**Designing for Safe, Fun and Informative Online Cross-Partisan Interactions**

by

Ashwin Rajadesingan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor Ceren Budak, Co-Chair
Professor Paul Resnick, Co-Chair
Professor Chris Bail
Associate Professor Brian Weeks

Ashwin Rajadesingan
arajades@umich.edu
ORCID iD: 0000-0001-5387-1350

# DEDICATION

I dedicate this dissertation to Oviya, Deepika, Siddharth and my parents, Rajadesingan and Aruna.

# ACKNOWLEDGMENTS

My graduate school experience, like many others', has been a mix of highs and lows. I have enjoyed the thrill of discovering something new and impactful while also being riddled with self-doubt and imposter syndrome. Thankfully, through it all, I have been lucky to lean on a wonderful crew of advisors, mentors, family and friends. I am extremely grateful to my advisors Ceren Budak and Paul Resnick, who have donned multiple hats through the years as exceptional mentors, role models, advocates and occasional confidantes. I will miss our weekly discussions about research and life in general, but I'm glad to have notes from over 200 such discussions to reflect on. I hope to imbibe their thoughtful, attentive-to-detail approach to research in my future endeavors. I also thank Brian Weeks for being my go-to person for all things communication studies, from job search advice and mentorship to providing invaluable feedback on my research over the years and especially, for serving on my dissertation committee. Many thanks also to committee member Chris Bail for providing insightful perspectives on both interaction design and social science research which helped immensely in shaping the dissertation.

Conducting interdisciplinary research as a graduate student is especially complicated as you must collate and tap into multiple research areas from scratch. I am thankful to have been part of the Political Communication Working Group (PCWG) at the University of Michigan, a fantastic interdisciplinary supportive group of scholars working on political communication research. I learned so much from the group! I also appreciate the faculty, administrators and staff at the School of Information (UMSI) for structuring the PhD program to provide the space for me to explore and experiment with interdisciplinary research. At UMSI, I also enjoyed and learned from research presented by colleagues at the Computational Social Science Seminar. Special thanks also to Kelly Garrett, my long-term political communication collaborator without whose advice and mentorship, I probably would not have got my dream job! I also thank David Jurgens for his support, especially during the early years of my PhD, co-writing my first PhD-level paper, which was also my first major interdisciplinary work. I am also grateful to have leaned on political communication and political science colleagues such as Ariel Hasell, Josh Pasek, Mara Ostfeld, Annie Franco and Pablo Barberá for advice and mentorship at different points of my PhD. I have also been fortunate to be a Facebook Fellow, which provided generous funding to explore multiple methods and theoretical approaches.

Doing a PhD is hard, even more so when labeled a Resident Alien by the US government. I am thankful to have had Ram Mahalingam and Joyojeet Pal as academic mentors who shared some of my lived experiences and to whom I could also relate at a personal level. They made me feel at home in Ann Arbor, especially in the initial years. I am also indebted to friends Vivek Veeriah, Vaishnav Kameswaran, Mauli Pandey, Chris Quarles, Sam Carton, Eshwar Chandrasekaran, Anmol Panda, Siqi Wu and so many others who have made PhD life fun and meaningful. I also had the absolute pleasure of working with many undergraduate students at Michigan. I thank Carolyn Duran, Daniel Choo, Mia Inakage and Jessica Zhang, who have all contributed majorly to my dissertation work. I also thank Jason Baumgartner for amazing his work building and maintaining the PushShift Reddit dataset and API, which saved me months of painstaking data collection time.

Words cannot adequately describe my wife Oviya's contribution to the PhD. She has been my intellectual sparring partner, my biggest cheerleader and my strongest support system. My parents, Rajadesingan and Aruna, my sister Deepika and nephew Siddharth have all contributed to and supported my PhD in so many tangible and intangible ways. My dad's effusive enthusiasm, my mom's can-do spirit and my sister's perceptive listening skills pulled me out of many mind-numbing research rabbit holes I found myself in. Even 11-year-old Siddy, my first playtester for the GuesSync game, contributed by giving me useful advice on how to make the game more fun! Simply put, I could not have hoped for a more supportive family for this journey.

# TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE

# LIST OF TABLES

TABLE

# ABSTRACT

The past decade in the US has been one of the most politically polarizing in recent memory. Increasingly, ordinary Democrats and Republicans fundamentally dislike and distrust each other, even when they agree on policy issues. Most Americans report believing that the opposing party is a "serious threat to the United States and its people". This extreme partisan hostility has wide-ranging consequences, even affecting how partisans respond to COVID-19 mitigation measures. In this context, this dissertation aims to reduce hostile interactions and attitudes towards ordinary Democrats and Republicans. I argue that we can reduce hostility by leveraging nonpolitical online spaces that cut through the partisan faultlines in uniquely engaging ways. I develop approaches to transform the currently hostile, uninspiring nature of online political interactions into not only a safe experience but also a fun and informative one. I take a mixed-methods approach to studying outpartisan hostility, combining computational social science with design methods. The dissertation progresses from a large-scale exploratory analysis of online political discussions to developing potential designs to reduce online partisan hostility and, finally, to designing and evaluating a fun party game that reduces outparty hostility.

In the first study, through large-scale computational analysis of billions of Reddit comments, I find that nearly half of all political discussions on Reddit take place in nonpolitical communities and that cross-partisan political conversations in these communities are less toxic than those in explicitly political communities. These findings suggest that shared nonpolitical interests can temper online partisan hostility. In the second study, through in-depth qualitative interviews and design probes, I explore approaches to surface these nonpolitical interests and identities during online political interactions on Reddit. I demonstrate that participants are comfortable knowing and revealing shared memberships in nonpolitical communities with outpartisan discussion partners which they expect to be humanizing, potentially reducing the hostility in those interactions. Through the interviews, I find that apart from serious deliberative discussions, participants also engage in light-hearted and casual political interactions where the motivation to simply entertain themselves and have fun. In the final study, drawing on insights from the prior study and extant political science research, I develop an online party game that combines the relaxed, playful non-partisan norms of casual games with corrective information about Democrats' and Republicans' political views that are often misperceived. Through an experiment, I find that playing the game

significantly reduces hostile attitudes toward outparty supporters among Democrats.

Overall, this dissertation demonstrates the potential of using nonpolitical context to facilitate quality online cross-partisan interactions that account for and mitigate the heightened levels of partisan animosity we observe today.

# CHAPTER 1

# Introduction

Over the past decade, the political climate in the US has been fraught with hostility and distrust. In fact, compared to other countries with long-standing democracies, the US ranks number one in affective polarization [24], a measure of how much partisans view copartisans favorably and outpartisans negatively [96]. Individuals' attachment to their party of choice is motivated not by an appreciation for their party but because they loathe the other party [93]. The impact of outparty hostility is felt far beyond the corridors of power in Washington as ordinary Democrats and Republicans now fundamentally dislike and distrust each other. Partisans routinely view supporters of the other party as cold, unpatriotic and closed-minded [51]. They are hostile towards the other party even when they agree on policy issues [141]. In social settings, they are reluctant to spend time with each other [34] or even talk about largely apolitical topics like music with each other [204]. Affectively polarized individuals also significantly contribute to the misinformation ecosystem; they are more likely to construe fake news as being political vendetta [220], consume more news from untrustworthy websites [78] and share more fake news online [170]. Most alarmingly, partisan animus politicizes ostensibly neutral or nonpartisan issues such as the COVID-19 pandemic [54] and impacts how partisans respond to mitigation measures, potentially resulting in excessive deaths [53].

Naturally, this hostility towards outpartisans shapes how Republicans and Democrats engage with each other online. About 70% of social media users find talking politics with someone they disagree with stressful and frustrating [7]. Worryingly, free-wheeling casual cross-partisan political interactions that are common on social media may exacerbate existing partisan faultlines. About 72% of users report finding out that they have less in common politically (than what they expected) with those they disagree with after engaging in social media political discussions with them [7]. Further, research suggests that exposure to outpartisans online, who are typically more extreme than the modal outpartisan, further exacerbates negative feelings towards outpartisans [13]. Thus, given the contentious political climate, most people understandably throw in the towel and refrain from engaging in online political interactions altogether. However, as deliberation scholars have demonstrated, informal political talk, such as discussions on social media, plays a crucial role in

supporting a deliberative democracy by fostering mutual understanding, political tolerance, and public-spiritedness [39]. Further, these interactions act as a rehearsal or training ground to empower citizens to voice their views and concerns in more formal deliberative forums in the future [40, 196]. Thus, opting out of political interactions may incur opportunity costs as well.

Therefore, in this dissertation, I aim to reduce hostile interactions and attitudes towards ordinary Democrats and Republicans. I argue that we can reduce both forms of hostility by leveraging non-political online spaces that cut through the partisan faultlines in uniquely engaging ways. Through in-depth qualitative interviews and design research, I develop approaches to transform the currently hostile, uninspiring nature of online political interactions into not only a safe experience but also a fun and informative one.

## 1.1 Reducing outparty hostility by leveraging nonpolitical spaces and identities

One explanation for the prevalent hostility is that by merely categorizing individuals into groups (here, Republicans and Democrats), group identities are activated, creating an 'us' versus 'them' group dynamic [216]. While partisan group behavior has always been a fixture of American politics, scholars suggest that this partisan psychology combined with selective partisan media coverage [124], negative campaigning [95] and a well-sorted electorate (fewer Conservative Democrats and Liberal Republicans) [123] have driven partisan hostility to extreme levels. These group-motivated behaviors may even be exacerbated in online spaces. The Social Identity model of Deindividuation Effects (SIDE) posits that the visual anonymity of online platforms increases the salience of group identities and adherence to group normative behavior because of the lack of individuating information about members from the group [183]. Following Self Categorization Theory, in the presence of an accessible group identity, individuals become depersonalized and view themselves and others less as individuals having distinct personalities but instead as interchangeable group members [222]. Although commonly used to explain behavior in intragroup contexts, similar group dynamics have also been observed in intergroup contexts. Through a series of experiments in intergroup online settings where participants were anonymous except for their group membership labels, research [178, 177] has shown that the depersonalization predicted by the SIDE model increased the relative salience of group boundaries and led to stereotyped perceptions of the outgroup. An important caveat is that group identity was accessible and salient during these intergroup interactions. However, in the context of social media interactions, prior research overwhelmingly points to the salience of partisan identities [203].

Since outparty hostility stems more from identity-related differences, focusing on nonpolitical

interests and identities to reduce the salience of partisan identity is a promising strategy. Social psychologists studying intergroup conflict have applied similar approaches by categorizing group members based on shared and superordinate social identities between group members to improve intergroup relations [45]. More recently, in the realm of politics, Levendusky [128] experimentally found that individuals exhibited significantly higher warmth (by over 20%) towards outpartisans when they were identified as supporting the same football team compared to no team identification. In another recent study [14], researchers found that participants matched on multiple nonpolitical "*soft* common features" (emphasis in original) such as hobbies, food choices and movies had increased feelings of closeness towards the matched individual and in turn, predicted openness toward the opposing views that the matched individual held. Finally, after the 2016 presidential campaigns left behind an angry and bitterly divided electorate, multiple initiatives such as the Weave project[1] have been started aiming to heal the divide by highlighting common, nonpolitical or at least apolitical ties that bind Americans. Such initiatives primarily center around bringing people together through facilitated dialogue, building a culture of listening and developing a shared sense of belonging. While these approaches facilitate high-quality interactions between opposing partisans, they are not directed at reducing hostility in the casual everyday political interactions that people already have currently online. In this context, this dissertation complements these approaches and contributes to reducing hostile attitudes and behavior by facilitating quality online political interactions in a few different ways as follows:

1. In Chapter 2, through a large-scale computational analysis, I explore how users in nonpolitical online spaces such as music and hobby communities engage in incidental political discussions that take place there and, in doing so, likely lean into their shared sense of community membership and the communities' casual conversational norms. I find that nearly half of all political discussions on Reddit take place in nonpolitical communities and that cross-partisan political conversations in these communities are less toxic than those in explicitly political communities. These findings suggest that shared nonpolitical interests can temper online partisan hostility. In the next chapter, I examine how to incorporate such hostility-reducing nonpolitical information into discussions that take place in explicitly political online communities.

