has been documented by dozens of prior studies on selective exposure (for reviews, see Hart et al., 2009; Sears & Freedman, 1968). A linear model predicted relative rates of engaging with outgroup views, centered around the midpoint of choosing opposing statements at a rate of 0.50. This model's intercept was significantly below the midpoint (b = -0.05, SE = .009, t(646) = -5.65, p < .001), revealing that people opted for ingroup views more often than outgroup ones. This effect (d = -0.44) was similar in size to most other studies of selective exposure (e.g., meta-analytic estimate of d = .36; Hart et al., 2009), but the rate (people choose congenial information 55% of the time) was not as large as in more recent studies of selective exposure to political information (e.g., people chose congenial information about two-thirds of the time in Frimer et al., 2017).

As a final validation check, I tested how my measure of engaging behavior correlated with perceived illegitimacy of outgroup views and with political homogeneity. As expected, people were less willing to choose opposing views when they thought these views were more illegitimate (absolute engaging: b = -0.13, SE = 0.066, t(645) = -2.04, p = .042, d = -0.16; relative engaging: b = -0.02, SE = 0.008, t(645) = -2.61, p = .009, d = -0.21), consistent with

theories that people refrain from engaging with opposing views to avoid feeling dissonance or generally upset (Dorison et al., 2019; Festinger, 1957; Frimer et al., 2017; Schwalbe et al., 2020). Likewise, people who reported consorting mostly with allies were also less willing to choose opposing views (absolute engaging: b = -0.14, SE = 0.076, t(645) = -1.79, p = .074, d = -0.14; relative engaging: b = -0.03, SE = 0.010, t(645) = -2.62, p = .009, d = -0.21), in line with prior work on homogeneity and exposure (Himelboim et al., 2013; Stroud & Collier, 2018). That said, these relationships were relatively weak.

4.3.2.2 Main analyses

Having confirmed that the manipulation was effective and tentatively validated the DV, I tested whether the polarization manipulation influenced people's engaging behavior. I tested two linear models (one for each engaging score) using as a predictor polarization video condition (high polarization = 0); in both models the key coefficient was null (absolute score: b = 0.11, SE = 0.14, t(645) = 0.81, p = .418, d = 0.06; relative score: b = 0.02, SE = 0.02, t(645) = 1.06, p = .292, d = 0.08). In other words, the video manipulation did not change engaging behavior, regardless of how it was operationalized. That said, I followed this with an exploratory test to see if perceived polarization—as operationalized by the manipulation check—was at least correlated with engaging behavior, even if not causally. Results supported this idea, as people were less willing to choose opposing views when they perceived larger partisan divides (absolute engaging: b = -0.09, SE = 0.035, t(645) = -2.66, p = .008, d = -0.21; relative engaging: b = -0.01, SE = 0.004, t(645) = -2.21, p = .028, d = -0.17). It therefore seems that perceived polarization is related to engaging, but this effect is small and not causal.

4.3.3 Discussion

I predicted that people, when induced to perceive low political polarization, would be more likely to engage with and hear out opposing views. Results did not support this hypothesis. People were equally likely to engage with opposing views when they were made to perceive high vs. low polarization.

Why did the videos fail to influence behavior? Perhaps the answer has to do with the study's methods. On one hand, this null effect was not because the video failed to change people's perceptions of polarization—it did. On the other hand, when validating my measure of engaging behavior, I found consistent but small relationships with expected constructs, so perhaps this measure is not particularly sensitive. If so, one might imagine that increasing sample size to compensate for measurement error could have yielded different results. At the same time, Study 9 already used a relatively large sample (about 325 people per video condition), powered to detect even quite small effects. And indeed, the study was powered enough to observe selective exposure: People sought out congenial views 55% of the time, compared to opposing views 45% of the time. This rate was lower than in other studies documenting political selective exposure, but this may owe to unique methodological features that are *strengths* of the present design: Those studies changed the choice incentives—they had people choose between accepting less (more) money to hear congenial (opposing) views (e.g., Frimer et al., 2017)—and/or did not allow them to express willingness to hear both sides (e.g., Minson et al., 2020; Wojcieszak et al., 2020). Taking these results together with findings from the selective exposure literature, it seems likely that people prefer to hear congenial views (Sears & Freedman, 1968), especially when forced to choose, but the existing literature might not portray people's willingness to seek out a variety of perspectives. As for explaining why these videos did not affect engaging behavior, it seems unlikely that this owes to this measure of engaging being invalid.

Another possibility is that the low polarization video *did* increase engagement, but that the high polarization video *also* increased engagement—a possibility that I cannot rule out because I did not measure people's default engagement rates in the absence of these video manipulations. For example, after seeing the large divides depicted in the high polarization video, perhaps people felt personally responsible for trying to bridge divides, compelling them to hear from opponents. Participants' open-ended responses do not provide positive evidence of this—no one mentioned it explicitly—but they also did not rule it out. That said, open-ended responses *did* reveal an ironic effect: Many participants took from the high polarization video that Americans agree *more* than they expected (maybe these people expected *no* agreement), which could have led them to engage through the same mechanism proposed for the low polarization video. On one hand, this suggests that the high polarization video should be revised to better manipulate what it was intended to. On the other hand, if both videos reduced perceived polarization relative to a control, this would be heartening because both videos could serve as potential interventions. In any case, future research should more thoroughly pre-test stimuli, include a control condition, and measure potential mechanisms.

A third possibility is that perceived polarization does not causally affect engaging behavior. I had predicted it would do so by influencing people's perceptions of what their allies approve of, which in turn would influence their behavior. The null result I observed could therefore be because both or just one of these sequential effects do not exist. There is already strong empirical and theoretical precedent to suspect that the second does: That people's perceptions of their group's preferences influence their behavior (Ajzen, 1991; Moore et al., 2021; Sherif & Sherif, 1953; Wojcieszak et al., 2020). Study 10 therefore narrowed in on the

more novel first effect, asking whether perceived polarization influences people's perceptions of what their allies want.

4.4 Study 10

Study 10 tested whether perceiving high versus low polarization changes perceptions of allies' support for engaging with opposing views. I combined Study 9's polarization videos with Study 8's dependent measure, in a 2 (high vs. low perceived polarization) × 2 (reporting own vs. predicting allies' reactions) ×2 (reacting to engaging or dismissing behavior) fully between-subjects design (without a control actor condition as in Study 8).

4.4.1 Method

4.4.1.1 Participants.

Like Study 9, Study 10 was not preregistered because I intended to follow up on any promising results (of which there were none) with a preregistered replication. I recruited 403 Americans from Prolific Academic aiming for equal numbers of Republicans, Democrats, and Independents or other-affiliates. From this initial sample, the survey ejected nine who failed an English comprehension check (as in Study 2); I further excluded three participants who self-reported providing low-quality data, 15 who reported technical difficulties preventing them from viewing any of the three videos included in the study, 13 who failed an attention check (within a scale, an item asked participants to select 'strongly disagree'), and three whose written responses indicated low English comprehension; some people failed multiple checks, leaving 362 (age M = 44.4, 49% female, 49% male (1 with trans experience), 1% nonbinary; 1 agender, 2 missing). Sensitivity analyses revealed that the per-condition sample size (about 43 people) provided 80% power to detect medium-sized effects or larger between any two conditions (d = .53).

4.4.1.2 Procedure.

The cover story for Study 10 was that participants would watch three short videos and answer questions about each. The first part of the study was identical to Study 9: comprehension check, demographics, polarization manipulation video. In line with the cover story, participants also reported their impressions of the video before proceeding with the survey. To avoid experimenter demand, I removed one brief section from the low polarization video that showed evidence that partisans want allies to be respectful and tolerant of opponents (this seemed too close to telling participants that their allies preferred engaging over dismissing—the key dependent measure in Study 10).

Participants then watched a second video which served as a manipulation check (for the exact video, see link). The video showed three Democrats and three Republicans gathering to discuss politics and ended before they interacted. Participants predicted how positively the interaction would go as well as how much the Americans in the video would "disagree with each other about politics" and "dislike people with whom they disagree about politics" (both items 1 = Not at all, 10 = Very much); the manipulation check was a composite of both items.

Next, participants watched the third and final video, which was Study 8's text messaging paradigm; I included only the conditions where the actor engaged with or dismissed shared opponents' views (that is, I excluded the control condition). As in Study 8, they imagined either participating in or observing the conversation, and reported the likelihood that they (or the other texter they were observing) would accept the (engaging or dismissing) actor's invitation to the bookstore.

Finally, participants completed the *perceived illegitimacy* and *political homogeneity* measures from Study 9, reported whether they watched all three videos without technical difficulties, and read a debrief.

4.4.2 Results

4.4.2.1 Manipulation check

As in Study 9, I used a linear model to predict manipulation check ratings from polarization video condition (low = 1, high = 0). Again, participants who saw the low polarization video perceived less dislike and disagreement than participants who saw the high polarization video, b = -0.25, SE = 0.08, $t(360)$ = -3.02, p = .003. This effect size was notably smaller (d = 0.32) than it was for Study 9's manipulation check (d = 0.88), perhaps because in Study 9 I asked participants to guess dislike and disagreement among Americans in general, about whom the video provided data, whereas in Study 10 I asked them about a specific group of individuals in a different video. In addition, contrasting with Study 9's findings, the video that participants watched did not affect their perceptions of opponents' views as illegitimate, b = -0.14, SE = 0.11, $t(360)$ = -1.25, p = .211, d = -0.13. This effect was in the same direction and reasonably similar sized to Study 9's significant effect (d = -0.21); it is possible Study 10's smaller sample size was insufficient to detect it.

4.4.2.2 Main analyses

To test my central prediction, a linear model predicted participants' responses from which video they watched (high polarization = -1, low = 1), actor behavior (dismiss = -1, engage = 1), reactor (perceived ally = -1, self = 1), and their interactions; see Table 4.4 and Figure 4.6. None of the effects involving the video manipulation were significant, ps > .518, suggesting perceived polarization did not play a role in participants' own attitudes toward engaging nor their perceptions of their allies' attitudes.

The only significant effect in the entire model was the interaction between actor behavior and reactor, which provided a third replication of the existence of misperceptions; see Table 4.4.

Participants reported that they themselves preferred an engaging ally, b = 3.86, SE = 1.59, t(354) = 2.42, p = .016, but that their ingroup would instead prefer an dismissing ally, b = -3.49, SE = 1.56, t(354) = -2.23, p = .026.

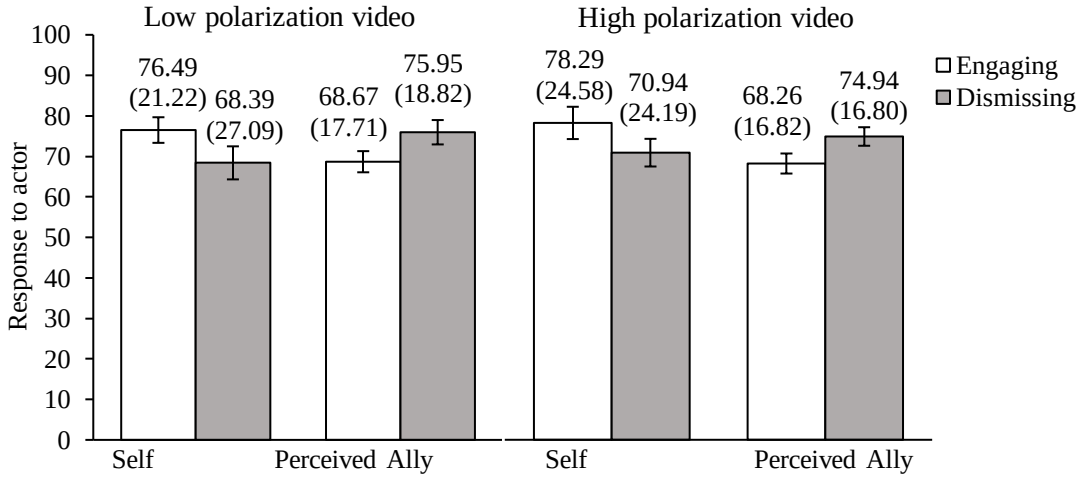


Figure 4.6. Own vs. perceived ally's attitudes toward engaging, dismissing actors broken down by high versus low polarization video condition, Study 8

Table 4.4

Regression results, Study 10

Effect	*b (SE)*	95% CI	*t*(354)	*p*
Intercept	72.74 (1.56)	70.55, 74.94	65.19	< .001
Polarization video (low = 1, high = -1)	-0.37 (1.56)	-2.56, 1.83	-0.32	.742
Reactor (self = 1, fellow ally = -1)	0.78 (1.56)	-1.41, 2.98	0.70	.482
Actor behavior (engaging = 1, dismissing = -1)	0.19 (1.56)	-2.01, 2.38	0.17	.867
Perceived polarization × reactor	-0.72 (1.56)	-2.92, 1.47	-0.65	.518
Perceived polarization × actor behavior	0.02 (1.56)	-2.17, 2.21	0.02	.986
Reactor × actor behavior	3.68 (1.56)	1.48, 5.87	3.29	.001
3-way interaction	0.17 (1.56)	-2.03, 2.36	0.15	.880

Note. *b* represents the beta for the difference between the two targets. SE represents standard error.

4.4.2.3 Exploratory individual difference moderators

Setting aside the manipulation variable, I conducted exploratory analyses to see if perceived illegitimacy and/or political homogeneity moderated misperceptions. That is, I ran two additional models similar to the one described above, but each one substituted one of the two (centered) individual difference variables in place of the manipulation variable. If some types of individuals are most prone to misperceptions, this could hint at a psychological source of misperceptions and suggests the need for tailored interventions.

In the model with perceived illegitimacy, the interaction between actor behavior and reactor was still (marginally) present, $b = 5.46$, $SE = 3.03$, $t(354) = 1.81$, $p = .072$. It was not moderated by perceived illegitimacy, $b = -0.68$, $SE = 1.01$, $t(354) = -0.67$, $p = .503$, but the main effect of the actor's behavior was, $b = -3.18$, $SE = 1.01$, $t(354) = -3.14$, $p = .002$. That is, regardless of whether people were reporting about themselves or their fellow allies, those who perceived opponents' views as more illegitimate reported a lower likelihood of agreeing to hang out with an ally who engages, $b = -3.58$, $SE = 1.46$, $t(354) = -2.45$, $p = .015$, but a higher likelihood of agreeing to hang out with one who dismisses, $b = 2.79$, $SE = 1.41$, $t(354) = 1.98$, $p = .048$. This faintly echoes Study 2, where people felt less favorably toward engaging with extreme (i.e., less legitimate) views.

In the model with political homogeneity in place of perceived illegitimacy, the interaction between actor behavior and reactor was also still (marginally) present, $b = 7.45$, $SE = 3.82$, $t(354) = 1.95$, $p = .052$. It was not moderated by political homogeneity, $b = -1.17$, $SE = 1.11$, $t(354) = -1.05$, $p = .292$, whose only effect was a main effect, $b = 2.39$, $SE = 1.11$, $t(354) = 2.16$, $p = .032$: People in more politically homogeneous social networks tended to report greater likelihoods of agreeing to hang out with the (allied) actor.

In sum, neither of these individual differences were related to perceived attitudes, so these results do not speak to other potential causes of misperceptions nor to the possibility of tailored interventions.

4.4.3 Discussion

Study 10 found no evidence that inducing people to perceive lower polarization helps correct their misperceptions about the reputational benefits of engaging. Regardless of whether they perceived high or low polarization, people preferred engaging and yet expected allies to

prefer dismissing. Study 10 thus replicated the key finding from Studies 7 and 8, further cementing the idea that people underestimate their allies' preference for engaging. But it provided no evidence about what causes this misperception.

In deciding how to interpret the non-significant effects of the polarization manipulation, it is worth considering some of Study 10's limitations: Power analyses reported above revealed that it was only powered to detect medium-sized differences between conditions, and the manipulation check—which would presumably show a larger effect than one could expect on a downstream DV—showed a small-to-medium effect. It is therefore possible that a higher-powered study could have revealed something different. That said, the predicted three-way interaction was not close to significant (p = .880), so the sample size would have had to have been much bigger, or the manipulation much stronger. The latter would likely be difficult: Manipulating perceived polarization is an uphill battle, as one must overcome people's lived experience of hearing more and more about perceived polarization in recent decades and years.

Together with Study 9, Study 10 sheds doubt on my proposed theorizing. I suggested that perceived polarization might shape how people expect their allies to respond to engaging with opposing views and, as a result, people's willingness to engage with such views. Instead, people induced to perceive low polarization were no more willing to engage with opposing views (Study 9) and no more likely to misperceive their allies' preferences for engaging (Study 10).

4.5 General discussion

4.5.1 Misperceiving allies' preferences

In Chapter 4, I proposed two hypotheses. First, I posited that people misperceive their allies' preferences for engaging over dismissing. Studies 7 and 8 tested this (as did Study 10, secondarily). This hypothesis received strong support in all tests. People preferred allies who

engaged with opponents' views over those who dismissively avoided them yet expected allies to either prefer the opposite (as in Studies 8 and 10) or have no preference (as in Study 7). Perhaps these studies' slightly different patterns reflect their different target viewpoints, such that people anticipated that allies are more accepting of engaging with an opponents' views on specific issues than their views broadly (as noted in Chapter 2's general discussion, observers indeed show this pattern). Differing patterns notwithstanding, these results are consistent with one piece of a pluralistic ignorance model (Prentice & Miller, 1996), by which most people feel one way but systematically misperceive others' feelings.

This conclusion may warrant further scrutiny, as Studies 7, 8, and 10 used convenience samples (Prolific Academic), rather than representative samples. Because of this, it is possible that my findings do not reflect a misperception: Perhaps Prolific workers are correct to say that *on average* their allies do not prefer engaging, even though their allies *who are also Prolific workers* do. That said, Prolific Academic samples are on the whole quite similar to the broader population's (Prolific Team, 2022), and respond similarly to political questions (Mullinix et al., 2015). I therefore deem it unlikely that the broader population in fact has no true preference for engaging, and tentatively conclude that what I observed in Chapter 4 is in fact evidence of a misperception.

This extends existing theory in several ways. First, this work connects to recent evidence on abundant political misperceptions, which for the most part has focused on people's perceptions of their opponents' attitudes (Lees & Cikara, 2020; Mernyk et al., 2021). The present work joins the few studies that have examined misperceptions of *allies'* attitudes (e.g., Dias et al., 2022; Lees & Cikara, 2020). Past work has likely focused on perceptions of opponents' attitudes because people misperceive their opponents' attitudes more than their allies (see

Westfall et al., 2015), perhaps because they have more direct knowledge of allies than of opponents (Huber & Malhotra, 2017; Mutz, 2006). But as political divides have grown, people have come to hold exaggerated stereotypes of even their allies, misperceiving their policy preferences and traits (Ahler & Sood, 2018; Kulibert et al., 2021; Levendusky & Malhotra, 2016b). Studies 7, 8, and 10 add to this by showing that people stereotype allies as having no preference or as preferring dismissing. And yet, Chapter 2's internal meta-analysis shows that even the most extreme partisans do not prefer dismissing. This fits with work suggesting that people perceive their allies as more extreme than they truly are (Kulibert et al., 2021), and specifically suggests that people's stereotypes of their allies' preferences go beyond even the most extreme of their real allies. This happened despite my taking precautions to avoid it in Study 8, by having the texting video specify to participants that the other texter shares their ideology and extremity. Nonetheless, due to this stereotyping, participants may have inferred that the actor was more extreme, which could driven these misperceptions.

4.5.2 Engaging behavior and the role of perceived polarization

Beyond the existence of a misperception, to more fully support a pluralistic ignorance account, I would need evidence that this misperception shapes engaging behavior. This was baked into my hypothesis for Study 9: That perceiving high polarization causes the misperception, which in turn inhibits engaging. I found no evidence for this overall hypothesis, and Study 10 in particular raised doubts about polarization causing misperceptions. (Neither study offered even a correlational test of whether people's perceptions of their allies' preferences are related to their own interparty behavior, but this idea has a strong precedent; Ajzen, 1991; Moore et al., 2021; Sherif & Sherif, 1953; Wojcieszak et al., 2020.)

The consistently null results I obtained from successfully manipulating perceived polarization suggest the need for theoretical refinement. One possibility is that it is important to consider polarization through both int*er*group and int*ra*group lenses. My theorizing had relied on the former: I reasoned that perceived polarization draws people's attention to divides *between* groups, such that members of one group expect ingroup backlash for engaging with the views of a more strongly disliked and disagreed with outgroup. Accordingly, my manipulation videos emphasized the degree of conflict *between* groups.

But perceived polarization might instead be important because it shapes the degree to which people feel their *own* group has become more extreme. In fact, decades ago, social psychological accounts of group polarization revolved around what happens when a single group becomes more extreme on some (non-political) dimension (Myers & Lamm, 1976); some more recent efforts in the context of politics have returned to the idea that sometimes a growing divide between groups is better understood as one or both groups radicalizing independently (see Jost et al., 2022).

In other words, perceived polarization could still be a key variable, if thought of as perceptions of one's *own* group's extremity. For one thing, people do clearly perceive (exaggeratedly) high polarization in this sense: Almost half of all Democrats (47%) and Republicans (45%) say most members of even their own party have become too extreme (Pew Research Center, 2019b). These perceptions are likely overblown: People overestimate how extreme their allies' policy positions are (Westfall et al., 2015; Levendusky & Malhotra, 2016b), and what stereotypical traits they embody (Ahler & Sood, 2018; Kulibert et al., 2021).

For another, this type of perceived polarization could shape both what people think their allies will prefer, and consequently their own behavioral choices. Extreme allies seem more

loyal, principled, close-minded, and unlikely to defect from group interests (Kulibert et al., 2021; Pew Research Center, 2019b)—all attributes that imply an unwillingness to engage with opposing views. Since people perceive many or most of their allies as extreme, they may infer that most of their allies would disapprove of engaging. In turn, this anticipated disapproval could cause people to inhibit any impulses they have toward engaging.

Indeed, at least one paper indeed finds that people engage with other views when political allies seem to endorse this behavior (Wojcieszak et al., 2020), though that paper did not provide accurate normative information. To manipulate perceived extremity of allies using accurate information, a video could visually convey evidence that only a minority of inpartisans dislike opponents and are politically engaged (Druckman et al., 2022; Klar et al., 2018) and that most people disidentify with their own party when copartisans behave uncivilly (Druckman et al., 2019). Future research might test this.

4.5.3 Generalizability across cultures and operationalizations

The generalizability of Chapter 4's studies are worth remarking on. Of course, Studies 7, 8, and 10 in part studied people's own responses to seekers, just as in Chapter 2, so the same generalizability points from that chapter apply here: People's own responses should better extend to places that share contemporary U.S. citizens' pro-democratic values and moral qualms with political opponents. As for *perceived* responses, my studies did not identify what causes people to perceive worse responses to engaging (versus dismissing), but if I am right that they stem from perceived intragroup polarization (i.e., perceiving allies as extremists), then generalizability depends on the extent to which perceived intragroup polarization owes to external, culturally variable sources (i.e., news and media coverage of allies) or internal, general processes (i.e., the tendency for people to see themselves as more moderate than a prototypically extreme ingroup

member). For countries where misperceptions exist between people's own preference for engaging and their perceptions of others' preferences, one might wish to intervene. Of course, my experimental videos in Studies 9 and 10 did not change perceptions or behavior, but would similar manipulations have worked better elsewhere? Perhaps so, in contexts where people have loosely-formed ideas about how polarized their compatriots are—Americans have heard much about this for the last decade or longer (Levendusky & Malhotra, 2016), so it may be harder to change their perceptions and downstream behavior.

Chapter 5: General Discussion

This dissertation investigated the actual and expected reputational benefits and costs that partisans face for engaging constructively with opponents' views, compared to dismissing them.

Chapter 2 examined why engaging usually reaps greater social rewards. This is because people who constructively engage with opponents seem admirably tolerant, rational, and cooperative, even though they also seem alarmingly willing to validate and potentially adopt opponents' views. This preference is robust but shrinks when the alarm bells get louder: When there are greater differences between observers' attitudes and the target views (i.e., when extremists observe an ally engaging, or when someone engages with a view that most observers find to be less legitimate).

Chapter 3 examined preferences in the real-world context of U.S. Senators' tweets. In doing so, it identified one case in which dismissing garners more social benefits: Senators' tweets receive more Likes and Retweets when they dismiss, rather than engage with, opponent's views. This discrepancy with Chapter 2's findings owes not to differences in who is engaging (Senators vs. citizens) nor the context in which engaging occurs (on Twitter vs. offline); rather, it occurs because people who frequently react to Senators' tweets—unlike everyone else—prefer allies who dismiss opponents' views. Moreover, although everyone else approves more of engaging, they nonetheless intend to Like and Retweet dismissing tweets more often than engaging ones. As a result, Like and Retweet patterns for Senators' tweets fail to represent most people's preferences.

Building on this, Chapter 4 tested whether people fail to realize that their allies prefer engaging and if so, how to eliminate this misperception. Across multiple studies I find that

people do indeed underestimate others' preferences for engaging over dismissing. I proposed that these misperceptions owe to perceived polarization, but I found no support for this idea.

These studies used different operationalizations of engaging and dismissing, thereby better representing the breadth of ways these core constructs can emerge. I did not directly compare how this influenced my results, but the broadest pattern—that most people typically prefer engaging—emerged across operationalizations, boosting confidence in this finding.

5.1 Implications

5.1.1 Benefits and costs of engaging with opposing perspectives

The present work builds on existing literatures on the reputational effects of engaging with opposing political perspectives, which are important to consider alongside other individual and societal effects of this behavior. Below, I consider how my findings fit with prior work on engaging's reputational and non-reputational effects.

5.1.1.1 Reputational benefits and costs for individuals

My findings in Chapter 2 and 3 reconcile seemingly conflicting findings in past work, some of which suggest people would gain more reputationally from engaging while others suggest the opposite. For example, studies find that people like allies who show respect and civility toward opponents (Druckman et al., 2019; Frimer & Skitka, 2018), yet dislike allied politicians who compromise with opponents (Ryan, 2017) as well as allied laypersons who empathize with immoral opponents (Wang & Todd, 2020). Dovetailing with those findings, mine here highlight some reputational costs to engaging—such behavior makes people seem willing to violate moral concerns by legitimizing and maybe even adopting opposing views—but suggest that on the whole, these costs are outweighed by the reputational benefits of upholding

democratic values of tolerance and rationality, even among audiences (e.g., extremist observers) and in cases (e.g., engaging with extreme views) in which these costs are more salient.

My findings in Chapter 3 resolve a different but related paradox in the literature and, in doing so, outline additional conditions under which engaging might have reputational costs. Recent literature highlights a paradox whereby social media rewards things—negativity, incivility, outgroup derogation—which people do not approve of (Brady et al., 2017; Frimer et al., 2018, 2020, 2022; Rathje et al., 2021, 2022; Schöne et al., 2021). Study 5 reconciles this, suggesting that social media patterns reflect the genuine but atypical preferences of a small group of active users, as well as reflecting most people's willingness to Like and Retweet posts that they do not approve of. Concretely, those results suggest that Senators aiming to boost their reputation on Twitter might be better off dismissing opposing views, but that most people nonetheless disapprove of this behavior and do not reward Senators for it.

5.1.1.2 Non-reputational benefits and costs for individuals

Of course, dismissing's occasional reputational benefits should be considered alongside its other individual- and societal-level effects. Prior work speaks to many of the individual-level effects—psychological costs and benefits—of engaging with opposing political perspectives, which are worth considering given recent calls for citizens to engage with opposing views. For example, existing evidence reveals intrapersonal benefits of engaging: When people try to understand opponents, they develop more tolerance for these opponents (Mutz, 2002; Stanley et al., 2020), which may help people avoid feeling angry and upset in the long-term. Likewise, engaging can help people fulfill their motivations to understand the world more accurately (Golman et al., 2017; Hart et al., 2009; Sharot & Sunstein, 2020). At the same time, prior work notes intrapersonal costs of engaging in the sense that people who engage with opponents' views

may hear offensive ideas that damage their short-term emotional wellbeing (Dorison et al., 2019; Frimer et al., 2017; Hart et al., 2009). Thus, practically speaking, individuals who are deciding whether to heed calls to engage with opposing views can consider the costs and benefits outlined in this and other work.

5.1.1.3 Benefits and costs for societies.

Institutions and leaders calling on citizens to engage with opposing views can consider these effects on individual citizens, along with societal-level effects when many citizens heed this call. For instance, the effects of engaging on individuals—increased tolerance and familiarity with alternative policy positions—when scaled to the masses would bode well for democracies, facilitating the sort of deliberation and discourse required for these democracies to thrive (McCoy et al., 2018; Mutz, 2002). At the same time, dismissing might be preferable for societies looking to make quick social or moral progress: By publicly disregarding an opposing view, one can catalyze social action by strongly signaling one's allegiance amid political conflict (Descioli & Kurzban, 2013; Spring et al., 2018; Van Zomeren et al., 2011). The democratic route for broader social change, by which proponents of a cause must bring others on board through persuasion or compromise, may be more difficult, slower, and unacceptable to proponents with strong moral convictions.

5.1.2 Responses to engaging on social media inform interactionist theories and perceived norms.

Although I initially set out to study reputational effects of engaging *in general*, in Chapter 3 I examined this question in a specific context and ended up finding that individual differences play an important role in explaining patterns of behavior in this context. In doing so, I can speak

to larger debates about the role of persons and situations in social media behavior (e.g., Lewin, 1951; Mischel & Shoda, 1995).

Social media has only been around for slightly more than a decade. While we still have much to learn about how human psychology responds to this new context, evidence reviewed above paints social media as a sinkhole of negativity. Early theorizing explained this by pointing to how social media is a unique context (e.g., Crockett, 2017), with features such as anonymity and psychological distance that may facilitate negative social interactions. Some recent work has pushed back on this, showing that the negativity on social media is not a main effect across people (which would support the unique context account) but rather comes from a subset of individuals who are hostile on- and offline (Bor & Petersen, 2022; Kumar et al., 2023; Mukerjee et al., 2022). My results align more with effects being driven by persons, as it suggests that some individuals are responsible for at least promoting and amplifying interparty division on social media. My data cannot directly speak to whether frequent reactors would also behave similarly offline, but I observed that frequent reactors are more extreme; since extremists tend to be more politically active on- and offline (Birkhead & Hershey, 2019), they might look for and select into situations in which they can express their strongly held political beliefs.

Chapter 3's finding that the patterns seen on Twitter are unrepresentative also has practical implications for the norms people infer. Researchers and journalists might think twice about using Twitter as a barometer of public opinion: Findings based on Twitter users may not generalize, even when the finding itself (i.e., the popularity of online negativity) has been shown to be highly robust. Crowdsourced participants' feelings about political stimuli may better mirror the general population's (Mullinix et al., 2015). Still, Twitter behaviors are worth studying in their own right (e.g., Warzel, 2020): Because politicians are prototypical group members, their

partisan followers might infer norms from their tweets, gleaning that dismissing is appropriate behavior. Even those citizens who do not use Twitter might still hear about social media on the news (McGregor, 2019), and so may infer their peers' attitudes from reports about popular tweets. They may incorrectly assume that most of their allies approve of dismissing opponents' views on Twitter. They might even internalize, or at least outwardly conform to, this perceived social pressure (Prentice & Miller, 1996). This idea formed the basis for Chapter 4, and Studies 7, 8, and 10 showed misperceptions consistent with this idea. That said, these studies showed only that misperceptions exist. They do not tell us whether these misperceptions come from (social) media or elsewhere—an open question that I return to later in this discussion.