2. In Chapter 3, through semi-structured qualitative interviews and design probes, I explore approaches to surface nonpolitical interests and identities shared by ordinary Democrats and Republicans during an online political interaction on Reddit. I demonstrate that participants are comfortable knowing (and revealing) shared memberships in nonpolitical communities with outpartisan discussion partners and expect that information to be humanizing and po-

[1] https://weareweavers.org/

tentially reducing the hostility in those interactions [2]. I also find that not all discussions participants engage in are serious and deliberative. Participants also engage in light-hearted and casual political interactions where the motivation to simply entertain themselves and have fun. In the next chapter, I build on these findings and incorporate a hostility-reducing intervention in a fun and engaging party game.

3. In Chapter 4, drawing on insights from the prior study and extant political science research, I propose an alternate design for online political interactions that combines the relaxed, playful norms of nonpolitical casual games with corrective information about Democrats' and Republicans' political views that are often misperceived. Through an experimental study, I find that playing the game significantly reduces hostile attitudes towards outparty supporters among Democrats. Importantly, comparing versions of the game that contained varying levels of political content, I find that adding more political information to the game did not significantly change how fun and engaging the players thought the game was, providing cause for optimism for future iterations of games that combine politics and play.

Overall, I take a mixed-methods approach to study outpartisan hostility, combining computational social science with design methods. The dissertation progresses from a large-scale exploratory analysis of online political discussions in Chapter 2 to developing potential designs to reduce hostility in cross-partisan interactions in Chapter 3 and finally, to designing and evaluating a fun party game that reduces outparty hostility in Chapter 4.

## 1.2 Facilitating Safe, Fun and Informative online political interactions

As discussed earlier, throughout the dissertation, I focus on making political interactions online less hostile. Yet, less hostile interactions are not necessarily great interactions. In my interviews with people who interact with opposing partisans online, participants expressed their many motivations for participating in these interactions (Section 3.4.1). Two, in particular, stood out to me: (i) participants looked for informative discussions which did not contain unverifiable claims or simply party talking points, (ii) since participants were online primarily to have fun and entertain themselves, they also looked for casual and fun engagement, even in political interactions. Therefore, I focused on designing to facilitate safe interactions that were both informative and fun. Safe, fun and informative are admittedly more modest aspirations compared to the traditional ideals of

[2]However, women and minority participants are skeptical about sharing other information such as membership in other (non-shared) nonpolitical communities and their past comments, as that information might be used to fuel personal attacks and disparage them.

deliberation such as reason-giving, equality, mutual respect and consensus-building [3][10] . But my objective is not to make casual online interactions more deliberative but to facilitate cross-partisan interactions that users would want to participate in. As Mansbridge and colleagues conceptualize in their foundational work on *deliberative systems* [137], political talk takes place across multiple sites such as homes, forums, schools and legislatures. There is a division of labor between the different sites of the system. Each site need not uphold all deliberative ideals, and the quality of deliberation should be considered as a whole. For example (paraphrased from [137]), interactions between activists in social movements are usually extremely partisan and closed to dissenting ideas. But this kind of interaction is conducive to fostering counter-hegemonic ideas, which may feed into public discourse and eventually shape an eventual democratic decision taken in more formal deliberative forums such as legislatures. In such a deliberative system, casual political talk, even without it being particularly deliberative, functions as a vital conduit to participating in more formal discussion forums later [196]. Next, I highlight past work on safe, fun and informative political interactions and how these qualities are explored in the dissertation.

### 1.2.1 Safe

*I tried to engage once with somebody that vehement, and they just were, they just attacked. It was like, you know, it was like getting a text that's like, all caps from your mom. And it's just, you know, who needs that? - Beth, interview participant*

I use 'safe' to mean being able to engage in online political interactions without becoming the target of hostile personal attacks, toxic name-calling, and harassment. According to a Pew survey, nearly 70% of social media users rarely or never post anything political on social media, and a significant reason they cite is not wanting to be attacked for their views (27% Democrats and 36% Republicans) [146]. This reasoning was evident in my interviews with Reddit users such as Beth[4], a New York high school teacher in her mid-40s, who described a particularly harsh exchange (see quote above) and how she had since become more selective about whom she engaged with. Therefore, designing for safe interactions is a theme that resonates throughout the studies discussed in this dissertation.

Researchers have developed numerous platform-based solutions to reduce hostility online by targeting offending posts [97], users [98], and even entire communities [30]. These approaches have reduced caustic online interactions to some extent. I focus on a different issue. Rather than minimizing the number of hostile interactions, I focus on turning average or good quality interactions into great ones. In other words, the interventions proposed in this dissertation are not aimed

[3]Theorists also expanded these ideals to include other forms of communication such as storytelling, humor, rhetoric and other more inclusive communicative practices [242].

[4]name changed to protect identity

at bad actors who are determined to engage in personal attacks but at individuals who want to engage with others' views in good faith but have not been given the necessary support to overcome the partisan faultlines that have come to dominate politics today. Further, as prior studies suggest, even ordinary people can sometimes engage in troll-like behavior in online interactions depending on their mood [35] or the local community norms [180], carefully designed interventions can help reduce such occurrences.

Hostility against outpartisans is motivated more by partisan identity than policy disagreements [50]. It is driven by multiple group-oriented factors such as maintaining positive group status [88], responding to threats to group status (such as during election time) [149] and burnishing in-group credentials [182]. Also, hostility expressed online is most among individuals who are most affectively polarized [181]. Thus, I focus on approaches that manage the salience of partisan identity by either highlighting alternate common identities (in Chapter 3) or by not cueing partisan identities entirely (in Chapter 4). Since this us-vs-them orientation also results in partisans harboring exaggerated views of the outpartisans leading to more hostile attitudes towards them, in Chapter 4, I introduce corrective information that provides an accurate view of the other side within the interaction context itself. Lastly, unlike protected attributes such as race, where group-related behaviors are moderated by strong social norms and laws against discrimination, no such norms temper partisan hostility. However, my research on toxicity norms in political communities on Reddit indicates that [180] users very quickly adopt the local conversational norms of that community, resulting in some communities being able to maintain low toxicity levels even during the 2016 US presidential elections. This suggests that local community norms can potentially temper hostile partisan norms. I explore this phenomenon in Chapter 2, where I demonstrate how cross-partisan political discussions in nonpolitical communities are less toxic than those in explicitly political spaces, likely because the largely cordial nonpolitical communities' norms temper hostile partisan norms.

### 1.2.2 Fun

Fun is a relatively uncommon adjective to describe political interactions, especially in today's hyperpartisan political climate. Yet, for generations of Americans fed with late-night political satires such as The Colbert Report and The Daily Show, that politics, even the hostile grab-your-throat kind, can be a source of fun and entertainment is not entirely far-fetched. Moreover, most online political interactions take place on social media platforms where most users engage for fun, to relax, to entertain themselves and pass the time [228]. Also, many people do not actively seek political content online; instead, their encounters with political content online are often incidental [106]. Their Facebook newsfeed or Twitter timeline just happens to include political content. Thus, users engage with political content often, not intending to have deep deliberations but to casually

engage in discussions like they do on other topics before moving on. In my interviews in Chapter 3 with Reddit users, participants describe how they often engage with political content on Reddit when they take a break from work, when they eat, or when they commute. They simply scroll through the comments in those situations, occasionally posting replies but leaving the conversation without serious engagement. This lightweight, casual engagement parallels the increasingly common practice of online news snacking, where users spend short bursts scanning headlines or blurbs without intending to engage deeply with the topic [44]. My interviews also indicate that participants derive fun by creating memes, joining in casual banter and posting witty rejoinders in their everyday political interactions. Drawing on these insights, in Chapter 4, I design a game that experiments with making this kind of lightweight engagement with political content fun and engaging but also informative and persuasive.

Note that not all fun is harmless, though. Scholars have studied how memes, through ironic humor, play a central role in normalizing hateful content, introducing the uninitiated to a pathway towards alt-right radicalization [157]. This medium of satire and edgy memes mainstreams racism and antisemitism while crucially providing plausible deniability. It allows "people to disclaim a real commitment to far-right ideas while still espousing them" (Marwick quoted in [229]). Similarly, studying the right-wing BJP party's vast network of Twitter volunteers in India, Udupa highlights how the visceral aspects of fun mobilize right-wingers to participate in the painstaking task of fact-checking and contesting mainstream media narratives, archiving content for future confrontations, making Hindu nationalist hashtags trend and aggressively ridiculing anti-Hindutva views [223]. In Chapter 4, I try to channel the innocuous fun experienced when playing a party game into meaningful, informative political interactions.

### 1.2.3 Informative

A significant outcome of deliberation touted by deliberative theorists is knowledge gained from interacting with outpartisans [11]. Extensive empirical research on heterogeneous offline discussion backs this claim [58, 184]. Online, outside of facilitated initiatives such as deliberative polling [63] and mini-publics [12], it is unclear if everyday cross-partisan political interactions with strangers on the internet in this hyperpartisan environment produce gains in political knowledge.

Increasingly, Republicans and Democrats trust different news sources and worryingly, Republicans trust few mainstream sources [74]. Further, the Republican party elite such as Donald Trump continues to peddle misinformation about critical issues such as voter fraud and ballot tampering. Thus, many ordinary Republicans and Democrats disagree on even basic facts. The impact of political misinformation is exacerbated as individuals primarily share and spread fake news motivated primarily by partisan goals [170]. When deciding to share a fake story that derogates the other

side, individuals do not appear to be concerned if the story is true or not, driven by their animosity towards political opponents.

Another kind of misinformation that partisans inadvertently harbor is in their perceptions about outpartisans (meta-perceptions[5]). Republicans and Democrats believe that the outparty supporters are more extreme than they actually are, a phenomenon called perceived or, more lately, misperceived polarization [119]. Each group considers the other to be more ideologically extreme and politically engaged [52], hold more negative opinions about them [118] and be more supportive of violence against them [148]. This exaggerated perception reduces opportunities to build common ground, increasing prejudice and dehumanization of the political outgroup [151]. In Chapter 3, interview participants described how viewing the interaction designs highlighting shared nonpolitical group memberships might help them view outpartisans more as individuals with a different political view rather than as caricatured political extremists. In Chapter 4, I explicitly correct inaccurate meta-perceptions through casual gameplay.

If this introduction has made you despair about online cross-partisan interactions, you are not alone in feeling this way. What purpose could this dissertation serve in facilitating online interactions between Republicans and Democrats when they have every incentive to be hostile to each other, when they do not even agree on basic facts, and when their conversations are not even serious? The practical, straightforward answer is that people already engage in these interactions, and we cannot prevent people from engaging with each other. Left to its own, this form of engagement will likely sow more distrust and discontent among the electorate [13]. So, we may as well intervene to make these interactions a bit better. The other answer is that I believe interaction designs have not kept pace with the current reality of increasing outparty hostility. We are not providing the tools for users to have quality cross-partisan political interactions in this hostile environment. For example, political discussion communities on mainstream platforms such as Facebook and Reddit have the same design as communities that discuss sports or food choices. But, studies show that even a simple button label change from 'Like' to 'Respect' can positively impact how users engage with counter-attitudinal political content [212]. Designs prioritizing active listening and reflection rather than privileging speaking can make users pause and provide more thoughtful empathetic responses [112, 113]. Layouts of the discussion systems can also provide de-stereotyping information about outpartisans to reduce partisan cues [103, 59]. Small-scale, one-on-one anonymous interactions can also improve the quality of interactions by lowering incentives to derogate individuals to reaffirm their group status [13]. The design space for potential improvements is vast, and these enhancements can meaningfully impact cross-partisan relationships and interactions.