5.2 Unanswered questions

This dissertation has helped to answer some questions but left others unanswered and unveiled further new questions. I consider these below.

5.2.1 How people can like those who engage yet frequently choose to dismiss

Given that people prefer allies who engage constructively with opponents and their views, why do people more often dismissively avoid those views (e.g., Frimer et al., 2017; Dorison et al., 2019; Iyengar & Hahn, 2009)? Such behavior is not only inconsistent with what they want others to do, but also their own tolerant, open-minded values (Brown, 2006; Fiske et al., 2006). Nonetheless, people might feel free to contravene their values for at least three reasons (see Guan et al., 2023). For one, they might not notice it happening: People might passively drift *towards* congenial ideas rather than actively averting their eyes from dissident ones (e.g., Motyl et al., 2014), with the unintentional effect that they fail to hear opposing views. For another, people may notice their close-minded behavior and its inconsistency with their values but justify it by blaming the communication breakdown on opponents (e.g., Yang et al.,

2016), citing that they are close-minded (Iyengar et al., 2019) and unwilling to listen (Yeomans et al., 2020), or that they already know what opponents would say. For yet another, even if people notice their behavior and do not blame opponents for initiating discussions, they could still feel opponents are immoral so dismissing them is the right thing to do (Cole Wright et al., 2008).

The explanations above deal with how people reconcile their democratic values with their reluctance to try to understand opponents, but another way of answering why people refrain from engaging with opposing views is to consider what drives people's behavior more broadly, beyond their internal values. The theory of planned behavior (Ajzen, 1991) suggests that people plan to behave in ways that they endorse, but also in ways that believe they can control and that their peers approve of. Existing work has not examined whether people feel they can control their engaging behavior, but I speculate they do: At least, people can easily choose to engage impersonally by reading about opposing views online or listening to opponents' recorded speeches (e.g., Minson et al., 2020), though they might feel less easily able to find and directly talk and listen to opponents.

Rather, evidence from Chapter 4 suggests that perceived peer support is a likely impediment to engaging, as people seem to routinely underestimate their peers' approval of engaging. By suggesting that misperceptions of allies' attitudes affect behavior, I break from recent efforts to study misperceptions of opponents (e.g., Lees & Cikara, 2020; Mernyk et al., 2021; Moore-Berg et al., 2020), and instead draw on theorizing that suggests ingroup attitudes shape intergroup behaviors (Crandall et al., 2002; Sherif & Sherif, 1953; Vial et al., 2019), as well as evidence in the political domain specifically that allies' attitudes shape interparty contact (Moore et al., 2021; Wojcieszak et al., 2020). That said, I only observed misperceptions, not their

influence on behavior—future work should test whether correcting these misperceptions can encourage interparty contact, as those theories and findings suggest. This is practically important given contact could mitigate political polarization (Guan et al., forthcoming; Combs et al., 2023), which itself impedes progress on various other societal issues (Heltzel & Laurin, 2020).

This social explanation for why people fail to engage with opposing views contrasts existing purely intrapsychic explanations and joins others that emphasize external factors. Intrapsychic explanations suggest that a person might have a close-minded disposition (e.g., Webster & Kruglanski, 1994), or wish to avoid potential mental and emotional burden of hearing such views (Dorison et al., 2019; Frimer et al., 2017; Minson & Dorison, 2022). External explanations have typically emphasized the actor-opponent dyad. For example, they suggest a person may refrain from engaging with an opponents' views because they want to prevent seemingly inevitable conflict with an opponent (e.g., Frimer et al., 2017), or because they are matching what they assume the opponent would do (Minson & Chen, 2022).

5.2.2 Why perceived polarization did not influence engaging behavior

In addition, there are also insights to be gleaned from perceived polarization failing to affect engaging behavior in Study 9. I predicted that perceived polarization would change this behavior by changing misperceptions, but existing theories offer additional mechanisms through which in particular the low polarization videos should have increased engaging (see Minson & Dorison, 2022). This makes it even more remarkable that it did not.

For one, classic research suggests that people engage with information when they are curious and motivated to learn more, usually because they think the information will be useful (Golman et al., 2017; Hart et al., 2009; Minson et al., 2020; Sears & Freedman, 1968). Since people already perceive high polarization (Pew Research Center, 2020b), the low polarization

video should have undermined their confidence in their knowledge of political divides and where each party stands, motivating them to learn more about everyone's views but especially opponents', which they are likely less familiar with. But it did not.

For another, cognitive dissonance theory suggests people refrain from engaging with opposing views because hearing valid arguments against one's own stance on identity-relevant issues elicits negative feelings of dissonance, anxiety, and discomfort (Dorison et al., 2019; Festinger, 1957; Frimer et al., 2017). The low polarization video emphasized Democrats' and Republicans' surprisingly high agreement on many issues, so participants would presumably expect to find more agreement with opponents' views. Again, this should disinhibit engaging, but it did not.

For yet another related reason, naïve realism theory suggests people perceive their own beliefs to be objectively true such that anyone who believes otherwise is factually misled or morally corrupt (Minson & Dorison, 2022). But the low polarization video addressed this too: This video led people to see opponents as less irrational and immoral. Despite this, they were no more willing to engage with those opponents' views.

Taken together, the low polarization video had several features that presumably should have motivated people to engage with opposing views, even beyond correcting misperceptions. And yet, it did not. We can be confident in the foundational research programs reviewed above (e.g., on selective exposure and avoidance), so this suggests a need to consider other possible explanations for the null effect.

5.2.2.1 Failure to address active mechanisms.

Aside from measuring perceptions of opponents' immorality and irrationality, I did not directly measure the other potential mechanisms mentioned above: people's uncertainty of their

political knowledge or motivations to learn about other political views, or whether they expected to feel less upset upon hearing opposing views. Perhaps my video did not affect these theorized mechanisms; some prior work has precisely manipulated these mechanisms and observed higher rates of engaging (e.g., Dorison et al., 2019), and meta-analyses have supported this and other proposed mechanisms (Hart et al., 2009), so we can still feel confident in them.

5.2.2.2 Possible issues with manipulation or dependent measure

Methodological features could also contribute to this null result. For example, there may have been issues with the manipulation, such that its effects owed to demand characteristics: Perhaps participants guessed the proximal purpose of the video and responded accordingly to the check item, but not to the dependent measure (because it came later and was not so obviously related to the content of the video). Study 9's between-subjects design should have helped disguise the manipulation, but participants likely already perceived high polarization (Pew Research Center, 2020b), and the low polarization video explicitly discredited this assumption by blaming it on the media. As a result, participants probably realized that the low polarization video aimed to change these perceptions and so responded accordingly. Alternatively, the manipulation might have had a real effect on engaging, but this effect may have been too small to affect this downstream dependent measure. Or perhaps the measure of engaging itself was to blame—there were conceptual advantages to this novel measure, but it showed only modest relationships with expected constructs, leaving open the possibility that it poorly approximated engaging behavior.

5.2.3 How could interventions increase engagement with opposing views?

In Chapter 4, I failed to support my hypothesis that reducing perceived polarization disinhibits engaging. I put this idea forth with hopes of designing a potential intervention. What

do the present results say about what a successful intervention might look like? Earlier, I suggested that a successful intervention might manipulate perceived intragroup, rather than intergroup, polarization, which is in line with some existing theory and evidence. Aside from this, I offer a few other possibilities below.

For one, researchers could directly correct misperceptions of engaging's desirability and see if that influences behavior. For example, participants could report their preferences for engaging and estimate their peers', then learn correct information about their peers' preferences. Of course, if there was an external factor causing these misperceptions (e.g., social or news media), it might be more effective to intervene on that cause directly rather than addressing its downstream effects.

For another, pluralistic ignorance theorizing suggests that people assume others' attitudes from their behaviors, so a potential intervention could emphasize that most people try to engage with opposing views. For example, I could tell participants that people try to seek a balanced viewpoint diet and support this by citing the results of Study 9, whereby people sought opposing views only slightly less often (about 45% of the time) than congenial ones (55%). I could supplement this by emphasizing that other intolerant behaviors are absolutely rare and support this by citing evidence that people overestimate the prevalence of canceling and violence (Dias et al., 2022; Mernyk et al., 2022), both of which imply intense opposition to hearing other views. Pluralistic ignorance theorizing suggests that by correcting perceptions of how prevalent these behaviors are, people will infer less peer support for them and conform more to these norms.

Another solution involves identifying and intervening on the cause(s) of the misperceptions documented in Studies 7, 8, and 10, but exploratory analyses in those studies have produced no clear leads. For instance, Study 10 revealed that perceived illegitimacy was

related to responses, but not differently for actual versus anticipated responses, which would be necessary to explain divergence between the two. In pilot data not presented here, I observed null relationships such that consuming more media (social and other formats) was unrelated to misperceptions. Nonetheless, future research might consider individual differences or general psychological processes (e.g., egocentric anchoring and adjustment) that could cause these misperceptions. Namely, it could experimentally test which of the possible causes is most responsible for misperceptions and develop interventions to address this cause.

5.3 Generalizability

5.3.1 Generalizability of these specific results

In considering what we learn from these results, it is important to consider whether they generalize to the populations we care about, telling us what the broader population truly prefers. On one hand, by primarily recruiting convenience samples from Prolific Academic, my results would appear to have questionable generalizability, especially given that Prolific workers are more likely to be democrats and women than the population (Prolific Team, 2022). But in all studies, I used screens to recruit approximately equal portions of Democrats, Republicans, and Independents or other-affiliates, which is within 5% of their true population proportions (Pew Research Center, 2019c). Although not perfectly representative, these samples are still likely to generalize to the population, given that samples from crowdsourcing platforms often respond similarly to representative samples (Mullinix et al., 2015). Likewise, if my samples underrepresented a specific type of participant whose opinions differed meaningfully, this would be an issue. However, the only meaningful individual differences I have observed—extremity and frequently reacting to politicians' tweets—were represented at levels comparable to the population (in my samples, about 6-7% of participants are frequent reactors and 1/3 were

independents, which approximates the population prevalence, Pew Research Center, 2019c). In other words, these samples approximated the broader population on demographic factors known to influence the effects I studied.

It is also worth remarking on which broader population I have generalized to—contemporary Americans—and whether their psychological responses would generalize to people in other times and places. America provides a great case study of conflict between democratic and moral values because it is a democracy (albeit one in decline; EIU, 2016, 2021); countries with other sociopolitical systems might be less likely to encourage tolerance and rationality, so their citizens likely see less to like in engaging and may be more likely to resort to passive intolerance or active condemnation to handle morally infused political disagreement. America is hardly the only democracy, but it is more polarized than it has been in the past (for a review, see Heltzel & Laurin, 2020; Jost et al., 2022), and is perhaps more polarized than other countries (Boxell et al., 2022; Finkel et al., 2020; Fletcher et al., 2019), though some scholars disagree (Westwood et al., 2018; Gidron et al., 2019; Wagner, 2021).

If America is not uniquely polarized, people's responses to engaging versus dismissing should generalize to other countries with roughly similar degrees of democratic values and polarization. I did not identify the specific cause of misperceptions but if they indeed come from external sources like the media, those patterns should generalize best to cultural contexts with similar media dynamics. If instead America is uniquely polarized, the present results might not generalize to other cultures and/or times, but they would still provide a theoretically informative case study of how Democracies fare amid record-high levels of polarization: Few times in history (e.g., the U.S. civil war; the red scare) have citizens of a democracy grappled with such an explicit tradeoff between their commitment to democratic values of political tolerance and

their moral opposition to other political views. Practically, regardless of whether U.S. polarization is a special case or not, America is a global power and other countries pay attention to its sociopolitical situation, so its dynamics could influence people elsewhere.

5.3.1 Generalizability of these broader dynamics

Given that democracies are non-universal sociopolitical systems and political polarization varies greatly across cultures, do the present findings speak at all to more general human psychological processes? I believe they speak to the basic tendency for (perceived) social pressures to shape intergroup relations. Prior theorizing suggests individuals will engage with an outgroup to the extent that their ingroup supports such engagement (Sherif & Sherif, 1953), but that theorizing says little about what factors shape perceived ingroup support. One such factor is the tendency to assume one's ingroup values loyalty to and promotion of the ingroup over outgroups, so people may generally feel pressure to not engage constructively with outgroups. Two other factors come from each opposing pathway mechanism I proposed: Valuing tolerance generally and being concerned with validating a specific outgroup. That is, when a person's ingroup prescribes tolerance (e.g., in democracies and other egalitarian social arrangements), there will be more social pressure to engage with outgroups, and when their ingroup seems concerned with validating an outgroup, there will be less pressure to engage. In this way, I have proposed moderators that determine when social approval encourages vs. discourages people from bridging intergroup divides—moderators that may operate in other intergroup divides.

In addition, there are unique insights to be gained from studying engaging across moralized *political* divides, compared to other moralized group divides. For example, religious groups might conflict over perceived moral differences, but religious teachings rarely if ever provide a framework for how to peacefully coexist or compromise with followers of other

religions. Even in a democracy, religious groups are required to merely tolerate each other and peacefully coexist, but they are not compelled to work together or find grounds for compromise. Political groups can similarly be mired in moral conflict and in non-democratic countries, they might similarly lack a framework for handling their disagreements. But things change when political groups feud within democracies: Each political group separately agrees to superordinate values that help them to peacefully and cooperatively navigate intergroup conflict. This illustrates how democracies are a psychologically unique sociopolitical arrangement in that they can mitigate intergroup conflict—the source of much turmoil in human history. Democracies can continue to do so as long as citizens continue to champion democratic values, even in the face of moral divides; the present results suggest Americans have done precisely this, providing good news for fans of democracy.

Bibliography

Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *Journal of Politics, 70*(2), 542–555. https://doi.org/10.1017/S0022381608080493

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

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes, 50*(2), 179-211.

Anderson, S., & Cameron, C. D. (2023). How the self guides empathy choice. *Journal of Experimental Social Psychology, 106,* 104444.

Badaan, V., & Jost, J. T. (2020). Conceptual, empirical, and practical problems with the claim that intolerance, prejudice, and discrimination are equivalent on the political left and right. *Current Opinion in Behavioral Sciences, 34,* 229-238.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., … & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, 115,* 9216-9221.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. Com's Mechanical Turk. *Political analysis, 20,* 351-368.

Birkhead, N. A., & Hershey, M. R. (2019). Assessing the ideological extremism of American

party activists. *Party Politics*, *25*(4), 495–506.

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review*, *116*, 1-18.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2022). Cross-country trends in affective polarization. *Review of Economics and Statistics*, 1-60.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*, eabe5641.

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 of the United States of America*, 201618923. https://doi.org/10.1073/pnas.1618923114

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The Ideological-Conflict Hypothesis: Intolerance Among Both Liberals and Conservatives. *Current Directions in Psychological Science*, *23*(1), 27–34. https://doi.org/10.1177/0963721413510932

Brown, W. (2009). *Regulating aversion: Tolerance in the age of identity and empire*. Princeton University Press.

Casas, A., Menchen-Trevino, E., & Wojcieszak, M. (2022). Exposure to extremely partisan news from the other political side shows scarce boomerang effects. *Political Behavior*, 1-40.

Christiano, T. (2008). *The constitution of equality: Democratic authority and its limits*. Oxford University Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591-621.

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific reports*, *11*(1), 22083.

Cole Wright, J., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin*, *34*(11), 1461–1476. https://doi.org/10.1177/0146167208322557

Combs, A., Tierney, G., Guay, B., Merhout, F., Bail, C. A., Hillygus, D. S., & Volfovsky, A. (2023). Reducing political polarization in the United States with a mobile chat platform. *Nature Human Behaviour*, 1-8.

Crandall, C. S., Eshleman, A., & O'brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of personality and social psychology*, *82*(3), 359.