Lest a gusto of technological determinism sweeps us, I acknowledge that design is not des-

[5]Like [148], I use meta-perceptions to mean perceptions of others' perceptions rather than [118]'s usage which refers specifically to how group members think the outgroup members perceive them.

tiny. A better-designed interaction interface will not magically transform online spaces into sites of democratic deliberation. There are numerous structural reasons for the prevailing hostility between ordinary Democrats and Republicans, including deep differences in ideology [185] and moral values [68], a fusing of partisan, religious, ideological and racial identities [142], historic levels of inequality [72] and racial resentment [1]. These factors naturally affect outparty attitudes and interactions, and a better design cannot solve these underlying concerns. Still, it can facilitate cross-partisan interactions around them, a necessary first step towards achieving political legitimacy for any potential policy solutions.

In the following three chapters, I describe three studies I conducted on this topic. Each study can be read independently of the others and contains its own related work that contextualizes its contributions. In the concluding chapter, I provide a conceptual summary of these studies, think through answers to some thorny questions about the dissertation and highlight multiple paths forward to design safe, fun and informative online political interactions that reduce hostile attitudes and interactions.

# CHAPTER 2

# Political Discussion Is Abundant In Nonpolitical Subreddits (And Less Toxic)

## 2.1 Introduction

Casual everyday political conversations are central to a vibrant deliberative democracy. Through these conversations, individuals learn new perspectives, form informed opinions and update their preferences [104]. These interactions may take place in explicitly political spaces such as city townhalls and civic committees but also in seemingly non-political spaces such as book readings, workplaces and social gatherings [38]. Importantly, this kind of everyday political talk is significantly correlated with opinion quality and political participation which are central to forming a well-informed electorate [237]. In this work, we explore this phenomenon online, particularly studying political discussions in communities on Reddit that are not explicitly political.

Most research on political discussions has primarily focused on explicitly political spaces, examining communities around political news groups, figures or ideologies [210, 85, 5]. However, survey research suggests that most people encounter political content online not in explicitly political spaces but in hobby and leisure groups where politics is incidental to the conversation [231]. Further, recent years have seen increased political engagement among the electorate perhaps due to high levels of partisanship [88] and growing social movements [22] such as Black Lives Matter. This heightened level of political engagement can also be observed online. For example, on Reddit, many communities that would not be typically construed as being 'political' such as r/EDM and r/MaleFashionAdvice protested against the platform's hate speech policies and police brutality in the US.[1] Further, in recent years, scholars have observed increasing politicization of typically non-political spaces [47]. The most prominent example is the politicization of emerging science and technology where inherent uncertainties are harnessed by political actors to cast doubt

[1] https://www.theverge.com/2020/6/3/21279601/reddit-dark-subreddits-protest-police-violence-racism-hate-speech-policies

on the existence of scientific consensus [21]. On Reddit, this phenomenon manifests in the formation of multiple communities on the same topic along partisan lines, for example, r/China_Flu and r/Coronavirus [244]. Thus, these developments call for expanding analysis of political discussions outside of typical political communities to communities that aren't explicitly political.

It is important to note that expanding the study of political discussions to include non-political spaces does not merely increase the volume of discussions for analysis. The dynamics of political discussions in these spaces may also be fundamentally different. Political discussions in these spaces may be moderated not by partisan identity but by participants' shared non-political interests and identities that drew them to the same community in the first place [67]. Thus, we might expect political conversations in non-political spaces, including cross-partisan ones, to be less toxic. However, shared non-political group identity may fail to offset, and might even exacerbate, the animosity generated by partisan identity [108]. Further, norms in these non-political spaces may not be designed to foster political discourse. Indeed, there may be norms against having political conversations at all, and thus when they occur, they may be even more toxic.

In this work, we focus on two primary questions: (i) What is the prevalence of political discussions in communities that are not explicitly political? (ii) Are cross-partisan political discussions in these spaces less toxic than ones in explicitly political spaces? We estimate that 49.26% $\pm 3.59\%$ of all political discussions on Reddit takes place in communities that host political discussions less than 25% of the time. This finding is not simply the result of a few very large non-political communities hosting some political content. It is instead due to a long tail of small communities that host some political content each. Our toxicity analysis reveals that political conversations in non-political spaces, including cross-partisan political interactions, are indeed less toxic than such interactions in political spaces. Interestingly, we find that there is an uptick in toxicity levels when talking politics, but even with this increase, the toxicity levels in non-political subreddits are still much lower than the toxicity in political subreddits.

## 2.2 Background

Political scientists have long highlighted the presence and importance of casual political talk in everyday social interactions taking place in spaces that are not explicitly political (see [38] for a review). In fact, research suggests that most political conversations take place at work or with neighbors, with more than 70% of American survey respondents reporting that they have never or only rarely even attended public meetings explicitly designed for political discussions [41]. Similarly online, early survey research suggests that most people encounter political talk in message boards and chatrooms designed not for political discussions but for hobby and leisure related discussions [231]. Thus, research limited to studying only political discussion spaces may overlook

other spaces where a significant amount of such interactions may be taking place. Such everyday political talk, although not always deliberative and conducive to rational-critical argumentation, have important positive outcomes such as increased political knowledge [173], political participation [197], refined opinions [104] and higher tolerance [173, 161].

Recently, scholars have analyzed political discussions taking place in online "third spaces" a term derived from sociologist Ray Oldenburg's conceptualization of the 'third place', referring to public spaces outside of work and home such as cafes, parks and libraries where people meet and interact informally, fostering community ties and political participation [234, 167]. Graham et al. [76] found that political discussions in the three UK-based non-political forums they analyzed were as likely to emerge from non-political, personally-oriented discussions as from discussions that were about politics from the start, with users explicitly linking their personal experiences to public policy. In contrast to discussions in political spaces, they found that the discursive culture in these discussions centers around help and support rather than being competitive and combative. Yan et al. [239] found that the political arguments made on transnational online cricket forums were typically short, unsubstantiated by external sources and occasionally uncivil. However, there was high exposure to cross-cutting political discussions with some engagement with opposing views in the form of question exchange and mutual acknowledgement. Analyzing a reality television discussion forum, Graham found that most political exchanges were driven by users' life experiences representing a more "lifestyle-oriented, personal form of politics" [75]. While exhibiting deliberative features such as the exchange of reasoned claims (as opposed to assertions) and reciprocity, participants also employed affirming, supportive and empathetic communicative practices fostering genuineness and civility in the discussions.

Political discussions in these non-political spaces may also be more civil and social compared to discussions in explicitly political communities. A significant factor contributing to hostility commonly observed in online political discussions is the increased levels of affective polarization [90], the tendency of partisans to view opposing partisans negatively and co-partisans positively [96]. This increased out-party animosity is explained by Social Identity Theory which argues that by merely categorizing individuals into groups (here, Republicans and Democrats), group identities are activated, creating an 'us' versus 'them' group dynamic [216]. Crucially, unlike race, gender and other protected attributes where group-related behaviors are mediated by strong social norms (and laws) against discrimination, there are no norms that temper hostility towards out-partisans [94]. In fact, the open hostility displayed by political elites towards their political opponents demonstrates that such behavior is appropriate [15]. Given the social identity underpinnings of affective polarization, researchers have explored ways to offset partisan identity drawing from prior research on intergroup conflict. One successful approach to reducing out-partisan animosity is by priming a superordinate identity. Based on the Common Ingroup Identity Model [67],

Levendusky showed that priming Republicans and Democrats to think of each other as Americans rather than outgroup members recategorized them as being part of the same common ingroup, resulting in reduced animosity and warmer attitudes towards each other [126]. Although our study is not a direct test of this theory, we expect that interactions in non-political subreddits likely increase the salience of shared common non-political group memberships. This may mediate how cross-partisan political discussions are conducted in these spaces.

Though not conclusive, the prior literature provides two compelling arguments: (i) political conversations are abundant in non-political spaces. (ii) quality of discourse in these conversations may be different and in some cases, better than political conversations in explicitly political communities. In this work, we assess these claims empirically in the context of Reddit. First, we quantify the relative contribution of non-political subreddits to the overall political content on Reddit. In this aspect, our work is similar to Munson et al.'s work on estimating the prevalence of political content in non-political blogs. Analyzing a sample of blogs from Blogger.com, they found that "25% of all political posts are from blogs that post about politics less than 20% of the time" [158]. Second, we examine a specific marker of conversation quality: toxicity in cross-partisan political interactions. Scholars have suggested that political talk in these third spaces are likely to be less polarized, since users participate in these spaces because of shared interests such as a soccer team or fast fashion which are not aligned politically [231, 235]. Thus, mediated by shared non-political identities, these spaces could facilitate respectful and civil cross-partisan interactions. In this work, we examine this hypothesis by quantifying the toxicity levels of cross-partisan political discussions in non-political spaces and comparing them to toxicity levels in other settings on Reddit.

## 2.3 Reddit Data

Reddit, a collection of communities of varied and diverse topics, provides us with an ideal platform to examine the prevalence of political content in non-political spaces. We use the PushShift Reddit dataset [17] to perform our analysis on comments posted from 2016 to 2019. We exclude comments from subreddits that have hosted less than 1000 comments over the four years. We also remove comments from known bots and moderators from the analysis. To allow for robust estimation of political prevalence, we only consider comments which are 50 characters or more in length in this analysis. In total, we examined 2.8 billion comments posted in 30,899 subreddits.

## 2.4 Estimating the Prevalence of Political Content in Non-political Spaces

Our basic approach to estimating the prevalence of political content is to train a classifier that yields, for each comment, a predicted probability that it would be judged as political by a panel of three MTurk raters. If the classifier's outputs are properly calibrated, the average of those outputs for all the comments in a subreddit is an estimate of the prevalence of political content in the subreddit.

Our training data consists of a sample of 10,000 comments, each rated by three people on MTurk as either political or not. We do not use these labels to directly train a classifier that predicts what how each item will be labeled. Following the quantification approach [66, 73], if the goal is to estimate prevalence rather than to correctly classify individual comments, it can be more effective to use ground-truth data to perform calibration on a crudely trained classifier than to use up the training data on improving the classifier.

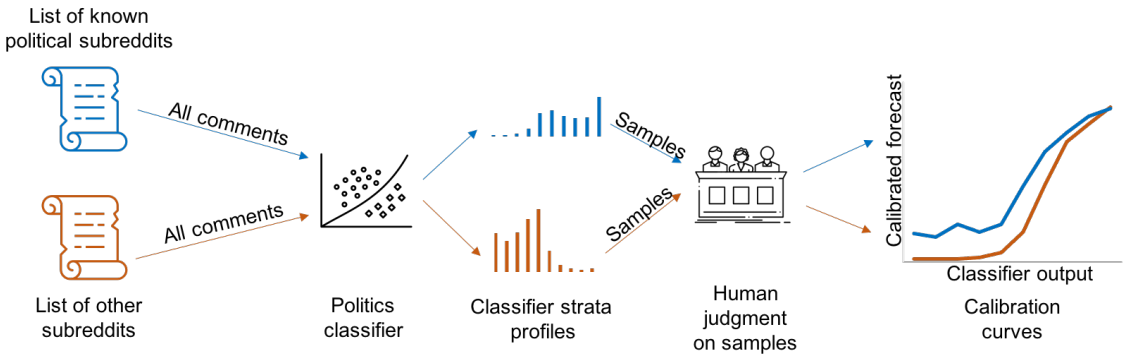
Section 2.4.1 describes a process for training a classifier to distinguish between comments from two baskets of subreddits, one of which consists of a hand-selected set of subreddits that are overtly political. This text-based classifier outputs a probability that the comment is from one of the political subreddits. The middle of Figure 2.1 shows the distribution of classifier outputs for all comments from the two baskets of subreddits. Some comments from the political subreddits contain phrases that are more common in the other subreddits. Such comments get a low score from the classifier. However, most comments originating in the political subreddits get higher scores (the blue distribution, on top) and most comments originating in the other subreddits get lower scores (the orange distribution, below).