Crawford, J. T. (2014). Ideological symmetries and asymmetries in political intolerance and prejudice toward political activist groups. *Journal of Experimental Social Psychology*, *55*, 284–298. https://doi.org/10.1016/j.jesp.2014.08.002

Crawford, J. T., & Pilanski, J. M. (2014). Political Intolerance, Right and Left. *Political Psychology*, *35*(6), 841–851. https://doi.org/10.1111/j.1467-9221.2012.00926.x

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Cut. (2015a, October 27). *Democrats and Republicans Part 1 | Dirty Data - Ep 1 | Cut* [Video]. Youtube. https://youtu.be/qAQt0ICZAr8

Cut. (2015b, November 11). *Democrats and Republicans Part 2 | Dirty Data - Ep 2 | Cut* [Video]. Youtube. https://youtu.be/ANh0iZ6Fqew

Cut. (2015c, November 18). *Democrats and Republicans Part 3 | Dirty Data - Ep 3 | Cut* [Video]. Youtube. https://youtu.be/0Sn-r6YLVtg

Descioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*(2), 477–496. https://doi.org/10.1037/a0029065

Dias, N. C., Druckman, J. N., & Levendusky, M. (2022). How and Why Americans Misperceive the Prevalence of, and Motives Behind,"Cancel Culture". *Cancel Culture"(October 2, 2022)*.

Ditto, P. H., & Koleva, S. (2011). Moral Empathy Gaps and the American Culture War. In *Emotion Review* (Vol. 3, Issue 3, pp. 331–332). https://doi.org/10.1177/1754073911402393

Dorison, C. A., Minson, J. A., & Rogers, T. (2019). Selective exposure partly relies on faulty affective forecasts. *Cognition, 188*, 98-107.

Druckman, J. N., Gubitz, S. R., Levendusky, M. S., & Lloyd, A. M. (2019). How Incivility on Partisan Media (De)Polarizes the Electorate. *The Journal of Politics*, *81*(1), 291–295. https://doi.org/10.1086/699912

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis) estimating affective polarization. *The Journal of Politics*, *84*(2), 1106–1117.

Druckman, J. N., Kang, S., Chu, J., N. Stagnaro, M., Voelkel, J. G., Mernyk, J. S., ... & Willer, R. (2023). Correcting misperceptions of out-partisans decreases American legislators' support for undemocratic practices. *Proceedings of the National Academy of Sciences*, *120*(23), e2301836120.

Dunaway, J., & Graber, D. A. (2022). *Mass media and American politics*. Cq Press.

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes, 129*, 59-69.

Evans, A. M., Stavrova, O., Rosenbusch, H., & Brandt, M. J. (2023). Expressions of doubt in online news discussions. *Social Science Computer Review*, *41*(1), 163-180.

Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., & Kralj Novak, P. (2022). Retweet communities reveal the main sources of hate speech. *PloS one*, *17*(3), e0265602.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12), e12501.

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

Festinger, L. (1957). Social comparison theory. *Selective Exposure Theory*, *16*, 401.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Fiorina, M. P., & Abrams, S. J. (2008). Political Polarization in the American Public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2006). Universal dimensions of social cognition:

warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fletcher, R., Cornia, A., & Nielsen, R. K. (2020). How polarized are online and offline news audiences? A comparative analysis of twelve countries. *The International Journal of Press/Politics*, *25*(2), 169-195.

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2022). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, 19485506221083812.

Frimer, J. A., & Skitka, L. J. (2018). The Montagu principle: Incivility decreases politicians' public approval, even with their political base. *Journal of Personality and Social Psychology*, *115*(5), 845–866. https://doi.org/10.1037/pspi0000140

Frimer, J. A., & Skitka, L. J. (2020). Americans hold their political leaders to a higher discursive standard than rank-and-file allies. *Journal of Experimental Social Psychology*, *86*(November 2019), 103907. https://doi.org/10.1016/j.jesp.2019.103907

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*(April), 1–12. https://doi.org/10.1016/j.jesp.2017.04.003

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group Processes and Intergroup Relations*, *8*(2 SPEC. ISS.), 109–124. https://doi.org/10.1177/1368430205051060

Garrett, K. N., & Bankert, A. (2020). The moral roots of partisan division: How moral conviction heightens affective polarization. *British Journal of Political Science*, *50*(2), 621-640.

Gidron, N., Adams, J., & Horne, W. (2020). *American affective polarization in comparative perspective*. Cambridge University Press.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information Avoidance. *Journal of Economic Literature*, *55*(1), 96–135. https://doi.org/10.1257/jel.20151245

Goodwin, G. P. (2018). The objectivity of moral beliefs. *Atlas of moral psychology*, 310-319.

Goodwin, G. P., & Darley, J. M. (2008). *The psychology of meta-ethics: Exploring objectivism*. *106*, 1339–1366. https://doi.org/10.1016/j.cognition.2007.06.007

Guan, K., Heltzel, G., & Laurin, K. (forthcoming). Moral Dimensions of Political Attitudes and Behavior. In P. A. Robbins & B. F. Malle (Eds.), *Cambridge Handbook of Moral Psychology*. Cambridge: Cambridge University Press.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29-48.

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

Haidt, J., Rosenberg, E., & Hom, H. (2003). Differentiating Diversities: Moral Diversity Is Not Like Other Kinds1. *Journal of Applied Social Psychology*, *33*(1), 1–36. https://doi.org/10.1111/j.1559-1816.2003.tb02071.x

Han, J., & Yzer, M. (2020). Media-induced misperception further divides public opinion: A test of self-categorization theory of attitude polarization. *Journal of Media Psychology: Theories, Methods, and Applications*, *32*(2), 70.

Harbridge, L., & Malhotra, N. (2011). Electoral incentives and partisan conflict in Congress: Evidence from survey experiments. *American Journal of Political Science*, *55*(3), 494-510.

Harel, T. O., Maoz, I., & Halperin, E. (2020). A conflict within a conflict: intragroup ideological

polarization and intergroup intractable conflict. *Current Opinion in Behavioral Sciences*, *34*, 52–57.

Harris, E. A., & Van Bavel, J. J. (2021). Preregistered Replication of "Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority". *Psychological Science*, *32*(3), 451-458.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Hatemi, P. K., Crabtree, C., & Smith, K. B. (2019). Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, *63*(4), 788-806.

Heltzel, G., & Laurin, K. (2020). Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/0375-9474(93)90395-E

Heltzel, G., & Laurin, K. (2021). Seek and Ye Shall Be Fine: Attitudes Toward Political-Perspective Seekers. *Psychological Science*, *32*(11), 1782–1800.

Hetherington, M. J. (2001). Resurgent mass partisanship: The role of elite polarization. *American political science review*, *95*(3), 619-631.

Himelboim, I., Smith, M., & Shneiderman, B. (2013). Tweeting apart: Applying network analysis to detect selective exposure clusters in Twitter. *Communication methods and measures*, *7*(3-4), 195-223.

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *Journal of Politics*, *79*(1), 269–283. https://doi.org/10.1086/687533

Huffpost. (2015, October 27). *Yes / No: Democrats and Republicans* [Video]. Youtube. https://youtu.be/tOUNMmdk7W0

Hussein, M. A., & Wheeler, S. C. (2023). Reputational Costs of Receptiveness: When and Why Being Receptive to Opposing Political Views Backfires.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of personality and social psychology*, *100*(4), 719.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, *59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jacoby, J., & Sassenberg, K. (2011). Interactions do not only tell us when, but can also tell us how: Testing process hypotheses by interaction. *European Journal of Social Psychology*, *41*(2), 180-190.

Joseph, K., Shugars, S., Gallagher, R., Green, J., Mathé, A. Q., An, Z., & Lazer, D. (2021). (Mis) alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. arXiv preprint arXiv:2109.01762.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political psychology*, *38*(2), 167-208.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560-576.

Kalmoe, N. P., & Mason, L. (2022). *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. University of Chicago Press.

Kearney, M. W. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of open source software*, *4*(42), 1829.

Kelman, H. C. (1958). Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, *2*(1), 51-60.

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, 922-946.

Kingzette, J. (2021). Who do you loathe? Feelings toward politicians vs. ordinary people in the opposing party. *Journal of Experimental Political Science*, *8*(1), 75-84.

Klar, S., Krupnikov, Y., & Ryan, J. B. (2018). Affective polarization or partisan disdain? Untangling a dislike for the opposing party from a dislike of partisanship. *Public Opinion Quarterly*, *82*(2), 379–390. https://doi.org/10.1093/poq/nfy014

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality*, *46*(2), 184–194. https://doi.org/10.1016/j.jrp.2012.01.006

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.

Kulibert, D., Moss, A. J., Appleby, J., & O'Brien, L. (2021). *Perceptions of Political Deviants: A Lay Theory of Subjective Group Dynamics*. Retrieved from https://psyarxiv.com/aq652/

Kumar, D., Hancock, J., Thomas, K., & Durumeric, Z. (2023, April). Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference (WWW)*.

Lacey, M., & Herszenhorn, D. M. (2011, January 8). In Attack's Wake, Political Repercussions. *The New York Times.* *https://www.nytimes.com/2011/01/09/us/politics/09giffords.html?smid=url-share*

Lees, J., & Cikara, M. (2020). *Understanding and combatting false polarization*. 1–41.

Lerner, M. J. (1980). The belief in a just world. In *The Belief in a just World* (pp. 9-30). Springer, Boston, MA.

Levendusky, M. S., & Malhotra, N. (2016a). Does Media Coverage of Partisan Polarization Affect Political Attitudes? *Political Communication*, *33*(2), 283–301. https://doi.org/10.1080/10584609.2015.1038455

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

Lewin, K. (1951). In Dorwin Cartwright (Ed.), *Field theory in social science: selected theoretical papers*. Harpers.

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., and Sonnet, L. (2022). *Voteview: Congressional Roll-Call Votes Database*. https://voteview.com/

Lijphart, A. 2010. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries.* New Haven: Yale University Press.

Marie, A., & Petersen, M. B. (2023). Motivations to affiliate with audiences drive the sharing of partisan (mis) information on social media. Preprint retrieved from https://osf.io/nmg9h

Martherus, J. L., Martinez, A. G., Piff, P. K., & Theodoridis, A. G. (2019). Party Animals? Extreme Partisan Polarization and Dehumanization. *Political Behavior*, *0123456789*. https://doi.org/10.1007/s11109-019-09559-4

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist*, *62*(1), 16–42. https://doi.org/10.1177/0002764218759576

McGregor, S. C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, *20*, 1070-1086.

Meindl, P., Johnson, K. M., & Graham, J. (2016). The Immoral Assumption Effect: Moralization Drives Negative Trait Attributions. *Personality and Social Psychology Bulletin*, *42*(4), 540–553. https://doi.org/10.1177/0146167216636625

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(16), 1–9. https://doi.org/10.1073/pnas.2116851119

Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015). What do retweets indicate? Results from user survey and meta-review of research. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 658–661.

Mill, J. S. (1861). *Representative government*. Kessinger Publishing.

Minson, J. A., & Dorison, C. A. (2022). Why is exposure to opposing views aversive? Reconciling three theoretical perspectives. *Current Opinion in Psychology*, 101435.

Minson, J. A., & Chen, F. S. (2022). Receptiveness to Opposing Views: Conceptualization and Integrative Review. *Personality and Social Psychology Review*, *26*(2), 93-111.

Minson, J. A., Chen, F. S., & Tinsley, C. H. (2020). Why won't you listen to me? Measuring receptiveness to opposing views. *Management Science*, *66*(7), 3069-3094.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, *102*(2), 246.

Moore, M., Dorison, C., & Minson, J. (2021). The Role of Social Signaling in Selective Exposure to Information. Preprint retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3953333

Moore-Berg, S. L., Ankori-Karlinsky, L. O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, *117*(26), 14864-14872.

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1–14. https://doi.org/10.1016/j.jesp.2013.10.010

Mukerjee, S., Jaidka, K., & Lelkes, Y. (2022). The Political Landscape of the US Twitterverse. *Political Communication*, 1-31.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*, 109-138.

Mutz, D. C. (2002). Cross-Cutting Social Networks: Testing Democratic Theory in Practice. *The American Political Science Review*, *96*(1), 111–126.

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

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological*

bulletin, *83*(4), 602.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. In *Cognitive Science* (Vol. 26, Issue 5). https://doi.org/10.1016/S0364-0213(02)00082-4

Oosterhoff, B., Poppler, A., & Palmer, C. A. (2021). Early Adolescents Demonstrate Peer-Network Homophily in Political Attitudes and Values. *Psychological Science*, 09567976211063912.

Parker, M. T., & Janoff-Bulman, R. (2013). Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love". *Social Justice Research*, *26*(1), 81-96.

Peterson, E., & Iyengar, S. (2020). Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading. *The American Journal of Political Science*, *00*(Forthcoming), 1–15. https://doi.org/10.1111/ajps.12535

Pew Research Center. (2019a). Partisan Antipathy: More Intense, More Personal. Retrieved from https://www.pewresearch.org/politics/2019/10/10/partisan-antipathy-more-intense-more-personal/

Pew Research Center. (2019b). In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. Retrieved from https://www.pewresearch.org/politics/2019/12/17/in-a-politically-polarized-era-sharp-divides-in-both-partisan-coalitions/

Pew Research Center. (2019b). In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. Retrieved from https://www.pewresearch.org/politics/2019/12/17/in-a-politically-polarized-era-sharp-divides-in-both-partisan-coalitions/

Pew Research Center. (2019c). Political Independents: Who they are, what they think. Retrieved from https://www.pewresearch.org/politics/2019/03/14/political-independents-who-they-

are-what-they-think/

Pew Research Center. (2020b). Far more Americans see ‘very strong’ partisan conflicts now than in the last two presidential election years. Retrieved from https://www.pewresearch.org/fact-tank/2020/03/04/far-more-americans-see-very-strong-partisan-conflicts-now-than-in-the-last-two-presidential-election-years/

Pew Research Center. (2023). Social Media Use in 2021. Retrieved from https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Pew Research Center. (2024). News on Twitter: Consumed by Most Users and Trusted by Many. Retrieved from https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/

Pew Research Center. (2021c). The Behaviors and Attitudes of U.S. Adults on Twitter. Retrieved from https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/

Pew Research Center. (2022a). Americans at the ends of the ideological spectrum are the most active in national politics. Retrieved from https://www.pewresearch.org/fact-tank/2022/01/05/americans-at-the-ends-of-the-ideological-spectrum-are-the-most-active-in-national-politics/

Pew Research Center. (2022b). Politics on Twitter: One-Third of Tweets From U.S. Adults Are Political. Retrieved from https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/

Pew Research Center. (2022c). A growing share of Americans are familiar with ‘cancel culture’. Retrieved from https://www.pewresearch.org/short-reads/2022/06/09/a-growing-share-of-americans-are-familiar-with-cancel-culture/

Popper, K. R. (1945). *The open society and its enemies*. Princeton University Press.

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In *Advances in experimental social psychology* (Vol. 28, pp. 161–209). Elsevier.

Prolific Team. (2022). What are the advantages and limitations of an online sample? Retrieved from https://researcher-help.prolific.co/hc/en-gb/articles/360009501473-What-are-the-advantages-and-limitations-of-an-online-sample-

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

Przeworski, A., Alvarez, R. M., Alvarez, M. E., Cheibub, J. A., Limongi, F., & Neto, F. P. L. (2000). *Democracy and development: Political institutions and well-being in the world, 1950-1990* (Issue 3). Cambridge University Press.

Puryear, C., Kubin, E., Schein, C., Bigman, Y., & Gray, K. (2022). *Bridging Political Divides by Correcting the Basic Morality Bias*. Preprint retrieved from https://psyarxiv.com/fk8g6

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences*, *114*(32), 8511–8516. https://doi.org/10.1073/pnas.1705238114

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118.