The classifier does not have perfect accuracy. Moreover, it may make different kinds of errors on content from different subreddits. So we do not directly use it to classify and count political comments in each subreddit. The second step, as described in section 2.4.2, is to build two calibration curves for the classifier, one based on human ratings of a sample of comments from the set of known political subreddits and the other based on human ratings for comments from other subreddits. Each calibrator provides a mapping from a classifier output stratum (e.g., 0.5-0.6) to a calibrated forecast, the fraction of comments that are political when the classifier gives an output in that stratum. The right side of Figure 2.1 shows the two calibration curves. For all classifier outputs below 0.9, comments originating in the political subreddits were more likely to be judged as political, as indicated by the gap between the blue calibration curve and the orange one.

Section 2.4.3 then describes a process for selecting a calibrator to use for a particular subreddit. Section 2.4.4 describes how to use the calibrator to generate an estimate of the fraction of political content in that subreddit. Finally, section 2.4.5 describes how we combine the estimates for

individual subreddits to yield overall prevalence.

Figure 2.1: Training a classifier that distinguishes between comments from political and non-political subreddits, then calibrating it to produce predictions of whether comments are political.




### 2.4.1 Training a classifier

We built an L1-regularized logistic regression classifier trained on bigrams and trigrams from a random sample of 500,000 comments from known political subreddits (positive or "politics" class) and 500,000 comments from all other subreddits (negative or "not politics" class). We used a list of 277 political subreddits provided by [180] and updated the list to include more recently created subreddits supporting Democratic primary candidates such as r/YangForPresidentHQ and r/JoeBiden before the 2020 US presidential election.

Assessed through 5-fold cross-validation, we obtained an accuracy of 81.56% with a false positive rate of 14.41% and a false negative rate of 22.45%. Note that the false positive and false negative rates are for predicting the source of a comment, not whether the comment itself is truly political. While this classifier performed reasonably well in identifying content from political subreddits, it was trained not on political and non-political comments but on *comments from political and non-political subreddits*. This is particularly problematic since our goal is to estimate the fraction of *political comments* in non-political subreddits. If the classifier were perfectly accurate at distinguishing between content from the two types of subreddits, and we used it as if it were distinguishing political comments, it would tell us that there were zero political comments in non-political subreddits, which might or might not be the case. Thus, a further calibration step is needed in order to estimate the error rates of this classifier at predicting whether a comment is truly political, and then adjust for those error rates in our prevalence estimates.

Table 2.1: Population proportions, Neyman samples, percent of samples identified as political by human judgment and standard deviation per strata for political and non-political subreddits.

| Strata | Political subreddits | | | | Non-political subreddits | | | |
|---|---|---|---|---|---|---|---|---|
| | **Pop. proportion** $W_{k,\text{pol}}$ | **Samples** | **Political%** $p_{k,\text{pol}}$ | **Std dev.** $s_{k,pol}$ | **Pop. proportion** $W_{k,\text{nonpol}}$ | **Samples** | **Political%** $p_{k,\text{nonpol}}$ | **Std dev.** $s_{k,nonpol}$ |
| 1 | 0.004 | 50 | 0.180 | 0.054 | 0.148 | 615 | 0.021 | 0.006 |
| 2 | 0.007 | 50 | 0.160 | 0.052 | 0.117 | 797 | 0.024 | 0.005 |
| 3 | 0.017 | 50 | 0.240 | 0.060 | 0.150 | 1239 | 0.023 | 0.004 |
| 4 | 0.047 | 107 | 0.187 | 0.038 | 0.199 | 1810 | 0.031 | 0.004 |
| 5 | 0.145 | 346 | 0.237 | 0.023 | 0.242 | 2296 | 0.057 | 0.005 |
| 6 | 0.165 | 394 | 0.475 | 0.025 | 0.083 | 788 | 0.189 | 0.014 |
| 7 | 0.129 | 295 | 0.695 | 0.027 | 0.026 | 237 | 0.485 | 0.032 |
| 8 | 0.116 | 241 | 0.821 | 0.025 | 0.013 | 107 | 0.757 | 0.041 |
| 9 | 0.119 | 204 | 0.917 | 0.019 | 0.008 | 61 | 0.869 | 0.043 |
| 10 | 0.252 | 263 | 0.970 | 0.011 | 0.012 | 50 | 0.980 | 0.020 |

The standard deviations for both political and non-political subreddits are lower in the strata where their population proportions are higher. This effect is by design (using Neyman allocation) to ensure that the confidence intervals for the prevalence estimates are lower.

### 2.4.2 Building calibrators

The right side of Figure 2.1 outlines the calibration process. We use the classifier to produce a probability estimate for each comment. Then, we allocate comments into ten strata, 0-10%, 10-20%, etc., based on the classifier outputs. That yields a profile of classifier strata: the proportion of comments that fall into each stratum. We compute two separate classifier strata profiles, one based on comments from the list of known political subreddits, the other based on comments from other subreddits. As might be expected, the classifier assigns many more comments from the political subreddits to the higher strata (higher probability of being political).

Then, we calibrate the classifier outputs against human judgments of the comments, separately for comments from each source. Below we first describe the human judgment process and then explain the rationale and details behind each of the steps in the calibration process.

#### 2.4.2.1 Human judgments

When asked to identify topics they considered political from a list, Fitzgerald [64] found systematic demographic differences with partisans, liberals and men identifying significantly more topics as political compared to non-partisans, conservatives and women respectively. In order to reduce such differences in interpretation, Fitzgerald suggests providing human raters with an explicit definition to follow.

For this work, we modify [154]'s political discussion definition, which is predominantly based on political issues, to also include references to political figures, parties and institutions. We consider a comment to be political if it is about (i) political figures, parties and institutions, (ii) Broad cultural and social issues (e.g., civil rights, moral values, and the environment), (iii) National issues (e.g., healthcare, welfare policy, and foreign affairs), (iv) Local and state concerns (e.g., school board disputes and sales taxes) or (v) neighborhood and community affairs (e.g., decisions about a neighborhood watch crime prevention program).

Even with an explicit definition, however, whether a particular comment is political or not is open to interpretation. Conceptually, we take the ground truth classification of a comment to be the label that the majority of people who ever read online comments would apply to that comment, if they all were asked to judge it according to the explicit definition. Of course, this ground truth is a counterfactual; no such survey of all readers of comments can ever be conducted for any comment. Instead, we rely on a proxy for this ground truth, a survey of three raters on Amazon Mechanical Turk. To elicit high quality labels, we limit the task to crowdworkers with high performance in prior tasks [2] who also correctly labeled at least 4 out of 5 items in a qualification task where raters

[2]Raters must be based in the US, previously completed 1000 tasks and have at least 98% acceptance rate on the tasks that they have previously completed.

were shown sample comments and were asked to identify if the comment was political or not. Such qualification tasks are shown to improve crowdsourcing label quality [29].

The inter-rater agreement score as computed by Krippendorff's alpha was 0.55. While an alpha score of 0.55 is just below the threshold used for conventional content analysis, this agreement is relatively higher compared to other cases of crowd coding (e.g. [133]). The most common outcome (66.85%) was for a comment to be unanimously labeled as not political. An additional 10.81% of comments were unanimously labeled as political. The remainder were split decisions: 7.75% were labels as political by two of three raters and 14.59% by one of three raters.

Following common practice in treatment of crowd labels, our primary analysis treats a comment as truly political if two or three of the raters label it as such. Given the relatively low agreement among raters, for robustness, we also report analyses in the Appendix that treat a comment as political if any of the three label it as political, or only if all three label it as political. The different aggregation strategies produced largely similar results and provide additional informative bounds on our estimates.

#### 2.4.2.2 Classifier strata

We group comments into ten strata based on the classifier probability output. For example, stratum 1 consists of all comments whose classifier output is between 0 and 0.1; stratum 10 consists of all comments whose classifier output is between 0.9 and 1. By stratifying comments into multiple relatively homogeneous groups based on classifier probability, we require fewer samples to estimate true prevalence per stratum as within-group variance reduces in more homogeneous groups [37].

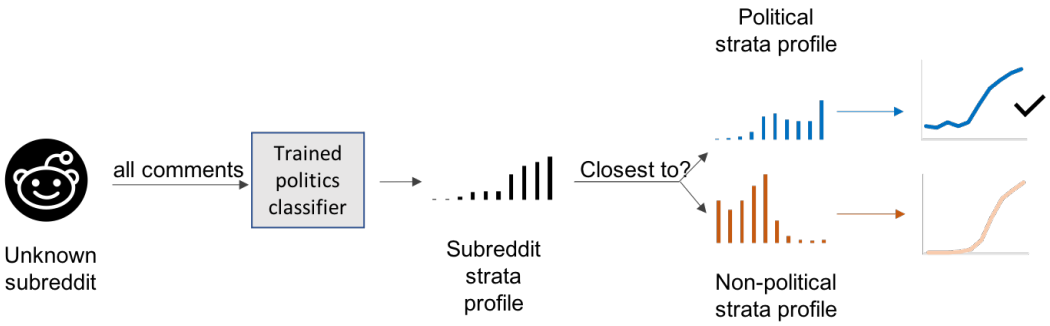
#### 2.4.2.3 One calibration for each subreddit type

We expect the per-stratum prevalence estimates to be different in different subreddits. That is, comments in the 60-70% stratum in political subreddits could be judged 70% of time to be political subreddits, while this number can be, say, 45% in non-political ones. This would not be a concern if we were trying to estimate the overall prevalence of political comments, since we could just estimate the classifier's error rates on a random sample of comments. However, our task demands accurate estimates of political prevalence in each subreddit; if the classifier is more prone to err on content from the stratum that originates in one subreddit than another, it would throw off our cumulative prevalence estimate.

Estimating separate per-stratum prevalence rates for each subreddit is practically infeasible as it would require human judgments for samples from each subreddit. Instead, we compute per-stratum error rates separately for each *subreddit type*: a sample of comments from known political subreddits and a sample of comments from other subreddits (same as those used to train the clas-

sifier).

Figure 2.2: Selecting the political or nonpoltical calibrator depending on if the subreddit strata profile is similar to the political or non-political strata profile




#### 2.4.2.4 Optimum strata sampling for human judgments

For each subreddit type, we must sample comments from each stratum for human judgments to obtain stratum specific prevalence estimates. The more comments we sample from a stratum, the less variance there will be in our estimate of the true prevalence of political comments in that stratum. Intuitively, fewer samples should be taken from a stratum with very few comments for calibration. A high variance in our estimated prevalence of political comments in such a stratum will not affect our overall estimate very much because it affects very few comments. Formally, for a fixed number of comments that we can afford to send for human rating, Neyman allocation provides the optimal allocation strategy which minimizes the variance of the overall prevalence estimate. Under Neyman allocation, the number of samples allocated to each stratum is given by:

$$n_k = n \frac{W_k * S_k}{\sum_{i=1}^{K} W_i * S_i}$$

- $n$ is the total number of comments to be rated
- $K$ is the number of strata (10 in our case)
- $n_k$ is the number of comments to sample from the $k$-th stratum
- $W_k$ is the weight of the $k$-th stratum in the classifier strata profile, i.e. the fraction of comments that are in that stratum
- $S_k$ is the standard deviation of stratum $k$.

Note that $S_k = \sqrt{P_k(1-P_k)}$ is unknown before sampling, where $P_k$ is the political prevalence in stratum $k$. Instead, we use our best estimate, the mean of the range limits of each stratum to calculate the approximate standard deviation expected in each stratum ($P_k = 0.05$ for Stratum 1, $P_k = 0.15$ for Stratum 2 and so on). Since our aim is to accurately estimate prevalence for each subreddit type, we perform separate stratified samplings, choosing two different $W_k$ for each stratum $k$ to match the overall comment proportions over strata (classifier strata profiles) for the two subreddit types.

We modified the Neyman allocation to include a minimum threshold to sample at least 50 comments in each stratum. We added this threshold to reflect the relative uncertainty in our initial estimates of $W_k$ and $S_k$. We had a total budget for rating 10,000 comments. We used $n = 2000$ comments from political and $n = 8000$ for other subreddits. The rationale behind the uneven breakdown is that there are likely fewer political comments in non-political subreddits, meaning that a similar sized confidence interval for both estimates would result in significantly larger levels of relative uncertainty for prevalence estimates in non-political spaces. The fraction of comments that fall into each stratum (the classifier strata profile) are shown in the "Pop. proportions" columns of Table 2.1 and graphically as a histograms in Figure 2.1. The number of comments selected per stratum are reported in the "Samples" columns of Table 2.1.

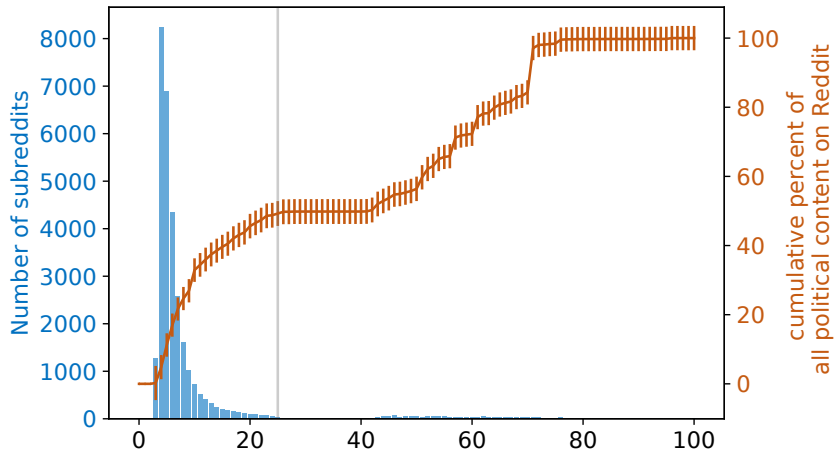
#### 2.4.2.5 Results of stratum-specific prevalence estimation using human judgments

The prevalence estimates per stratum from labeling comments for the two subreddit types are shown in Table 2.1 under "Political %" ($p_{k,\text{pol}}$ and $p_{k,\text{nonpol}}$). They are also shown graphically in the calibration curves in Figure 2.1, where x-axis is the stratum (classifier output) and y-axis is the calibrated prevalence in that stratum.

### 2.4.3 Selecting a calibrator

Given that the prevalence estimates for the same strata are quite different for the two subreddit types, it is important to determine which set of prevalence estimates to use for each subreddit. Figure 2.2 outlines this process. For each subreddit, we obtain the classifier probabilities of all its comments to build its strata profile. If the subreddit strata profile is more similar to the known political strata profile than to the other subreddits strata profile, we use the calibration curve of the known political subreddits, else we use the other calibration curve. We use the Jenson-Shannon divergence (JSD) to make these comparisons between profiles. Lower JSD values imply higher similarity between distributions.

Figure 2.3: Distribution of subreddits over the percentage of political content in them




The line graph shows the cumulative percentage of all political comments posted in subreddits that host political comments less than $x\%$ of the time. The bar graph shows the number of subreddits that host political comments $x\%$ of the time. The grey line marks the 25% threshold we use to identify subreddits as political or non-political.

$$\text{diff}_{subr} = JSD(D_{subr}||D_{pol}) - JSD(D_{subr}||D_{nonpol})$$

$$f(subr) = \begin{cases} \text{political calibrator}, & \text{if diff}_{subr} \leq 0 \\ \text{non-political calibrator}, & \text{otherwise} \end{cases}$$

where $D_{subr}$, $D_{pol}$ and $D_{nonpol}$ are strata profiles of subreddit $subr$, political and non-political subreddits respectively.

The political and non-political strata profiles shown in Figure 2.2 correspond to the actual prevalence estimates reported in Table 2.1. Comparing the two profiles, we expect that subreddits with strata profiles that are either uniformly distributed or are peaked at around the middle strata will have smaller $\text{diff}_{subr}$ scores, leading to potentially incorrectly assigning calibrators for those subreddits. Reassuringly, we find that less than 2.5% of all subreddits that we analyze have an absolute $\text{diff}_{subr}$ less than 0.1, for comparison, the distance between the two strata profiles ($JSD(D_{pol}||D_{nonpol})$) is 0.40. We include the heatmap of subreddits based on $JSD(D_{subr}||D_{pol})$ and $JSD(D_{subr}||D_{nonpol})$ scores in Figure 2.5 in the Appendix. Further, since in this work it is especially important to not overestimate the prevalence of political content in non-political subreddits, we experimented with a more conservative approach of assigning subreddits to the non-political calibrator (detailed in the Appendix section 2.9.2); we did not find a major difference in

the prevalence estimates using this more conservative approach.

### 2.4.4 Corrected "classify and count" estimation

We quantify the prevalence of political content in each subreddit $subr$ according to the proportion of content in each stratum and the corresponding forecast from the calibration curve. We calculate the estimated prevalence of political content in a subreddit (subr) as:

$$p_{\text{subr}} = \sum_{k=1}^{K} W_{k,\text{subr}} * p_{k,f(subr)}$$

where $W_{k,\text{subr}}$ is the proportion of total comments in stratum $k$ for the subreddit and $p_{k,f(subr)}$ is prevalence estimate of the $k$-th stratum of the calibrator selected by $f(subr)$.

### 2.4.5 Estimates of cumulative counts of political comments

We estimate the total prevalence of political content on Reddit as the weighted sum of the prevalence in each subreddit.

$$p = \sum \left( \frac{N_{\text{subr}}}{N} \right) * p_{\text{subr}}$$

where $N_{\text{subr}}$ is the total comments in subreddit $subr$ and $N$ is the total comments across all of the subreddits.

We can estimate the variance of this prevalence estimate by combining the variance estimates across strata. The weight for each stratum is computed from the fraction of comments that the classifier assigns to that stratum. For political subreddits:

$$s^2_{\text{pol}} = \sum_{k=1}^{K} \left( \frac{N_{k,pol}}{N} \right)^2 * s^2_{k,pol}$$

where

$$N_{k,pol} = \sum_{\text{subr} \in \text{pol}} N_{k,\text{subr}}$$

$N_{k,pol}$ is the sum of the comments in each stratum $k$ for subreddits similar to the political strata profile. $s^2_{k,pol}$ is the variance estimated for political strata profiles from Table 2.1. Similarly, we calculate $s^2_{\text{nonpol}}$ for non-political subreddits. The overall variance is just the sum.

$$s^2 = s^2_{\text{pol}} + s^2_{\text{nonpol}}$$

Finally, we define subreddits that are not explicitly political as those that host fewer than some threshold $y$ percentage of political content and calculate the aggregate prevalence and variance of political content in all subreddits that host less than y% of political content. A higher cutoff of y will, of course, treat more subreddits as non-political and thus yield a higher estimate of the proportion of all political content that is in non-political subreddits.

### 2.4.6 Prevalence estimation results

In total, we estimate that 12.84% $\pm 0.45\%$ of all comments on Reddit are political. To study the prevalence of political content in subreddits that are not explicitly political we construct Figure 2.3. The blue histogram shows the frequency of subreddits with x-coordinate percentage of political content. Of the 1399 subreddits whose classifier strata profile was closer to the profile for known political subreddits, almost all (99.71%) were estimated to have 40% or more political content. Of the 29,500 subreddits that were closer to other classifier strata profile, almost all (99.79%) were estimated to have less than 25% political content.

Each point on the orange line graph represents the cumulative percent of all political content on Reddit contributed by subreddits that host political comments less than x-coordinate percent of time. We find that 49.26% $\pm 3.59\%$ of all political content on Reddit are from subreddits that host political content less than 25% of the time. Most subreddits on Reddit host very little political content, but cumulatively the non-political subreddits contribute nearly half of all political comments. This could be driven by the few most popular non-political subreddits having far more comments overall than the political subreddits. We examine this possibility by removing the top 10 non-political subreddits (see Table 2.3 in the Appendix) that contribute the most political comments. After removing these subreddits, we find that, similar to our original estimates, about 44.82% $\pm 3.42\%$ of all political content on Reddit are from subreddits that are not explicitly political. These results suggest that the large fraction of political content in non-political subreddits is primarily driven by a large number of relatively small subreddits that each host a small percentage of political content. Robustness checks using different human judgment aggregation strategies and calibrator selection approaches yield similar estimates (see Appendix sections 2.9.1 and 2.9.2).

## 2.5 Quantifying Toxicity of Cross-partisan Political Discussions in Non-political Spaces

Our main goal in this section is to identify the toxicity levels of cross-partisan discussions on political topics in non-political spaces. Our secondary goal is to compare that toxicity to toxicity observed in other settings to better contextualize our findings. To do so, we determine how toxicity

on Reddit varies according to the following attributes: (i) political leaning of the users participating in the discussion, (ii) nature of the discussion, and (iii) type of the subreddit where the conversation is taking place.

For (i), we define a cross-partisan discussion as a left leaning user replying to a right leaning user or vice-versa. Our analysis is focused on the *reply* comments for each parent-reply discussion pair as the parent comment could be directed at a co-partisan or may not be directed at anyone if it is a top-level comment. For (ii), we rely on our calibrated classifier to determine the probability of a reply being political. Finally, for (iii), we classify any subreddit that contains fewer than 25% political content as not being explicitly political as per Section 2.4.6.

### 2.5.1 Identifying political leaning of users

To identify political leaning of users, we adopt a simple heuristic similar to ones that have been used in prior Reddit political studies [5, 210]. First, we identify the well known subreddits r/politics, r/Liberal, r/progressive as left-leaning and r/The_Donald, r/Conservative, r/Republican as right-leaning. Then, we identify a user as left leaning only if all three of the following conditions are satisfied:

1. They post more comments in left-leaning subreddits than right-leaning subreddits.
2. The mean karma points score of their comments in left-leaning subreddits is higher than their mean score in right-leaning subreddits.
3. Their mean karma score in left-leaning subreddits is greater than 1. 1 is the default score that any comment receives on Reddit. So, a higher than 1 karma score implies that the comment has met net approval by the community.

Similarly, we identify right leaning users. Among users who posted at least once in these subreddits, we have 1,223,229 left leaning and 367,363 right leaning users. We cannot identify political leanings of other users and do not include them in this analysis.