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2022). People think that social media platforms do (but should not) amplify divisive content. Preprint retrieved from https://psyarxiv.com/gmun4

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual Versus Assumed Differences in Construal: "Naive Realism" in Intergroup Perception and Conflict. *Journal of Personality and Social Psychology*, *68*(3), 404–417. https://doi.org/10.1037/0022-3514.68.3.404

Robison, J., & Mullinix, K. J. (2016). Elite polarization and public opinion: How polarization is communicated and its effects. *Political Communication*, *33*(2), 261–282.

Rodriguez, C. G., Moskowitz, J. P., Salem, R. M., & Ditto, P. H. (2017). Partisan selective exposure: The role of party, ideology and ideological extremity over time. *Translational Issues in Psychological Science*, *3*(3), 254–271. https://doi.org/10.1037/tps0000121

Roose, K., Isaac, M., & Frenkel, S. (2020, November 24). Facebook Struggles to Balance Civility and Growth. *The New York Times*. nytimes.com/2020/11/24/technology/facebook-election-misinformation

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software*, *48*, 1-36.

Rozin, P., Lowery, L., Haidt, J., & Imada, S. (1999). The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586. http://faculty.virginia.edu/haidtlab/articles/rozin.lowery.1999.moral-emotion-triad-hypothesis.pub012.pdf

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296-320.

Ryan, T. J. (2017). No Compromise: Political Consequences of Moralized Attitudes. *American Journal of Political Science*, *61*(2), 409–423. https://doi.org/10.1111/ajps.12248

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affective Science, 2*, 379-390.

Schwalbe, M. C., Cohen, G. L., & Ross, L. D. (2020). The objectivity illusion and voter polarization in the 2016 presidential election. *Proceedings of the National Academy of Sciences*, *117*(35), 21218-21229.

Sears, D. O., & Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, *31*(2), 194-213.

Sharot, T., & Sunstein, C. R. (2020). How people decide what they want to know. *Nature Human Behaviour*, *4*(1), 14–19.

Sherif, M., & Sherif, C. W. (1953). Groups in harmony and tension; an integration of studies of intergroup relations.

Silver, I., & Shaw, A. (2022). When and why "staying out of it" backfires in moral and political disagreements. *Journal of Experimental Psychology: General*.

Skitka, L. J., & Bauman, C. W. (2008). Moral conviction and political engagement. *Political Psychology*, *29*(1), 29–54. https://doi.org/10.1111/j.1467-9221.2007.00611.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral Conviction: Another Contributor to Attitude Strength or Something More? *Journal of Personality and Social Psychology*, *88*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology, 72*, 347–366.

Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. *Political Psychology*, *35*(SUPPL.1), 95–110. https://doi.org/10.1111/pops.12166

Skitka, L. J., Washburn, A. N., & Carsel, T. S. (2015). The psychological foundations and consequences of moral conviction. *Current Opinion in Psychology, 6,* 41–44. https://doi.org/10.1016/j.copsyc.2015.03.025

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1), 19-41.

Soroka, S., & McAdams, S. (2015). News, Politics, and Negativity. *Political Communication*, *32*(1), 1–22. https://doi.org/10.1080/10584609.2014.881942

Stroud, N. J., & Collier, J. R. (2018). Selective exposure and homophily during the 2016 presidential campaign. *An unprecedented election: Media, communication, and the electorate in the 2016 campaign*, 21-39.

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The Upside of Outrage. *Trends in Cognitive Sciences*, *22*(12), 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

Ståhl, T., Zaal, M. P., & Skitka, L. J. (2016). Moralized Rationality: Relying on Logic and Evidence in the Formation and Evaluation of Belief Can Be Seen as a Moral Issue. *PLoS ONE*, *11*, 1–38. https://doi.org/10.1371/journal.pone.0166332

Stanley, M. L., Whitehead, P. S., Sinnott-Armstrong, W., & Seli, P. (2020). Exposure to opposing reasons reduces negative impressions of ideological opponents. *Journal of Experimental Social Psychology*, *91*(July), 104030. https://doi.org/10.1016/j.jesp.2020.104030

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*, 130.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in*

cognitive sciences, *7*(7), 320-324.

The Economic Intelligence Unit. (2016). Democracy Index 2016. Retrieved July 12, 2022, from https://www.eiu.com/public/topical_report.aspx?campaignid=DemocracyIndex2016

The Economic Intelligence Unit. (2022). Democracy Index 2022. Retrieved July 31, 2023, from https://www.eiu.com/n/campaigns/democracy-index-2022/

Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, *24*, 2454-2462.

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. Personality and Individual Differences, 185, 111269.

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018)*.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124-1131.

Van Boven, L. (2000). Pluralistic ignorance and political correctness: The case of affirmative action. *Political Psychology*, *21*(2), 267-276.

Van Boven, L., Judd, C. M., & Sherman, D. K. (2012). Political Polarization Projection: Social Projection of Partisan Attitude Extremity and Attitudinal Processes. *Journal of Personality and Social Psychology*, *103*(1), 84–100. https://doi.org/10.1037/a0028145

Van Zomeren, M., Postmes, T., Spears, R., & Bettache, K. (2011). Can moral convictions motivate the advantaged to challenge social inequality? Extending the social identity model of collective action. *Group Processes & Intergroup Relations*, *14*(5), 735-753.

Vancouver, J. B., & Carlson, B. W. (2015). All things in moderation, including tests of mediation (at least some of the time). *Organizational Research Methods*, *18*(1), 70-91.

Vial, A. C., Brescoll, V. L., & Dovidio, J. F. (2019). Third-party prejudice accommodation increases gender discrimination. *Journal of Personality and Social Psychology,* 117(1), 73–98.

Viciana, H., Hannikainen, I. R., & Gaitan Torres, A. (2019). The dual nature of partisan prejudice: Morality and identity in a multiparty system. *PloS one*, *14*(7), e0219509.

Voelkel, J. G., Chu, J., Stagnaro, M. N., Mernyk, J. S., Redekopp, C., Pink, S. L., ... & Willer, R. (2023). Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes. *Nature human behaviour*, *7*(1), 55-64.

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, 102199.

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, *125*(5), 617.

Wang, Y. A., & Todd, A. R. (2020). Evaluations of empathizers depend on the target of empathy. *Journal of Personality and Social Psychology*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/pspi0000341

Ward, D. G., & Tavits, M. (2019). How partisan affect shapes citizens' perception of the political world. *Electoral Studies*, *60*, 102045.

Warzel, C. (2020, February 2). Twitter Is Real Life: Elites just pretend it's not. *The New York Times.* https://www.nytimes.com/2020/02/19/opinion/twitter-debates-real-life.html

Webster, D. M., Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. Journal of Personality and Social Psychology, 67(6), 1049–1062.

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide. *Perspectives on Psychological Science*, *10*(2), 145–158. https://doi.org/10.1177/1745691615569849

Wilson, A. E., Parker, V., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, *34*, 223–228. https://doi.org/10.1016/j.cobeha.2020.07.005

Woitzel, J., & Koch, A. (2022). Ideological prejudice is stronger in ideological extremists (vs. moderates). *Group Processes & Intergroup Relations*, 13684302221135083.

Wojcieszak, M., Azrout, R., & De Vreese, C. (2018). Waving the red cloth: Media coverage of a contentious issue triggers polarization. *Public Opinion Quarterly*, *82*(1), 87–109.

Wojcieszak, M., Winter, S., & Yu, X. (2020). Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. *Mass Communication and Society*, *23*(4), 455-483.

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2022). Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science advances*, *8*(39), eabn9418.

Wolf, M. R., Strachan, J. C., & Shea, D. M. (2012). Incivility and standing firm: A second layer of partisan division. *PS: Political Science & Politics*, *45*(3), 428–434.

Yang, J., Rojas, H., Wojcieszak, M., Aalberg, T., Coen, S., Curran, J., ... & Tiffen, R. (2016). Why are "others" so polarized? Perceived political polarization and media use in 10

countries. *Journal of Computer-Mediated Communication*, *21*(5), 349-367.

Yeomans, M., Minson, J., Collins, H., Chen, F., Gino, F., & States, U. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, *160*(May), 131–148. https://doi.org/10.1016/j.obhdp.2020.03.011

Yu, X., Wojcieszak, M., & Casas, A. (2021). Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users.

Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge university press.

Zhou, S. & Barbaro, N. (2023). *Understanding Student Expression Across Higher Ed: Heterodox Academy's Annual Campus Expression Survey*. Heterodox Academy.

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, *12*, 415-438.

Zlatev, J. J. (2019). I May Not Agree With You, but I Trust You: Caring About Social Issues Signals Integrity. *Psychological Science*, 095679761983794. https://doi.org/10.1177/0956797619837948

Appendix: Supplemental Material

A.1 Supplemental analyses for Studies 3 and 4

A.1.1 Robustness checks

I ran key analyses in additional ways, to ensure the robustness of my results. I ran models described in the main text with two different modifications. First, I ran models changing the DV to be raw rather than log-transformed Like and Retweet counts, varying the dummy codes to obtain all pairwise comparisons (see Table A1.1). Next, I ran models eliminating the 20 tweets that were present in both Studies 3 and 4, again varying the dummy codes to obtain all pairwise comparisons (see Table A1.2). In all cases results replicate what I report in the main text, though for the raw counts significance was sometimes marginal.

Table A1.1

Comparing raw positive feedback to Senators' neither vs. engaging and neither vs. dismissing tweets

Model Comparison	DV	Covariates	Study	*b (SE)*	*t*	*p*	*d*
Engaging (1) versus dismissing (0)	Likes	No	3	-1225.57 (505.38)	-2.43	.015	-0.33
			4	-2666.44 (1303.14)	-2.05	.041	-0.35
		Yes	3	-1138.80 (499.89)	-2.28	.023	-0.31
			4	-2277.42 (1282.69)	-1.78	.076	-0.30
	Retweets	No	3	-351.41 (166.33)	-2.11	.035	-0.29
			4	-693.59 (287.40)	-2.41	.016	-0.42
		Yes	3	-321.00 (165.15)	-1.94	.052	-0.27
			4	-610.55 (284.62)	-2.15	.032	-0.37
Engaging (1) vs. neither (0)	Likes	No	3	41.10 (270.81)	0.15	.879	0.01
			4	-2341.76 (1105.77)	-2.12	.034	-0.31
		Yes	3	-103.87 (267.87)	-0.39	.698	-0.03
			4	-2012.36 (1087.43)	-1.85	.064	-0.27
	Retweets	No	3	3.06 (89.13)	0.03	.973	0.00
			4	-665.05 (243.85)	-2.73	.006	-0.40
		Yes	3	-38.90 (88.50)	-0.44	.660	-0.03
			4	-600.24 (241.27)	-2.49	.013	-0.35
Dismissing (1) vs. neither (0)	Likes	No	3	1266.67 (432.76)	2.93	.003	0.34
			4	324.68 (764.12)	0.42	.671	0.04
		Yes	3	1036.59 (429.09)	2.42	.016	0.28
			4	265.07 (754.93)	0.35	.726	0.04
	Retweets	No	3	354.46 (142.42)	2.49	.013	0.29
			4	28.55 (168.47)	0.17	.865	0.02

	Yes	3	283.01 (141.75)	2.00	.046	0.24
		4	15.22 (167.27)	0.09	.928	0.01

Note. With (and without) covariates, the models for Study 3 had 4976 (4995) *df*; those for Study 4 had 1121 (1130).

Table A1.2

Comparing positive feedback for Senators' neither vs. engaging and neither vs. dismissing tweets, after removing overlapping tweets from each dataset

Comparison	DV	Covariates	Study	*b (SE)*	*t*	*p*	*d*
Engaging (1) versus dismissing (0)	Likes	No	3	-0.79 (0.19)	-4.07	< .001	-0.57
			4	-0.81 (0.24)	-3.39	< .001	-0.61
		Yes	3	-0.65 (0.16)	-3.99	< .001	-0.56
			4	-0.77 (0.23)	-3.30	< .001	-0.59
	Retweets	No	3	-0.96 (0.17)	-5.64	< .001	-0.78
			4	-1.14 (0.23)	-4.86	< .001	-0.87
		Yes	3	-0.85 (0.16)	-5.36	< .001	-0.75
			4	-1.09 (0.23)	-4.72	< .001	-0.84
Engaging (1) vs. neither (0)	Likes	No	3	0.10 (0.10)	0.96	.339	0.07
			4	-0.66 (0.20)	-3.25	.001	-0.49
		Yes	3	0.02 (0.09)	0.18	.861	-0.01
			4	-0.64 (0.20)	-3.23	.001	-0.49
	Retweets	No	3	0.00 (0.09)	0.05	.961	0.00
			4	-0.98 (0.20)	-4.96	< .001	-0.75
		Yes	3	-0.06 (0.08)	-0.73	.467	-0.05
			4	-0.97 (0.20)	-4.96	< .001	-0.75
Dismissing (1) vs. neither (0)	Likes	No	3	0.89 (0.17)	5.33	< .001	0.64
			4	0.15 (0.14)	1.09	.276	0.11
		Yes	3	0.67 (0.14)	4.74	< .001	0.57
			4	0.14 (0.14)	1.01	.315	0.11
	Retweets	No	3	0.96 (0.15)	6.60	< .001	0.79
			4	0.15 (0.14)	1.07	.283	0.11
		Yes	3	0.79 (0.14)	5.79	< .001	0.69
			4	0.12 (0.13)	0.91	.362	0.09

Note. Likes and retweets are log-transformed. With (and without) covariates, the models for Study 3 had 4955 (4974) *df*; those for Study 4 had 1103 (1110).

A.1.2 Test of mediation by linguistic features

Prior work has identified some linguistic markers of tweets that get a lot of positive feedback and that might naturally occur more (or less) in dismissive than engaging tweets. When tweeting their dismissal of opponents, Senators might naturally express more incivility, outrage, negative emotions, and certainty, as well as fewer positive emotions, all of which are linked with more Likes and Retweets (Brady et al., 2017; Evans et al., 2023; Frimer et al., 2022, Schöne et al., 2021). I considered whether any of these played a role in the popularity of Senators'

dismissive tweets, examining both whether these linguistic markers explain the popularity of dismissing tweets, and whether this popularity persists over and above linguistic markers.

We measured linguistic markers using the same dictionaries that have been used in prior work on Twitter reactions: *incivility* (i.e., toxicity scores from PerspectiveAPI; see Frimer et al., 2022) and *positive*, *negative* non-moral, and *moral outrage* (i.e., negative moral) emotional sentiments (from AFINN sentiment dictionaries; see Nielsen, 2011). For *certainty*, I used the LIWC certitude dictionary (Boyd et al., 2022). Since word count methods are imperfect, I also considered six alternative measures related to these constructs: LIWC's dictionaries for politeness (inversely related to incivility), moralization, positive tone and negative tone (related to moral outrage, positive, and negative emotions, respectively, but without necessarily being emotional in nature), and all-or-none language (related to certainty).

We ran separate multilevel mediation models for each possible combination of DV (Likes or Retweets) and linguistic mediator (incivility and politeness, outrage, negative non-moral sentiments, negative tone and moralization, positive sentiments and positive tone, and certainty and all-or-none language), with dismissing (0) versus engaging (1) as the IV. Each model also included a random intercept for Senator. Table A1.3 reports for each model the a-link (do engaging and dismissing tweets differ on the linguistic marker?), the indirect effect (does the linguistic marker help explain why dismissing tweets receive more positive feedback), and the direct effect (do dismissing tweets receive more positive feedback even accounting for the linguistic marker?).

Table A1.3

Role of linguistic markers in explaining the popularity of Senators' dismissing tweets

		a-link				Indirect effect			Direct effect		
Mediator	Study	*b*	*SE*	*p*	DV	*b*	*SE*	*p*	*b*	*SE*	*p*
Incivility	3	-0.10	0.008	< .001	Likes	-0.85	0.197	< .001	-1.16	0.321	< .001
					Retweets	-0.83	0.175	< .001	-1.19	0.280	< .001
	4	-0.16	0.016	< .001	Likes	-0.57	0.252	.023	-0.43	0.405	.287

				Retweets	-0.53	0.233	.022	-0.83	0.375	.027
Politeness	3	0.27	0.242	.072	Likes	-0.03	0.034	.389	-1.98	0.275	< .001
					Retweets	-0.03	0.032	.306	-1.99	0.242	< .001
	4	0.22	0.086	.012	Likes	-0.04	0.066	.546	-1.06	0.336	.002
					Retweets	-0.03	0.060	.583	-1.42	0.311	< .001
Outrage (negative moral emotions)	3	-0.17	0.089	.058	Likes	-0.01	0.032	.794	-2.00	0.276	< .001
					Retweets	-0.00	0.028	.978	-2.02	0.243	< .001
	4	-0.50	0.120	< .001	Likes	-0.04	0.106	.687	-1.05	0.344	.002
					Retweets	-0.05	0.098	.613	-1.41	0.318	< .001
Moralization	3	-1.37	0.253	< .001	Likes	0.02	0.086	.779	-2.03	0.287	< .001
					Retweets	-0.03	0.076	.726	-2.00	0.253	< .001
	4	-2.05	0.387	< .001	Likes	-0.04	0.134	.743	-1.06	0.357	.003
					Retweets	-0.00	0.124	.975	-1.38	0.329	< .001
Negative non-moral emotions	3	-0.79	0.129	< .001	Likes	-0.19	0.105	.065	-1.82	0.290	< .001
					Retweets	-0.18	0.093	.051	-1.84	0.255	< .001
	4	-0.97	0.172	< .001	Likes	-0.03	0.142	.816	-1.04	0.357	.004
					Retweets	-0.05	0.131	.710	-1.41	0.330	< .001
Negative tone	3	-2.83	0.365	< .001	Likes	-0.09	0.129	.486	-1.92	0.302	< .001
					Retweets	-0.05	0.113	.676	-1.98	0.266	< .001
	4	-3.43	0.511	< .001	Likes	-0.21	0.182	.241	-0.81	0.375	.032
					Retweets	-0.19	0.168	.249	-1.19	0.346	.001
Positive emotions	3	-0.20	0.238	.407	Likes	-0.05	0.063	.418	-1.96	0.268	< .001
					Retweets	-0.03	0.041	.427	-1.99	0.238	< .001
	4	0.99	0.326	.002	Likes	-0.00	0.076	.954	-1.00	0.336	.003
					Retweets	-0.00	0.070	.999	-1.36	0.311	< .001
Positive tone	3	3.15	0.531	< .001	Likes	-0.20	0.103	.056	-1.81	0.289	< .001
					Retweets	-0.19	0.091	.038	-1.84	0.254	< .001
	4	2.90	0.511	< .001	Likes	-0.17	0.146	.260	-0.86	0.359	< .001
					Retweets	-0.14	0.135	.286	-1.24	0.332	< .001
Certainty	3	-0.63	0.191	< .001	Likes	-0.00	0.055	.977	-2.01	0.280	< .001
					Retweets	-0.01	0.048	.913	-2.03	0.246	< .001
	4	-0.08	0.199	.683	Likes	-0.00	0.013	.797	-1.02	0.331	.002
					Retweets	-0.00	0.013	.791	-1.39	0.305	< .001
All or none	3	-0.38	0.234	.103	Likes	-0.02	0.029	.539	-1.99	0.275	< .001
					Retweets	-0.02	0.026	.499	-2.01	0.242	< .001
	4	0.05	0.257	.849	Likes	-0.01	0.051	.850	-1.01	0.327	.002
					Retweets	-0.01	0.045	.850	-1.38	0.302	< .001

The linguistic markers tended to differ between tweet types in line with my expectations: Compared to engaging tweets, dismissing tweets expressed more incivility and less politeness, more outrage, moralization, negative emotion, negative tone and certainty (in Study 3 only, and no more absolute language), and less positive emotion (Study 4 only) and positive tone. However, these differences tended not to explain my findings: The only consistent indirect effects were through incivility, but politeness as an alternative marker of a similar conceptual construct did not replicate the pattern. Moreover, even accounting for incivility the direct effects

remained significant in all but one test. Incivility may thus contribute to dismissing tweets' popularity, but dismissing tweets seem more popular for reasons beyond this and all other linguistic differences I considered.

A.2 Supplemental Studies S1 and S2

Studies S1 and S2 both tested whether people's preferences for engaging versus dismissing shift depending on either the communicator's role (Senator vs. citizen), or the communication medium (tweet vs. offline). Each used a 2×2 design manipulating each of these factors independently between subjects. Participants saw multiple instances of both engaging and dismissing by a co-partisan, and in each case reported their feelings of warmth towards the communicator. For theoretical precision, Study S1's participants rated artificial (i.e., experimenter-generated) tweets or vignettes inspired by prior research; For realism, Study S2's participants rated a small set of real Senator tweets—or vignettes inspired by them—from Study 4.

Study S1 was preregistered at this link and Study S2 at this link; exclusions and analyses reported below follow my preregistration unless otherwise indicated. The preregistration made no explicit *a priori* hypotheses but given existing literature (summarized in the main text's introduction) and what would explain the discrepancy between Study 3 and 4's results and prior research, I might have expected worse reactions to engaging in the case of Twitter and/or Senators.

A.2.1 Method

A.2.1.1 Participants

We recruited American participants from Prolific Academic who had registered for Prolific before July 25, since reports indicated data quality had decreased due to an influx of

registrations after this date (Charalambides, 2021). I aimed for 1200 observations per study, before exclusions; in Study S1 (S2), each participant provided two (six) observations. I used Prolific screening to recruit approximately half self-identified Democrats and Republicans.

Table A2.1 reports demographics and exclusions. The survey automatically ejected any participants who failed an English comprehension check at the survey's start; this check described multiple actors in a short story and required that participants identify the referent of a particular pronoun. I also excluded participants who self-reported providing low-quality data. Study S1's sample had no ideological lean (M = 3.24, SD = 1.67, below the scale midpoint, $t(612) = -3.79$, $p < .001$), while Study S2 leaned liberal (M = 3.35, SD = 1.57, not different from the scale midpoint, $t(199) = -1.34$, $p = .183$).

Table A2.1
Demographics of Appendix Studies S1–S4

Study	N recruited	Low quality data	Final N	Observations	Gender	M_{age}
S1	659	2	657	1027	42% men (1 trans), 56% women, 1% non-binary 1% agender, gender-fluid, or gender-queer	39.9
S2	200	1	199	961	42% male, 55% female, 1% non-binary, 1% queer, 0.5% agender	37.7
S3	200	2	198	1967	40% men, 57% women, 1% non-binary, 0.5% agender	37.5
S4	200	0	200	1998	48% men (1 trans), 47.5% women, 3.5% non-binary, 0.5% agender	37.6

A.2.1.2.1 Stimuli: Study S1

Study S1 used carefully controlled, experimenter-generated stimuli to convey engaging and dismissing, thereby minimizing any extraneous confounds between the two. Each participant saw two instances of political communication, both randomly assigned from same cell in the 2 (Senator versus citizen) × 2 (tweet vs. offline) between-subjects matrix. Each participant saw, in randomized order, one communicator engaging constructively with and another dismissively dismissing opponents and their views, such that the design analyzed was a 2 (within subjects:

engaging vs. dismissing) × 2 (between subjects: Senator vs. citizen) × 2 (between subjects: tweet vs. offline).

For generalizability I also manipulated the scenario in which the communication took place (a TV interview vs. a conversation about recent interactions) and the political issue (randomly drawn from the same six issues used in Study S2: climate change, efforts to address COVID-19, suppressing violent outgroup extremists, tax plans, COVID-19 vaccine distribution priorities, guns). I randomly assigned levels of these variables within subjects (with the stipulation that they be different for each of the two communications each participant saw). Also, communicators' political beliefs always matched the participant's, adding another between-subjects variable (Republican vs. Democrat) that was not randomly assigned. Thus, the full design was a 2 (within subjects: engaging vs. dismissing) × 2 (between subjects: Senator vs. citizen) × 2 (between subjects: tweet vs. offline) × 2 (within subjects: café versus news report) × 2 (between subjects: Democratic vs. Republican) × 6 (within subjects: issue), requiring 192 different stimuli.

Figure A1.1 displays real stimuli from the study selected to show all variables levels and close parallels in language across variations. The top left shows a Rep. citizen on Twitter dismissing gun rights opponents in a TV scenario; the middle left shows a Rep. citizen offline dismissing climate change opponents in a TV scenario; the bottom left shows a Rep. Senator offline engaging gun control opponents in a recent interactions scenario; the top right shows a Rep. Senator dismissing climate change opponents in a TV scenario; the middle right shows a Rep. Senator offline engaging gun opponents in a TV scenario; the bottom right shows a Dem. Senator on Twitter dismissing climate opponents in a recent interactions scenario.

Figure A1.1 Examples of stimuli used in Study S1.

Note that Senators were portrayed as real Senators—Democrats Gary Peters (MI) or Jacky Rosen (NV); Republicans Mike Rounds (SD) or Marsha Blackburn (TN). Citizen communicators used the names and likenesses of the Senators to avoid any associated confounds, expecting that participants would not recognize them—we intentionally selected recently elected Senators with relatively low profiles, about whom participants would not have strong pre-existing opinions.

A.2.1.2.2 Stimuli: Study S2

We first selected six real tweets from Study 4 (three each to represent engaging and dismissing; see preregistered materials). My goal was to select tweets that were good exemplars of engaging and dismissing, that I could readily adapt into vignettes about citizens' offline interactions, and that I could transpose into parallel versions such that I could always show

Democratic (Republican) participants stimuli about fellow Democrats (Republicans). For variety, I selected tweets that focused on the six different political issues mentioned in Study S1: gun regulations, climate change, tax plans, whom to prioritize in COVID-19 vaccine distribution, efforts to address COVID-19, and suppressing violent outgroup extremists. These six tweets (alongside parallel versions attributed to an opposing party Senator, described below) served as stimuli for one of the between-subjects cells in this design. I created the other three cells by modifying these tweets in three ways.

To create the citizen tweets, I replaced the Senator's name and Twitter handle with an experimenter-generated name and handle, and the Senator's profile picture with a plain picture (i.e., not their official Senate photo) of another Senator from the other party. When the original tweets mentioned other Senators in a way that citizens' tweets would not (e.g., discussing policy with other Senators), I modified them accordingly (e.g., the citizen expressed the same sentiment as the Senator about the policy discussions).

To create both Senator and citizen offline vignettes, I stated the communicator's name and partisan affiliation, their policy beliefs on the issue mentioned in the tweet stimulus, and described their engaging or dismissing behavior.

Finally, to ensure I had versions I could show both conservative and liberal participants, I created a parallel version of each stimulus that depicted a Senator with the opposing party affiliation and stance on the issue (Pre-testing revealed near unanimous support for addressing the pandemic so both Democrats and Republicans saw the same stance for this issue). In selecting tweets, I prioritized ones posted by lower profile Senators (e.g., who had not run for President), to reduce the chance that participants' pre-existing feelings would influence their evaluations, and selected similarly low-profile Senators to be their counterparts in the parallel

opposing party versions. But I did ultimately include one tweet by Senator Ted Cruz, a prominent Republican figure at the time of this study; in creating the parallel Democratic version, I portrayed the communicator as Senator Bernie Sanders, a figure likely to be similarly well-known among Democrats. Figure A1.2 displays real stimuli selected to show all variables levels and close parallels in language across variations. Tweets on the left (middle / right) are real tweets from the Study 4 dataset (are citizen variations of the real tweet shown on the left / are versions of the real tweet altered to depict the opposing party and associated stance), and the offline version of each tweet is shown below it.

Figure A1.2. Example Stimuli, Study S2

A.2.1.3 Procedure

Studies S1 and S2 had nearly identical procedures, with the exception of the number of communicators participants rated. After completing the English comprehension check,

participants reported their demographics, including their ideological orientation using a scale with no midpoint: “In general, to what extent do you consider yourself to be liberal, moderate, or conservative?” (1 = Very liberal, 6 = Very conservative).

To summarize, participants in Study S1 (S2) read about two (six) communicators. In both studies, I randomly assigned the communications to take the form of tweets or vignettes, and to be by Senators or regular citizens. For each participant, half of the communicators that they saw engaged and the other half dismissed, in randomized order; all participants always saw communicators endorsing the issue stance typical of their own ideology (for example, a participant who selected a point on the conservative half of the scale would see only communicators who supported gun rights, or opposed violent left-wing extremists).

Beneath each communication, participants reported their feelings of warmth toward the communicator, using the same measure described in Study 1 in the main text.

Participants then reported their position on each of the six issues in the set (even in Study S1, where they would only have seen communicators mention two of the issues). For instance, to report their position on guns they responded to the following prompt:

> This issue concerns the government's role in regulating the manufacture, sale, transfer, possession, modification, or use of firearms by civilians. Those who favor increasing government regulation are described as supporting gun control, while those who favor less government regulation are described as supporting gun rights.

In a significant minority of cases (Study S1: 287 of 1314 observations, or 22%; Study S2: 233 of 1194 observations, or 20%), their stance on an assigned issue was atypical of their self-reported ideology (e.g., some participants identified as liberal but supported gun rights). I excluded such observations, so that I could test people’s attitudes toward *co-partisans* who engaged with or dismissed *shared opponents’* views. Finally, participants self-reported whether they provided quality data.

A.2.2 Results

A.2.2.1 Preferences for engaging versus dismissing: Are they different for Senators / on Twitter?

In each study, a multilevel model predicted feelings toward the Senator from action (dismissing = 0, engaging = 1), communication medium (offline = -1, Twitter = 1), communicator role (citizen = -1, Senator = 1), the interactions among them, and a random intercept for participant. As preregistered, I focused on interactions involving the communicator's behavior, to test whether Senators or Twitter elicited stronger preferences for dismissing. Three of the four relevant two-way interactions were significant or marginally so, but in the *opposite* direction from what prior literature would suggest; see Table A2.2 and Figure A2.3: Participants' relative preference for engaging over dismissing was *stronger* on Twitter versus offline, and for Senators versus regular citizens (Study S1 only). In neither study did participants report feeling more warmly toward Senators who tweeted about dismissing (versus engaging with) opponents. Nonetheless, one difference between the studies stood out: In S1, participants preferred engaging Senators who allegedly authored experimenter-generated tweets; in S2 they had no preference for engaging or dismissing Senators who authored real tweets drawn from Study 4. I return to this discrepancy later in this document.

Table A2.2

Interactions and simple slopes for key tests

Study	Effect	*b*	*SE*	95% CI	*t*	*p*
	Action	11.27	1.28	8.77, 13.78	8.84	< .001
	Medium x Action interaction	2.26	1.28	-0.25, 4.76	1.77	.077
	Tweet simple slope	13.53	1.79	10.01, 17.05	7.55	< .001
	Offline simple slope	9.02	1.82	5.45, 12.58	4.97	< .001
	Role x Action interaction	2.20	1.28	-0.30, 4.71	1.73	.085
S1	Senator simple slope	13.48	1.75	10.05, 16.90	7.72	< .001
	Citizen simple slope	9.07	1.86	5.42, 12.73	4.88	< .001
	Medium × Role x Action interaction	0.93	1.28	-1.57, 3.44	0.73	.464
	Senator tweet simple slope	16.67	2.47	11.82, 21.52	6.74	< .001
	Senator offline simple slope	10.29	2.47	5.45, 15.13	4.17	< .001
	Citizen tweet simple slope	10.39	2.60	5.30, 15.49	4.00	< .001

	Citizen offline simple slope	7.75	2.67	2.52, 12.99	2.91	.004
S2	Action	0.40	1.31	-2.17, 2.96	0.30	.761
	Medium x Action interaction	-2.78	1.31	-5.34, -0.21	-2.12	.034
	Tweet simple slope	3.17	1.83	-0.41, 6.76	1.74	.083
	Offline simple slope	-2.38	1.87	-6.05, 1.30	-1.27	.205
	Role x Action interaction	0.70	1.31	-1.87, 3.27	0.53	.593
	Senator simple slope	1.10	1.81	-2.45, 4.64	0.61	.544
	Citizen simple slope	-0.30	1.89	-4.02, 3.41	-0.16	.873
	Medium × Role x Action interaction	3.33	1.31	0.77, 5.90	2.55	.011
	Senator tweet simple slope	0.54	2.62	-4.61, 5.69	0.21	.837
	Senator offline simple slope	1.65	2.48	-3.22, 6.53	0.67	.506
	Citizen tweet simple slope	5.81	2.54	0.82, 10.80	2.28	.023
	Citizen offline simple slope	-6.41	2.80	-11.91, -0.91	-2.29	.022

Note. The Study S1 (S2) model had 1017 (951) *df*. Positive simple slope coefficients reflect a preference for engaging in that condition.