### 2.5.2 Quantifying toxicity of replies

We use the Perspective toxicity classifier to identify toxicity of a comment. The classifier provides the probability of a comment being toxic, defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" [236]. The Perspective classifier has been used in prior Reddit studies [150, 238]. Research evaluating its performance on comments from political communities shows that, on average, its toxicity classification is comparable to a single

human judgment of toxicity [180]. We have toxicity classifier probabilities only for comments posted in 2016 and 2017, so we limit this analysis to comments posted in that time interval [3].

We calculate the probability of a reply comment $r$ being toxic and political $TP_r$ as:

$$TP_r = toxicity(r) * political(r)$$

where $toxicity(r)$ is the toxicity probability given by the Perspective classifier and $political(r)$ is the probability that the comment is political, which is calculated using the calibrated classifier. Similarly, we calculate the probability of the reply comment $r$ being toxic and not political $TNP_r$ as:

$$TNP_r = toxicity(r) * (1 - political(r))$$

### 2.5.3 Comparing toxicity between discussion pairs

Our aim is to compare the mean toxicity levels of cross-partisan political interactions to toxicity levels in other settings. Replies from the same subreddit are clustered to perform semi-pooling. We conduct mixed effects logistic regression using the $lme4$ package [16] modeling the toxicity of replies with a random effect for subreddits. The count of toxic replies is modeled as the number of *successes* and the total replies as the number of Bernoulli *trials* in a binomial distribution. We estimate the following 3-way interaction model:

$$\begin{aligned} T_{s,polreply,cross} &= Binomial(P(toxicity), \\ &\quad N_{s,polreply,cross}) \\ P(toxicity) &= logit(\alpha_s + \beta_1 polsub + \beta_2 polreply \\ &\quad + \beta_3 cross + \beta_4 polsub * polreply \\ &\quad + \beta_5 polsub * cross \\ &\quad + \beta_6 polreply * cross \\ &\quad + \beta_7 polsub * polreply * cross) \end{aligned}$$

where, $polsub$ is an indicator variable for whether the subreddit $s$ is political, $polreply$ denotes whether the reply is political, $cross$ represents whether the reply is directed at an out-partisan. For each subreddit $s$, we identify the total number of replies ($N_{s,polreply,cross}$) for each $(polreply, cross)$ combination and the number of replies in $N_{s,polreply,cross}$ that are toxic ($T_{s,polreply,cross}$). We quantify

[3] We use the 5th version of the Perspective classifier

Table 2.2: Discussion data (in millions) for model estimation.

| Interaction type | Conversation type | Political subreddits | Non-political subreddits |
|---|---|---|---|
| Copartisan | Non-political | 13.83M | 57.55M |
| | Political | 21.48M | 6.52M |
| Cross-partisan | Non-political | 3.72M | 20.51M |
| | Political | 6.03M | 3.17M |

the total political cross-partisan replies and number of such replies that are toxic in subreddit $s$ as:

$$N_{s,polreply=1,cross=1} = \sum_{r \in XR_s} political(r)$$
$$T_{s,polreply=1,cross=1} = \sum_{r \in XR_s} TP_r$$

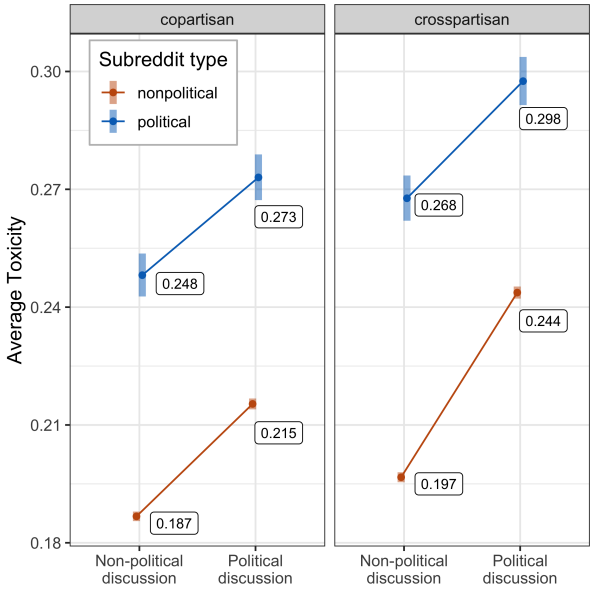
where $XR_s$ is the set of all cross-partisan replies in subreddit $s$. Similarly, we quantify the non-political co-partisan replies and number of such replies that are toxic in $s$ as:

$$N_{s,polreply=0,cross=0} = \sum_{r \in CR_s} (1 - political(r))$$
$$T_{s,polreply=0,cross=0} = \sum_{r \in CR_s} TNP_r$$

where $CR_s$ is the set of all copartisan replies in subreddit $s$. Similarly, we calculate $N_{s,polreply,cross}$ and $T_{s,polreply,cross}$ for all other combinations of $(polreply, cross)$.

Table 2.2 shows the number of replies for each $(polreply, polsub, cross)$ combination used to estimate the binomial model. Upon estimating the model, we are most interested in (i) comparing the average toxicity levels of cross-partisan political discussions in non-political spaces to such discussions in political spaces. We further (ii) compare the average toxicity levels of political and non-political cross-partisan interactions in non-political spaces, since any large increase in toxicity levels when talking politics has important implications for the health of non-political communities. Finally, we (iii) compare average toxicity levels of cross-partisan and co-partisan interactions as a sanity check. We would expect to see a higher level of toxicity in cross-partisan interactions than in co-partisan ones.

Figure 2.4: Interaction plot modeling the toxicity of discussions.




All observed pairwise differences are statistically significant at .05 level.

### 2.5.4 Toxicity analysis results

Since the coefficients in a three-way interaction are hard to interpret, we present our results in the form of interaction plots. In Figure 2.4, y-axis is the average toxicity of replies and x-axis is the indicator variable on whether the discussion is political or not. Orange and blue lines represent the toxicity levels in non-political and political subreddits respectively. Left and right subgraphs represent discussions between co-partisans and cross-partisans respectively.

First, we focus on the plot on the right side of Figure 2.4, considering only cross-partisan replies. Answering our primary research question, we find that cross-partisan replies are significantly less toxic in non-political subreddits (24.4% toxic) compared to such interactions in political subreddits (29.8% toxic). This holds both for political and non-political content. There is also a main effect of content type, with political discussions being more toxic than non-political ones. Finally, there is an interaction effect: the difference in toxicity between political and non-political discussion is significantly larger in non-political subreddits. Still, cross-partisan political replies in non-political subreddits are less toxic than even co-partisan ones, in political subreddits. A similar pattern holds for copartisan replies. However, co-partisan replies are less toxic than cross-partisan replies in all settings.

## 2.6 Discussion

Political discussions on Reddit, much like face-to-face discussions, appear to crop up incidentally in various social settings or communities [38]. We found that political discussions, while by definition uncommon in non-political subreddits, are cumulatively abundant. There are many more non-political subreddits than political ones. Due to their sheer number, non-political subreddits *cumulatively* host nearly half of all political comments on Reddit.

This result suggests the need to diversify where researchers are looking when they try to understand the nature of political discussion online. Importantly, political discussions were not limited to only a few big non-political communities. Rather, we found a large number of small-sized non-political subreddits that had occasional political comments. This surely poses an important challenge. Given the sheer scale of content in non-political communities, this necessarily requires building classifiers to accurately identify political content across a wide variety of communities. As is evident from our calibration exercise, commonly used classifiers generally include bias, making this task daunting.

Political conversations in non-political spaces not only add to the volume of total political discourse but also are qualitatively different from conversations in political communities. We find compelling evidence to support the theory that cross-partisan political discussions are indeed much less toxic in non-political spaces than such discussions are in political spaces [235]. There are multiple potential explanations for this finding. First, political discussions in non-political communities are more likely to be moderated by shared group identity [126] and social ties [20] instead of partisan identity which may reduce cross-partisan animosity [67]. Second, in general, the toxicity levels observed in non-political communities are much lower and it is likely that these low toxicity norms moderate and temper the tendency to indulge in harsh rhetoric in cross-partisan interactions [96]. Regardless of the cause, our findings pose an important caution for researchers: simply aggregating political discussions from political and non-political communities may obfuscate the differences in the types of conversations in these spaces.

There is one important nuance in our toxicity findings. While cross-partisan political discourse is indeed less toxic in non-political spaces, it is significantly more toxic than non-political discourse in the same non-political spaces. Thus, these conversations may have adverse effects on non-political communities. More research is required to understand the consequences of political interactions in these spaces. Further, we observe a larger increase in toxicity levels when talking politics in these spaces than when talking politics in political spaces. We speculate that the norms, rules and the style of moderation in place to foster conducive topic-specific conversations in non-political communities may not be as effective in handling toxicity stemming from cross-partisan political discussions. Further, a political comment in a non-political space can also be seen as a

norm violation, leading to aggression from other community members. Alternately, the smaller increase in toxicity levels in political subreddits could indicate a ceiling effect; the toxicity levels of non-political discussions in political subreddits may already be so high that they are near the upper limits of how toxic the discussions can be.

Finally, there are important open questions regarding how these political conversations in atypical spaces fit into the "deliberative system" and how they ought to be studied. The deliberative system consists of both formal spaces such as legislatures and townhalls as well as informal spaces such as social gatherings and online political discussions in social media sites [172]. Recently, deliberation theorists have highlighted the importance of everyday talk as a web that interconnects these diverse deliberation sites, urging empirical researchers to study discussions wherever they happen [136, 38]. Future work on how ideas, frames and narratives transition from these spaces to more explicitly political deliberation sites both online and offline will provide important insights on the role and importance of these conversations in non-political spaces.

## 2.7 Limitations and Future Work

The approach we followed to estimate prevalence is an improvement over a conventional classify and count approach in two important ways, but is still imperfect. The first improvement is that we employ a calibration process to map probabilistic outputs of the classifier into calibrated forecasts of the frequency of political comments. The second improvement is that, rather than assuming that the classifier performs equally well on comments originating in different subreddits, we separately calibrate the classifier on two samples of comments, one from known political subreddits and one from other subreddits. Indeed, we do find that the same classifier score is much more likely to indicate a political comment when the comment comes from a political subreddit, and this dual calibrator approach allows us to appropriately lower estimates of the prevalence of political comments in non-political subreddits.

The approach, however, is still imperfect. First, while creating two calibrators is better than one, there could be more than two types of subreddits, with the classifier having a different error profile for each. Second, our process for selecting the calibrator for each subreddit, by comparing its classifier strata profile to that of the known political subreddits and that of other subreddits, may itself be error prone. We have taken a conservative approach, with more subreddits using the calibrator that yields lower counts than the number of subreddits that are eventually classified as non-political based on their counts. This avoids overestimating the political content in non-political subreddits, but may undercount the political content in political subreddits.

In the first step, we use a simple n-grams based logistic regression classifier as opposed to using word-embeddings or deep learning techniques. Developing a more accurate classifier generally

will improve the effectiveness of the stratified sampling since each stratum is likely to be more homogeneous, leading to smaller confidence intervals for the overall prevalence estimate (see [116] for a similar argument using a simulation analysis). In our particular case, since we train the classifier on comments from political subreddits rather than on political comments, the gains from using a more accurate classifier are likely tempered by the extent to which comments from political subreddits accurately approximate political comments. Future research examining the gains of using more accurate classifiers in combination with calibrators will refine prevalence estimation techniques. Finally, our robustness checks suggest that the relatively low levels of agreement between raters did not majorly affect our prevalence estimates. Yet, raters disagreeing frequently on what is political indicates scope for improvement in the labeling process, perhaps by providing training examples and exercises.