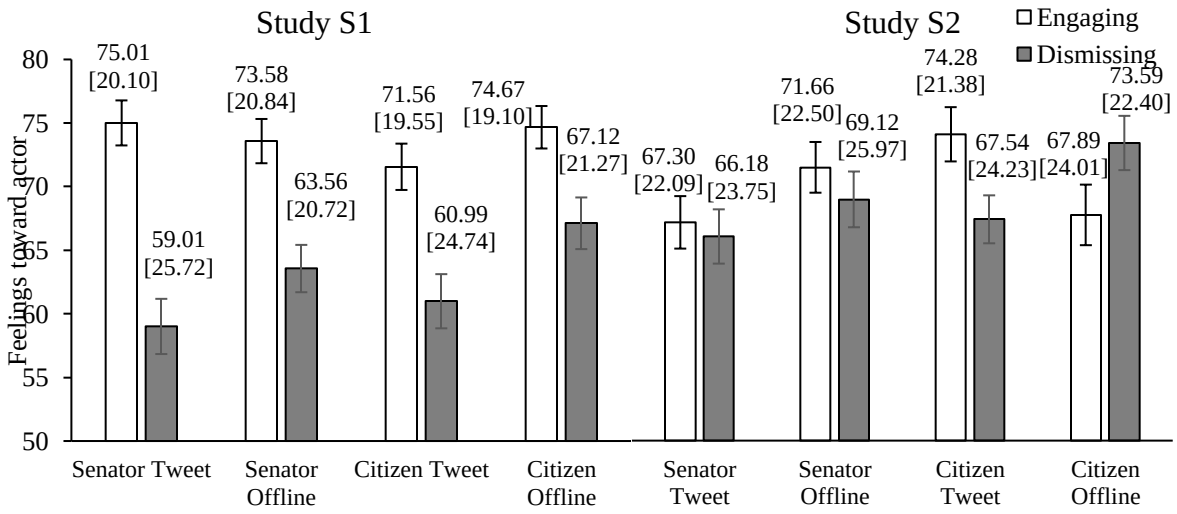


Figure S3. Participants' attitudes towards engaging, dismissing as a function of medium, role

A.2.2.2 Preferences for engaging versus dismissing: Do they differ among extremists vs. moderates?

Exploratory models tested whether political extremists find dismissing more appealing than political moderates. Multilevel models predicted attitudes from action (dismissing = 0, engaging = 1), as well as ideological orientation (midpoint-centered), ideological extremity (squared value of participant's midpoint-centered partisan affiliation), and their interactions with action; separately for each study (see Table A2.3). In Study S1 and S2, I found the same pattern of moderation by extremity reported in Study 5 in the main text.

Table A2.3

Preferences for Senators' engaging vs. dismissing tweets as a function of extremity

Extremity measure	Effect	*b*	*SE*	*t*	*df*	*p*
Study S1	Interaction with *Action*	-1.13	0.57	-2.00	1019	.046
	Most moderate simple slope	14.60	1.98	7.37	830	< .001
	Most extreme simple slope	7.78	2.27	3.43	839	.001
Study S2	Interaction with *Action*	-5.25	0.55	-9.51	953	< .001
	Most moderate simple slope	13.30	1.83	7.29	803	< .001
	Most extreme simple slope	-18.20	2.33	-7.81	773	< .001

Note. Positive coefficients indicate a relative preference for Senators' engaging tweets over their dismissing tweets.

A.3 Supplemental Study S3 and S4

Both Study S3 and S4 were preregistered (Study S3; Study S4). I ran Study S3 to test if crowdsourced participants feel more warmly toward co-partisan Senators when their real tweets engaged with or dismissed opponents. I ran Study S4 to as a confirmatory test of the moderating role of political extremity, which I reported for Studies S1 and S2 above and also found in Study S3. In both studies, I also assessed the frequency with which participants reacted to Senators' tweets, and exploratory analyses with this variable yielded the same results I report for Study 5 in the main text (but with only the feelings DV as that is all I measured).

A.3.1 Method

A.3.1.1 Participants

Table A2.1 (see Studies S1-S2) shows sample characteristics. I used Prolific Academic to recruit equal numbers of American participants who self-identified as Democrat, Republican, and Independents or other. In both studies, participants leaned Democrat (below implied scale midpoint of 3.5; S1: $M = 3.26$, $SD = 1.66$, $t(196) = -2.04$, $p = .043$; S2: $M = 3.12$, $SD = 1.69$, $t(199) = -3.13$, $p = .002$). A small minority of participants ($n_{S3} = 11$, $n_{S4} = 16$) reported frequently reacting to Senators' tweets; the remainder reported having done so once or twice ($n_{S3} = 78$, $n_{S4} = 61$), or never ($n_{S3} = 108$, $n_{S4} = 123$).

A.3.1.2 Procedure

Participants followed the same procedure as Study 5, with two exceptions. First, all participants reported their feelings of warmth toward the Senator who authored each tweet; I used no other measure of preferences. Second, participants did not report the three individual difference variables (endorsement of compromise, affective polarization, desire for party status); in Study S3, participants also did not report their ideological orientation (though they did report affiliation).

A.3.2 Results

A.3.2.1 Do participants prefer engaging or dismissing overall?

This was the analysis I preregistered for Study S3: a multilevel model predicted attitudes from tweet type (dismissing = 0, engaging = 1) with random intercepts for participant and tweet. Neither sample showed an overall preference: As in Study S2, participants overall felt equally warm toward Senators who posted engaging versus dismissing tweets (Study S3: b = -1.00, SE = 1.28, $t(1962)$ = -0.79, p = .434; Study S4: b = -0.75, SE = 1.40, $t(1993)$ = -0.53, p = .594).

A.3.2.2 Do frequent reactors respond differently than everyone else to engaging vs. dismissing, and is this related to their partisan extremity?

Because Studies S3 and S4 had so few frequent reactors, I collapsed across studies for these analyses; everything I report here also emerged in within-sample analyses. For analyses reported in this section, only those testing the moderating role of extremity were preregistered, and only for Study S4. All analyses, however, follow the strategy preregistered for Study 5 in the main text, without standardizing preferences because all participants completed the same measure (feelings toward the Senators).

Multilevel models predicted feelings of warmth from tweet type (dismissing = 0, engaging = 1), participant group (frequent reactors = 0, everyone else = 1), their interaction, and

random intercepts for tweet, participant, and study; see Figure A3.1. The same key interaction emerged as in Study 5, b = 13.86, SE = 3.52, t(3957) = 3.94, p < .001: Frequent reactors felt more warmly toward dismissing Senators compared to engaging Senators, b = -14.25, SE = 3.59, t(3957) = 3.97, p < .001, everyone else showed no preference, b = -0.39, SE = 1.11, t(3957) = -0.35, p = .726.

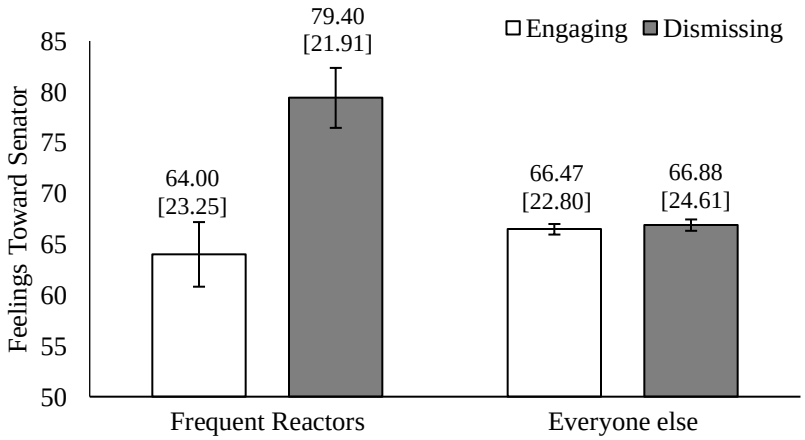


Figure A3.1. Participants' attitudes towards engaging and dismissing as a function of participant group.

We then tested whether extremists' preferences matched frequent reactors'; this was the analysis I preregistered for Study S2. A multilevel model predicted feelings from party affiliation (midpoint-centered), party extremity (squared value of participant's midpoint-centered partisan affiliation), and their interactions with tweet type (dismissing = 0, engaging = 1), and including random intercepts for tweet, participant, and study. The key party extremity × tweet type interaction was significant, b = -2.97, SE = 0.22, t(3956) = -13.29, p < .001: The most extreme participants felt more warmly toward co-partisan Senators whose tweets dismissed, rather than engaged, with opponents, b = -10.60, SE = 1.33, t(368) = -7.97, p < .001; the most moderate participants showed the opposite preference, b = 7.16, SE = 1.26, t(210) = 5.70, p < .001.

We did not measure the variables that in Study 5 played the role of second mediators in the serial mediation. Instead, I tested a model using participant group as the IV (frequent reactors = 0, everyone else = 1), preference for engaging (average feelings toward engaging Senators

minus average attitude toward dismissing Senators) as the DV, and party extremity as the mediator; see Figure A3.2. Similar to Study 5, frequent reactors' greater partisan extremity helped account for their preference for dismissing tweets, but only partially; a significant direct effect suggested that even controlling for extremity, frequent reactors' preferences still differed from other participants'.

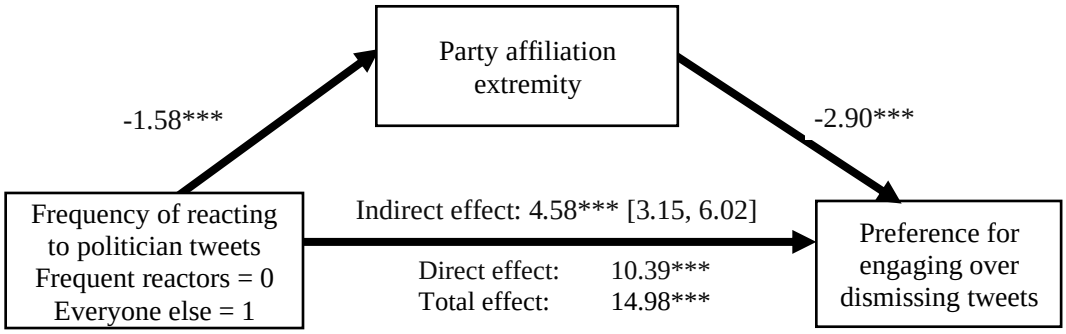


Figure A3.2. Extremity's role in explaining why frequent reactors preferred Senators' dismissing tweets.

A.4 Supplemental analyses for Study 5

A.4.1 Extremity results, moderated by response type

Additional preregistered analyses tested whether extremity moderated preferences for engaging versus dismissing tweets consistently across response measures. The model was identical to the extremity model reported in the main text with the addition of response measure (dummy coded with approval of the tweet as the reference group) and all its interactions (see Table A3.1). As mentioned in the text, the key tweet type × party extremity interaction was significant for both direct preference measures (approval: b = -0.10, SE = 0.02, t(4538) = -5.63, p < .001; b = -0.12, SE = 0.02, t(4538) = -6.78, p < .001), but not for intentions to Like and Retweet, b = -0.02, SE = 0.01, t(4538) = -1.50, p = 0.133).

Table A3.1

Interactions and simple slopes for tweet type by extremity by measure, Study 5

Effect	*b*	*SE*	*t*	*p*
Intercept	-0.38	0.11	-3.50	< **.001**
Tweet type	0.36	0.08	4.48	< **.001**
Party affiliation	-0.06	0.03	-1.87	.061

Feeling toward Senator (vs. approval of tweet)	-0.12	0.15	-0.82	.411
Intent to Like and Retweet (vs. approval of tweet)	0.15	0.15	1.01	.312
Extremity	0.10	0.02	4.20	< **.001**
Tweet type × party affiliation	-0.01	0.02	-0.32	.747
Tweet type × feeling toward Senator (vs. approval of tweet)	-0.11	0.11	-1.00	.316
Tweet type × intent to Like and Retweet (vs. approval of tweet)	-0.44	0.10	-451	< **.001**
Party affiliation × feeling toward Senator (vs. approval of tweet)	0.03	0.05	0.65	.515
Party affiliation × intent to Like and Retweet (vs. approval of tweet)	0.02	0.05	0.36	.716
Tweet type × extremity	-0.10	0.02	-5.63	< **.001**
Tweet type × feeling toward Senator (vs. approval of tweet)	0.07	0.04	1.97	**.049**
Tweet type × intent to Like and Retweet (vs. approval of tweet)	-0.03	0.03	-0.90	.369
Party affiliation × feeling toward Senator (vs. approval of tweet)	-0.01	0.03	-0.30	.764
Party affiliation × intent to Like and Retweet (vs. approval of tweet)	0.03	0.03	1.14	.253
Extremity × feeling toward Senator (vs. approval of tweet)	-0.03	0.03	-1.05	.292
Extremity × intent to Like and Retweet (vs. approval of tweet)	0.08	0.02	3.57	< **.001**

Note. The model had 4538 *df*. Positive simple slope coefficients reflect more favorable responses.

A.4.2 Robustness checks on the moderation by extremity

To test the robustness of the finding that extremists prefer dismissing tweets, I ran the same analyses described in the main text but using alternative operationalizations of extremity. First, I considered absolute (rather than squared) distance from midpoint, recoding participants' responses to the party affiliation variable so that strongly partisans received a 3, somewhat affiliated partisans received a 2, and leaning partisans received a 1. Second, I considered ideological extremity in place of affiliation extremity, squaring the distance from the midpoint of a 7-point scale (1 = Extremely Conservative, 4 = Moderate, 7 = Extremely Liberal). Table A3.2 reports the results; in all cases replicating what I observed in the main text.

Table A3.2

Preferences for Senators' engaging vs. dismissing tweets as a function of extremity

Extremity measure	Effect	*b*	*SE*	*t*	*df*	*p*
Absolute distance from midpoint	Interaction with *Tweet type*	-0.23	0.03	-7.97	4550	< .001
	Most moderate simple slope	0.15	0.04	3.43	3916	= .001
	Most extreme simple slope	-0.30	0.03	-8.83	3322	< .001
Ideological extremity	Interaction with *Tweet type*	-0.07	0.01	-10.08	4512	< .001
	Most moderate simple slope	0.21	0.04	5.04	3823	< .001
	Most extreme simple slope	-0.39	0.04	-10.48	3538	< .001

Note. Positive coefficients indicate a relative preference for Senators' engaging tweets over their dismissing tweets.

A.4.3 Analyses including only participants from pre-amendment sample

For transparency, I report results for all preregistered confirmatory tests using only the original sample recruited for Study 5, prior to the amendment I submitted when I realized that my smaller sample of frequent reactors had not rated all the tweets that everyone else had.

In the first model (across response measures), the key interaction was significant, b = 0.21, SE = 0.06, $t(3221) = 3.62$, $p < .001$: Frequent reactors responded more positively to dismissing tweets, b = -0.22, SE = 0.05, $t(3221) = -4.39$, $p < .001$, whereas infrequent reactors responded similarly to both types of tweet, b = -0.00, SE = 0.04, $t(3221) = -0.19$, $p = .847$.

In the second model (including moderations by response measure; see Table A3.3), only one of the three-way interactions was significant. Nonetheless, as in the main text, the key participant group × tweet type interaction was significant for approval (b = 0.40, SE = 0.11, $t(3213) = 3.55$, p = < .001) and marginal for feelings (b = 0.19, SE = 0.11, $t(3213) = 1.70$, p =.089); in both cases the direction of the interaction suggested frequent reactors had a stronger relative preference for dismissing than everyone else. The interaction was not significant for intentions to Like or Retweet (b = 0.11, SE = 0.08, $t(3213) = 1.32$, $p = .186$).

Table A3.3

Interactions and simple slopes for tweet type by reaction frequency by measure, Study 5

Effect	*b*	*SE*	*t*	*p*
Intercept	0.12	0.12	0.97	.330
Tweet type	-0.12	0.09	-1.32	.188
Participant group	-0.37	0.15	-2.50	.012
Feeling toward Senator (vs. approval of tweet)	0.13	0.17	0.75	.456
Intent to Like and Retweet (vs. approval of tweet)	0.23	0. 17	1.37	.170
Tweet type × participant group	0.40	0.11	3.55	<.001
Tweet type × feeling toward Senator (vs. approval of tweet)	-0.16	0.13	-1.19	.234
Tweet type × intent to Like and Retweet (vs. approval of tweet)	-0.10	0.11	-0.90	.367
Participant group × feeling toward Senator (vs. approval of tweet)	0.02	0.21	0.07	.942
Participant group × intent to Like and Retweet (vs. approval of tweet)	-0.32	0.21	-1.57	.116
3-way interaction (feeling vs. approval)	-0.21	0.16	-1.30	.194
3-way interaction (intent vs. approval)	-0.29	0.14	-2.10	.036

Note. The model had 3213 *df*. Positive simple slope coefficients reflect a preference for engagers in that condition.

In the third model (examining partisan extremity in place of participant group, across response measures), an interaction emerged between tweet type and partisan extremity, b = -0.08, SE = 0.01, t(3219) = -7.47, p < .001. Extremists preferred dismissing, b = -0.33, SE = 0.04, t(2645) = -7.25, p < .001, whereas moderates preferred engaging, b = 0.16, SE = 0.04, t(2562) = 3.68, p < .001. Moreover, being more extreme predicted better responses to dismissing tweets, b = 0.09, SE = 0.02, t(3219) = 5.67, p < .001, but not engaging ones, b = 0.01, SE = 0.02, t(3219) = 0.67, p = .503. As in the main text, there was no interaction between tweet type and partisanship, b = -0.02, SE = 0.01, t(3219) = -0.81, p = .421.

In the fourth model (adding response measure as a variable to the third model), I again find results similar to what I report for the full sample: The key participant group × tweet type interaction was significant for approval (b = -0.07, SE = 0.02, t(3207) = -3.29, p < .001) and feelings (b = -0.16, SE = 0.02, t(3207) = -7.61, p < .001); however it was also significant (but smaller) for intentions to Like or Retweet (b = -0.04, SE = 0.02, t(3207) = -2.86, p = .004). In all cases, the direction of the interaction suggested frequent reactors showed more relative preference for dismissing than everyone else; see Table A3.4.

Table A3.4

Interactions and simple slopes for tweet type by extremity by measure, Study 5

Effect	*b*	*SE*	*t*	*p*
Intercept	-0.41	0.12	-3.56	<.001
Tweet type	0.37	0.09	4.29	<.001
Party affiliation	-0.04	0.04	-0.91	.363
Feeling toward Senator (vs. approval of tweet)	-0.15	0.16	-0.93	.351
Intent to Like and Retweet (vs. approval of tweet)	0.21	0.16	1.31	.191
Extremity	0.09	0.03	3.06	.002
Tweet type × party affiliation	0.01	0.03	0.24	.812
Tweet type × feeling toward Senator	-0.02	0.12	-0.20	.843
Tweet type × intent to Like and Retweet	-0.38	0.10	-3.65	<.001
Party affiliation × feeling toward Senator	0.00	0.06	0.01	.991
Party affiliation × intent to Like and Retweet	0.05	0.06	0.80	.423
Tweet type × extremity	-0.07	0.02	-3.29	.001
Tweet type × feeling toward Senator	0.09	0.04	2.32	.020
Tweet type × intent to Like and Retweet	-0.06	0.04	-1.47	.143
Party affiliation × approval of tweet vs. feeling toward Senator	-0.04	0.04	-1.00	.318
Party affiliation × approval of tweet vs. intent to Like and Retweet	-0.02	0.04	-0.52	.605

Extremity × approval of tweet vs. feeling toward Senator	-0.09	0.03	-3.07	.002
Extremity × approval of tweet vs. intent to Like and Retweet	0.03	0.03	0.95	.343

Note. The model had 3207 *df*. Positive simple slope coefficients reflect more favorable responses.

A.4.4 What mediates extremists' preference for engaging tweets?

Additional exploratory models examined why extremists liked not only (their own Senators') dismissing tweets but also (their own Senators') engaging tweets more than moderates did, albeit to a lesser degree. I tested mediation models using extremity as the IV, affective polarization, desires for party status, and endorsing compromise as simultaneous mediators, and responses to either engaging or dismissing as the DV (in separate models); see Figure A4.1 (numbers before the slash reflect responses to engaging, after the slash reflect responses to dismissing). Each mediator showed its a unique pattern; in particular I highlight here affective polarization: Extremists' greater preference for their own party over opponents accounted for their preference for dismissing tweets, but *also* explained (marginally; indirect effect $p = .096$) their preference for engaging tweets: This suggests they like (their own Senators') engaging tweets more than do moderates (or no less, in Studies S3 and S4) out of ingroup favoritism.

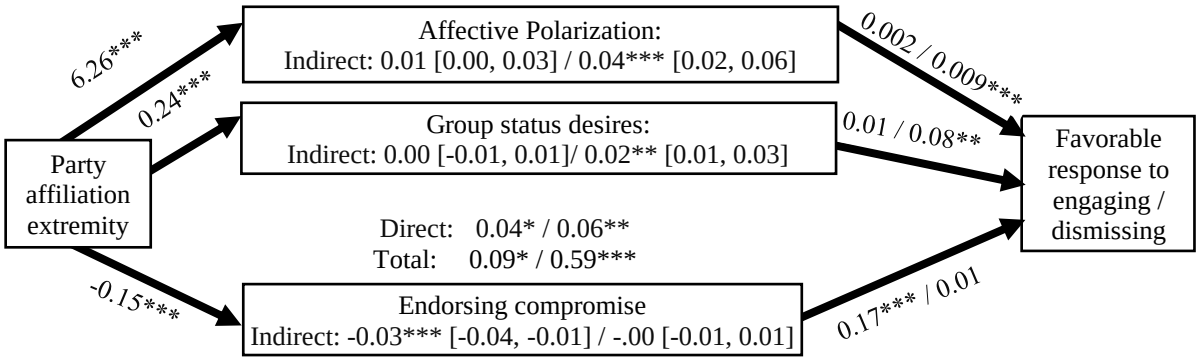


Figure A4.1 Explaining why extremists responded better to Senators' dismissing tweets.

A.5 Supplemental analyses for Study 6

Prior work suggests that most people should feel warmer toward Senators whose tweets engage with opponents; Study S1 replicated this prior pattern using artificial tweets. But in Study 5 and Studies S2-S4, which all used Senators' real tweets, most people felt similarly warm

toward Senators whose tweets dismissed versus engaged with opposing views (though Study 5 participants representing the majority did approve more of the engaging tweets themselves). This suggests that the tweets I created differed in some critical way from Senators' tweets I found in the wild. Thus, I collected additional data, measuring how Senators came across in the artificial tweets, to compare with my data from Study 6 (measuring how Senators came across in their real tweets).

A.5.1 Method

We recruited 81 participants from Prolific Academic such that roughly equal parts reported identifying as part of the Democratic or Republican parties or as Independents / other-affiliated. Following the same criteria as in Study 6, five participants who failed an English comprehension check were ejected from the study and I excluded one who self-reported providing low-quality data, leaving 74 (age M = 40.1; 49% women, 49% men, 2% nonbinary or trans).

As in Study 6, participants reported their party affiliation as part of a demographics form, which determined whether they would see tweets written by Democratic or Republican Senators. Each participant then saw and rated two of the 23 tweets I created for Study S1, one engaging and one dismissing. These participants rated only two tweets (compared to Study 6's ten) because the artificial tweets were made up of variations on a smaller set (e.g., similar language describing different issues or coming from the opposite side of the same issue).

Beneath each tweet, participants rated the Senator who wrote the tweet on the same five attributes, and using the same scale, as in Study 6. After this, participants reported whether I should use their data.

A.5.2 Results

We compared granular perceptions of Senators' real tweets to the artificial ones I generated, allowing me to identify similarities between them as well as differences that might explain the somewhat different results they produced. I tested five multilevel models, each predicting one of the five attributes from tweet type (dismissing = 0; engaging = 1), tweet source (artificial = 0, real = 1), their interaction, and random intercepts for participant and tweet. For key results, see Table A5.1 and Figure A5.1. As expected, in both tweet sources, engaging (compared to dismissing) tweets made the Senator who posted them seem more tolerant, cooperative and rational, but also more legitimizing of opponents' views and willing to change their mind. Thus, the two types of tweets I generated and the two types of tweets I coded differed along the same broad dimensions.

Table A5.1

Differences in attribute perceptions, engaging vs. dismissing / real vs. artificial tweets

Attribute	*Effect*	*subset*	*b* (*SE*)	*t*	*df*	*p*
Tolerant	Type × source	All tweets	-0.78 (0.14)	-5.47		< .001
	Engaging vs. dismissing	Real	1.26 (0.04)	28.48		< .001
		Artificial	2.04 (0.13)	15.12	2616	< .001
	Real vs. Artificial	Engaging	-0.21 (0.18)	-1.17		.241
		Dismissing	0.57 (0.17)	3.27		.001
Cooperative	Type × source	All tweets	-0.56 (0.14)	-4.02		< .001
	Engaging vs. dismissing	Real	1.28 (0.04)	30.21		< .001
		Artificial	1.84 (0.13)	13.82	2609	< .001
	Real vs. Artificial	Engaging	0.02 (0.17)	.11		.910
		Dismissing	0.58 (0.17)	3.51		< .001
Rational	Type × source	All tweets	-0.86 (0.14)	-6.24		< .001
	Engaging vs. dismissing	Real	0.50 (0.04)	11.90		< .001
		Artificial	1.36 (0.13)	10.38	2603	< .001
	Real vs. Artificial	Engaging	-0.09 (0.17)	-0.50		.620
		Dismissing	0.78 (0.17)	4.54		< .001
Legitimizing	Type × source	All tweets	0.12 (0.15)	0.81		.418
	Engaging vs. dismissing	Real	1.58 (0.05)	33.59		< .001
		Artificial	1.46 (0.14)	10.56	2614	< .001
	Real vs. Artificial	Engaging	0.16 (0.19)	0.87		.383
		Dismissing	0.04 (0.18)	0.24		.814
Open to changing	Type × source	All tweets	-0.33 (0.17)	-2.02		.043
	Engaging vs. dismissing	Real	1.00 (0.06)	17.44		< .001
		Artificial	1.34 (0.15)	8.64	2616	< .001
	Real vs. Artificial	Engaging	-0.03 (0.23)	-0.12		.901
		Dismissing	0.31 (0.23)	1.34		.182

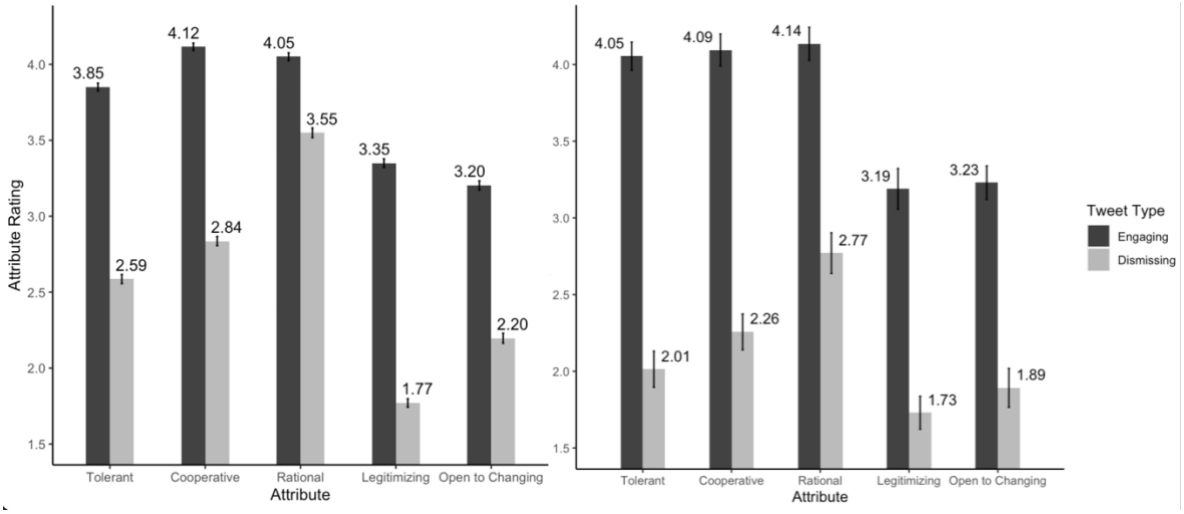


Figure A5.1. Trait ratings for real (left) and artificial (right) engaging and dismissing tweets

At the same time, there were some important differences: On all attributes except legitimizing, the difference between engaging and dismissing tweets was larger for artificial than real tweets. This was because, whereas the two kinds of engaging tweets were largely similar, the real dismissing tweets made Senators seem more attractive—that is, more rational, cooperative, and tolerant—compared to artificial tweets. As I found in Study 6, most people quite like others who are rational, cooperative and tolerant (e.g., Fiske et al., 2007; Heltzel & Laurin, 2021). This may help explain why, in my studies that used real tweets, most people felt just as warmly toward dismissing Senators as engaging ones, despite preferring engaging in Study S1 and in prior work (and despite approving more of engaging tweets themselves than dismissing ones in Study 5).

A.6 Supplemental analyses for Study 8

Following my preregistration, I first replicated the same key analysis from Study 7, setting aside data from the control condition. The predicted interaction was again significant (see Table A6.1 for simple slopes). Again replicating Chapter 2, participants themselves preferred the actor who engaged constructively with, rather than dismissed, opposing political perspectives. Consistent with misperceptions, participants assumed their allies would prefer dismissing. This

pattern is somewhat different from Study 7 (where participants had simply assumed allies would have no preference), but the overall presence and direction of the misperception was the same.

Table A6.1

Results of simple slopes tests, Study 8 supplemental tests.

Simple slope effect	*b (SE)*	95% CI	*t*	*p*	*d*
Interaction, engaging vs. dismissing	14.60 (4.13)		3.54	< .001	
Own response to engaging vs. dismissing	7.28 (3.03)	1.32, 13.23	2.40	.017	0.23
Perceived ally response to engaging vs. dismissing	-7.32 (2.80)	-11.24, 4.63	-2.62	.009	-0.25
Own vs. perceived ally response to engaging	13.20 (3.05)	7.22, 19.19	4.33	< .001	0.42
Own vs. perceived ally response to dismissing	-1.39 (2.78)	-6.86, 4.07	-0.50	.617	-0.05

Note. *b* refers to the beta for the difference between the two targets. *SE* represents standard error. There were 432 *df.*