In the toxicity analysis, we also did not perform a similar calibration process of the Perspective API. While a previous study showed that it was reasonably accurate on content from political subreddits [180], there was insufficient data provided for calibration, and we do not know whether the error profile of the Perspective API is different on content originating in political vs. non-political subreddits. Numerous other factors, in addition to the type of subreddit, political nature of the comment, and partisanship, can dictate toxicity of responses on Reddit (e.g. toxicity of the parent comment). In our analysis, we used a simplified model to only account for the select attributes of interest to our study. Finally, while Reddit is a popular online forum for political discussions, it surely is not the only one. Future work that determines the role non-political communities play in driving political discourse on other platforms can help political communication scholars better identify spaces to pay attention to.

## 2.8 Conclusion

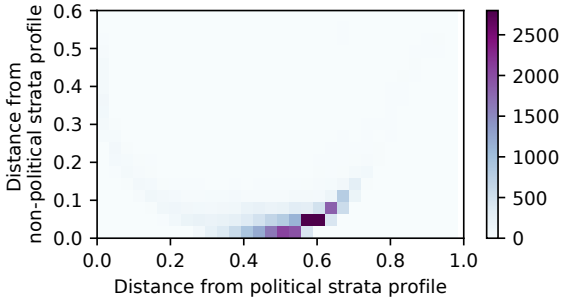
The subreddits where political comments are uncommon cumulatively produce almost 50% of all political comments. This is true even when we estimate prevalence based on conservative classifier calibrations. This large cumulative prevalence is not because of the volume of political comments in a few large non-political subreddits; instead, it is driven by a large number of non-political subreddits that host occasional political conversations. Importantly, political comments in non-political spaces seem to be less toxic on average. Thus, scholars looking at the promise and perils of online political deliberation would do well to focus their attention on political discussions that occur in venues that are not primarily organized to encourage political discussion.

# Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1717688. Ashwin Rajadesingan is supported by a Facebook Fellowship. We thank the ARC-TS team at Michigan for Cavium-Thunderx Hadoop cluster support. We also thank the anonymous reviewers for their invaluable feedback on this work.


## 2.9 Robustness Checks

Figure 2.5: Density of subreddits by $JSD(D_{subr}||D_{pol})$ and $JSD(D_{subr}||D_{nonpol})$ scores.




Few subreddits are equidistant from both strata profiles; most are close to the non-political strata profile.

Table 2.3: Non political subreddits that contain the most political comments

| Subreddit | Political percent | Political comments |
|---|---|---|
| AskReddit | 9.97 | 15,151,228 |
| pics | 19.15 | 2,979,427 |
| todayilearned | 16.57 | 2,537,532 |
| unpopularopinion | 22.49 | 1,995,731 |
| videos | 12.01 | 1,632,210 |
| funny | 10.03 | 1,614,550 |
| nba | 5.73 | 1,481,271 |
| nfl | 6.32 | 1,373,542 |
| soccer | 6.71 | 1,223,417 |
| AdviceAnimals | 20.94 | 1,080,184 |

### 2.9.1 Robustness of prevalence estimates varying label aggregation strategy

We estimate prevalence with two other aggregating strategies: (i) 'any one' strategy: comment is political if at least one rater labels it as political. (ii) 'all three' strategy: comment is political if all three raters label it as political. We estimate that 54.56% $\pm 3.28\%$ and 42.28% $\pm 3.95\%$ of the overall political content are from subreddits that are not explicitly political, based on the 'any one' and 'all three' strategies respectively.

### 2.9.2 Robustness of prevalence estimates varying political subreddit identification strategy

To ensure robustness of our prevalence estimates, we employ another identification strategy. We relax the $\text{diff}_{subr}$ criterion such that subreddits near the decision boundary will be calibrated using the political calibration curve. With this strategy, the prevalence estimates in political subreddits are expected to be higher.:

$$f(s) = \begin{cases} \text{political calibration curve}, & \text{if } \text{diff}_{subr} \leq 0.1 \\ \text{non-political calibration curve}, & \text{otherwise} \end{cases}$$

From this, we find that 43.29 $\pm 3.40\%$ of overall political content are from subreddits that are not explicitly political.

# CHAPTER 3

# ‘Walking Into A Fire Hoping You Don’t Catch’: Strategies And Designs To Facilitate Cross-Partisan Online Discussions

## 3.1 Introduction

Everyday casual political conversations through which individuals construct their identities, recognize others’ perspectives and form informed opinions are central to a vibrant deliberative democracy [105]. Many of these political interactions take place in social media where users discuss politics among other topics with friends, acquaintances and often, strangers. Although online political interactions are associated with some positive outcomes such as increased civic participation [206], they are often unpleasant experiences; about 70% of social media users report feeling stressed and frustrated when discussing politics with others on social media that they disagree with [6]. Worryingly, the tone of political discussions online tend to be less civil, less respectful and more angry than offline conversations [57].

A major factor contributing to hostility in both online and offline political discussions is the heightened levels of affective polarization that we observe today, a tendency of partisans to view opposing partisans negatively and co-partisans positively [95]. Increasingly, rank-and-file Republicans and Democrats view each other as selfish, hypocritical and close-minded [94]. This increased outparty animosity is explained by Social Identity Theory which argues that by merely categorizing individuals into groups (here, Republicans and Democrats), group identities are activated, creating an ‘us’ versus ‘them’ group dynamic [216]. Importantly, unlike protected attributes such as race where group-related behaviors are moderated by strong social norms and laws against discrimination, no such norms temper partisan hostility [94]. Thus, it is imperative that platform designers account for and mitigate the deleterious effects of partisan identity when building systems that facilitate cross-partisan discourse.

Most prior research on improving cross-partisan discourse has predominantly aimed at address-

ing partisan bias in information consumption to burst filter bubbles [171, 159, 162], with little emphasis on mitigating the role of partisan identity during interactions. In this work, we aim to reduce inter-group prejudice. We do so by designing interfaces showcasing user information to promote cross-categorization and decategorization—two relevant processes identified in social psychology research. Cross-categorization *increases awareness of cross-cutting identities with members of the outgroup* to reduce partisan animosity [27]. Decategorization *increases awareness of the distinctiveness of individual members of the ingroup and outgroup* [26]. We conduct a qualitative study using semi-structured interviews (i) to first understand the expectations, concerns and strategies of users who engage in cross-partisan interactions and (ii) to seek feedback on designs and evaluate the types of information to better facilitate cross-partisan discussions. We focus our analysis on Reddit, the popular social networking discussion site which hosts hundreds of political subcommunities (subreddits).

Our interviews reveal complex, and at times contradicting, motivations for participation in online cross-partisan talk, where participants look for serious deliberation but also entertainment and banter in political discussions. Participants also highlight varied concerns with engaging in cross-partisan discourse. As one participant succinctly put it, cross-partisan talk can sometimes feel like "walking into a fire hoping you don't catch", requiring refined strategies to increase the odds of having compelling discussions. However, our designs to decategorize and cross-categorize users produced mixed effects. Participants expressed strong support for the cross-categorization inspired "shared subreddit" component. They—especially women and minorities—expressed that the extra user information provided by the other components, while potentially humanizing, increased scrutiny on their profiles and would likely be used to attack or derail discussions. We discuss the implications of these findings and detail the design challenges and opportunities to improve cross-partisan discourse.

## 3.2 Related Work

### 3.2.1 Partisan identity in online deliberation

Normative theories of deliberation largely stem from Jürgen Habermas' conception of the public sphere, where citizens engage in rational-critical argumentation to form public opinion [79]. The presupposed conditions central to such argumentation such as inclusion, discursive equality, ideal role taking (impartiality and reciprocity) and absence of coercive power have been conceptualized as ideals of deliberation by deliberative theorists [10]. While these ideals aim to ensure that individuals are swayed only by the best of arguments, in practice, empirical research reveals how partisan identities play a consequential role in how people engage with out-partisans and their

arguments [83]. Motivated to maintain their party's positive distinctiveness and advance group status, partisans engage in in-party favoritism and out-party animosity [87]. This translates into increased partisan hostility which we review below.

Partisan hostility typically manifests in the form of incivility and abuse targeted at out-party supporters. Partisans are more willing to denigrate out-partisans while judging incivility expressed by opposing partisans more strongly than incivility by co-partisans [179]. Exposure to co-partisan attacks on out-partisans encourages copy-cat attacks while attacks by out-partisans result in stronger retaliation [71]. Further, this partisan hostility is not spurned, often it is favored, even in highly moderated online discussion spaces; studying New York Times comments, researchers observed that uncivil partisan comments received more "recommendations" from users than comments that only contain uncivil or partisan language [156]. Worryingly, exposure to partisan ad hominem criticism in news comments, which are exceedingly common online, result in more prejudiced attitudes towards out-partisans further exacerbating affective polarization [213].

These group motivated behaviors may even be exacerbated in online spaces. The Social Identity Theory of Deindividuation Effects (SIDE) posits that visual anonymity of online platforms increases the salience of group identities and adherence to group normative behavior because of the lack of individuating information about members from the group [183]. Following self categorization theory, in the presence of an accessible group identity, individuals become depersonalized and view themselves and others less as individuals having distinct personalities but instead as interchangeable group members [222]. Although commonly used to explain behavior in intragroup contexts, similar group dynamics have been observed in intergoup contexts as well. Through a series of experiments in intergroup online settings where participants were anonymous except for their group membership labels, Postmes et al. [178, 177] showed that the depersonalization predicted by SIDE theory increased the relative salience of group boundaries and led to stereotyped perception of the outgroup. An important caveat is that group identity was accessible and salient during these intergroup interactions. However, in the context of online political discussions, prior research overwhelmingly points to the salience of partisan identity in these interactions [203]. Moreover, in many subreddits such as r/AskTrumpSupporters and r/AskALiberal, users who are otherwise anonymous are required to use a flair to identify themselves as a 'Trump Supporter' or a 'Progressive', setting up the ideal conditions for group motivated behavior.

### 3.2.2 Cross-categorization and decategorization to reduce partisan hostility

Given that self categorization into partisan groups forms the basis for partisan hostility, we review two social psychology approaches aimed at changing individuals' level of categorization : cross-categorization [27] and decategorization [26]. These strategies rely on the fact that individuals

have multiple social identities apart from their partisan identities which may be activated to affect interaction dynamics [86].

#### 3.2.2.1 Cross-categorization

Cross-categorization aims to make individuals of a group aware that they share membership in another dimension with individuals of the out-group [27]. Revealing overlapping or shared group memberships makes social categorization more complex and reduces bias by increasing awareness of multiple subgroups within the out-group [46]. Further, by making cross cutting identities more salient, assimilation effects of the cross-cutting identity tend to offset the discriminatory nature of the partisan identity. Studying how other identities interact with partisan identity, Mason [140] observed "a cross-cutting calm", individuals with cross-cutting identities (for example, secular Republicans and evangelical Democrats) significantly reduced angry responses to party threats, exhibiting anger at even lower rates than weak partisans. Recently, testing the effects of shared non-political identities on partisan hostility, Levendusky experimentally found that individuals exhibited significantly higher warmth (by over 20%) towards out-partisans when they they were identified as supporting the same football team compared to no team identification ([128], Chapter 3). Based on these findings, we design an interface that surfaces "shared subreddits", users' shared membership in other nonpolitical communities, during their interaction to reduce hostility stemming from partisan identity. By explicitly highlighting shared group membership, we alert the user to the presence of "calming" cross-cutting identities.

#### 3.2.2.2 Decategorization

Decategorization is aimed at increasing the salience of intragroup variability by highlighting the distinctiveness of individual members [45]. By exposing individuals to information about multiple other group memberships of outgroup members, individuals are nudged to differentiate outgroup members from the out-group stereotype. Thus, by providing a more complex view of each outgroup member, individuals can evaluate them based on their personal merit rather than their stereotypical group memberships [26]. In politics, research suggests that people consistently overestimate the importance of partisan politics to others when no other information is provided, which exacerbates out-partisan hostility [109]; when out-partisans were described as talking politics rarely (as a proxy for importance), out-partisans were evaluated more positively [109]. Similarly, participants evaluated out-partisans who were less interested in politics more positively in a hypothetical roommate selection experiment [205]. These findings suggest that providing information contextualizing the extent of users' political versus non-political attachments may help reduce partisan hostility. Thus, in addition to highlighting shared subreddits in our design, we also pro-

vide non-political individuating information about out-partisans in the form of "active subreddits", non-political subreddits that the interlocutor has recently participated in. By explicitly highlighting the other group memberships, we aim to decategorize the user as solely a member of their partisan group, instead showcasing the user as a distinctive individual with varied interests and identities, unrelated to their political leanings.

Another intervention closely related to this work is the intergroup contact hypothesis. The contact hypothesis suggests that interpersonal interactions between outgroup members under certain conditions: equal status, common goals, cooperative, institutional support will reduce intergroup prejudice [176]. However, as Wojcieszak and Warner [232] note, the intergroup contact hypothesis has not been extensively tested in the context of partisanship. While intergroup contact is central to this study, we aim to facilitate positive intergroup contact by reducing partisan bias, whereas studies testing the intergroup contact hypothesis examine the effects of intergroup contact on reducing partisan bias.

### 3.2.3 Designing for online deliberation

Researchers aiming to improve political deliberation have typically focused on two aspects: diversifying information consumption [159, 160, 163] and facilitating deliberative interactions [112, 113]. As this work primarily concerns the latter, we review in detail the innovative interface designs that facilitate quality deliberation while reducing hostility in discussions. Early work on online deliberation centered around highly structured interactions mapping information into facts, positions, arguments and relationships between them [208]. In practice, these formal systems erected high barriers to usage as they required training to help users navigate complex predetermined interaction structures and argumentation schema [207]. Over the past decade, researchers have aimed to facilitate high quality deliberation while reducing such impediments, focusing on design considerations that center active contribution, navigability, usability, quality content and adoption [221]. Kriplean et al. [112] introduced ConsiderIt, a system that facilitates reflection of others' perspectives by allowing users to form their own pro/con list on a particular topic by also including pro/con points contributed by others. Kriplean and colleagues [113] also built Reflect, a commenting system that makes active listening the normative behavior for users of the system by including a small listening text box along with the comment for users to succinctly summarize the original comment. Another system, OpinionSpace [59] maps users to points onto a 2-D space based on their responses to five general value-based questions (answers to these questions map to either liberal or conservative leaning opinions), with the distance between the points representing similarity between user answers to question set. When a user clicks on a point, they can rate how much they agree and respect a comment posted by the user corresponding to the point.

These systems all aim to make conversations more reflective. Alternately, finding that users often used multiple social media platforms, Semaan et al. built Poli [200, 201], an integrated political deliberation environment that aggregates multiple social media.

#### 3.2.3.1 Managing hostility in online deliberation

Hostility stemming from interactions have been typically handled in two ways: (i) by structuring interactions to reduce direct contact (for example, ConsiderIt uses pro/con lists instead of facilitating back and forth interactions between users) and (ii) by removing or sanctioning problematic content, users or even entire communities [97, 199, 30]. More recently, researchers have aimed to design interfaces to proactively reduce hostility. Seering et al. [198] designed psychologically embedded CAPTCHAs to prime users (just before replying) to trigger positive emotions that increased positivity, analytical complexity and interpersonal connectedness even in cross-partisan situations. Grevet et al. [77] studying how weak ties manage political differences on Facebook, recommend another proactive approach; they suggest that "making common ground visible (i.e., highlighting past interactions and shared interests) during contentious discussions could alleviate in-the-moment tension." This lends further support to our design choice to highlight shared subreddits during interactions. Although Reddit users are unlikely to know each other unlike Facebook, we expect that showing shared non-political group memberships will likely still have an effect of alleviating tension. Somewhat paradoxically, many of Reddit's design choices such as up/down voting mechanisms and participation cultures such as circlejerking which contribute to the insularity of the subreddits may actually strengthen the effects of shared memberships in these communities by increasing users' bonds with other community members [4].

#### 3.2.3.2 Managing partisan identities in online deliberation

Despite the prominence of partisanship in political interactions, most systems or designs (barring a few notable exceptions such as ConsiderIt and OpinionSpace) do not specifically address the prevailing group dynamics in these interactions. ConsiderIt [112] takes the deliberate strategy of providing no information about users beyond their names so as to "not provide group cues to activate political identity". OpinionSpace [59] takes the opposite approach of displaying users according to their answers to values questions. It takes advantage of the fact that liberal and conservative users often have similar answers to the general values questions, resulting in closely spaced points in the 2-D space, contrary to expectations of seeing them on opposite ends of the space. This disrupts users' binary mental models and "conveys that the range of opinions do not fall along a single axis and that they are far more diverse." With both shared and active subreddits, we build on OpinionSpace's underlying principle that revealing information about users would

show that users are not as divided as they are projected to be. By showcasing non-political group memberships, users are presented with a more complicated picture about their interlocutors which we expect will disrupt the 'us' vs 'them' partisan group dynamics.

#### 3.2.3.3 Exposing user information in online deliberation

A significant concern with online deliberative systems is that the interactions are often between users know nothing about each other, leading to concerns about trust and the credibility of information exchanged [226]. For example, Kriplean and colleagues on evaluating ConsiderIt noted that "almost immediately after raising the issue of trust, user study participants would comment that they wanted to know more about the point author." However, as discussed above, they do not include user details to prevent priming partisan identity. In contrast, our design choice to show non-political group activity details to reduce partisan identity salience may also help to increase trust by providing individuating information. For example, Tanis et al. [217] found that, as predicted by SIDE, revealing individuating information about an anonymous outgroup member online, increased interpersonal trustworthiness as the member is seen less as an outgrouper and more as an individual. However, revealing information about group memberships comes with multiple concerns. Firstly, it raises concerns about inadvertently revealing sensitive private attributes [245]. Secondly, revealing this information may result in an asymmetrical disclosure, where one party knows information about the other but not vice-versa. Studies, albeit on dating practices, show that even when this information is obtained from public Facebook profiles, it is typically considered deceptive and norm violating to use them [80]. Finally, this information initiates a form of 'context collapse' [139]. On Reddit, usually user activity in one subreddit is not directly visible in another subreddit allowing users to relatively freely participate in subreddits related to unpopular or stigmatized topics without it affecting their other activities (although throwaway accounts are still common) [48]. Thus, disclosing this participation information can cause real harm and harassment, especially given Reddit's known toxic participatory cultures [143]. Therefore, we carefully evaluate if and when users consent to sharing their activity details with others.

## 3.3 Research Methods

### 3.3.1 Research context

We conduct this study on Reddit users in the lead up to the 2020 U.S. Presidential elections. Reddit is a popular social networking platform comprising hundreds of thousands of subcommunities called subreddits. Each subreddit is centered around a topic and independently run by volunteer moderators. Although there are some commonalities, the norms and rules enforced in these sub-

reddits may also vary significantly [31, 61]. For example, r/NeutralPolitics and r/moderatepolitics both host cross-partisan discussions but vary in how the discussions are conducted. While the former does not allow "bare expressions of opinion" and requires claims to be backed by sources, the latter has no such restrictions. Users interact with each other in these subreddits through a threaded comment system that allows users to directly reply to each other. This allows for prolonged interactions between pairs of users. Comments accumulate points (called karma) through up/down votes by other users which affect their visibility. Users accumulate karma points as well, which is the sum of their comments' karma points. A similar mechanism applies to the top-level posts in the subreddits called "submissions". Many cross-partisan interactions take place usually in relatively non-partisan subreddits such as r/PoliticalDiscussion, question-answer subreddits such as r/AskTrumpSupporters, ideology subreddits such as r/neoliberal and occasionally in partisan subreddits such as r/politics. As an indicator of the levels of partisan animosity prevalent on Reddit, many large political subreddits such as r/The_Donald and r/ChapoTrapHouse were banned for inciting hate just a few days before our first interview. It is in this context that we study the strategies that users engage in cross-partisan discussions and the potential effectiveness of our designs in facilitating quality discourse.

### 3.3.2 Participants and recruitment

The participants of this study are United States residents who actively use Reddit to have cross-partisan political discussions. Participants were recruited through: Reddit private messages (9 participants), recruitment posts on subreddits such as r/PaidStudies (3 participants) and multiple university mailing lists (6 participants). First, we tried recruiting by sending private messages on Reddit to users inviting them to participate in the study from a Reddit account created for this purpose; we did not get any responses. Speculating that the lack of response was due to the account being new and not trusted, we sent recruiting messages through the first author's personal account which was much older, had more karma points and a detailed history. This approach was more successful, 9 out of 83 ($> 10\%$) users we reached out agreed to participate in the study. We sent recruitment messages to users who actively engaged with opposing partisans in political subreddits such as r/politics, r/AskTrumpSupporters and r/moderatepolitics. However, this approach appeared to predominantly recruit White males, likely due to privacy and safety concerns. Therefore, we turned to two other channels: university mailing lists and subreddits such as r/PaidStudies. These are both popular recruiting avenues for academic research where we could more easily identify ourselves as university researchers to establish trust. We were able to recruit a more diverse set of participants using these approaches. The interviews were conducted from July to September of 2020. In total, we conducted interviews with 18 participants (11 males, 6 females and 1 nonbi-

Table 3.1: Demographic details of participants.

| Participant | Recruitment | Age | Gender | Ethnicity | Political Orientation | Years on Reddit |
|---|---|---|---|---|---|---|
| P01 | PM | 37 | Male | White | Left | 10 |
| P02 | PM | 19 | Male | East Asian | Left | 1 |
| P03 | PM | 23 | Male | White | Right | 3 |
| P04 | PM | 35 | Male | White | Left | 7 |
| P05 | PM | 36 | Male | Caucasian | Ind./Right-leaning | 10 |
| P06 | Univ. | 20 | Male | Caucasian | Right | 5 |
| P07 | PM | 21 | Male | White | Right | 3 |
| P08 | Univ. | 25 | Female | Chinese-American | Left | 6 |
| P09 | Univ. | 24 | Male | Hispanic / Latino and White | Left | 7 |
| P10 | Univ. | 28 | Male | Middle Eastern / Southwest Asian | Ind./Right-leaning | 15 |
| P11 | Post | - | Female | Black | Left | 2 |
| P12 | PM | 37 | Male | White | Right | 5.5 |
| P13 | PM | 48 | Female | Jewish | Left | 7 |
| P14 | Univ. | 23 | Nonbinary / Genderqueer | (Southeastern) Asian | Left | 8 |
| P15 | Univ. | 25 | Female | Asian-American | Left | 1 |
| P16 | PM | 62 | Female | Caucasian and Native American | Right/Never Trump | 1.5 |
| P17 | Post | 33 | Male | Black | Left | 1.5 |
| P18 | Post | 22 | Female | Caucasian | Left | 5 |

We report participant responses to a short open-ended demographic survey as submitted by them. P11 did not provide age details. Recruitment channels are PM (Reddit private message), Univ. (university mailing lists) and Post (post on subreddits).

nary/genderqueer). For the purpose of this paper, we exclude P15 who in her interview explained that she primarily only lurked on political subreddits and did not actually participate in them. Participants were required to be (i) US residents, (ii) 18 years or older and (iii) must have participated in cross-partisan discussions to be eligible for the study. Each participant was paid with a $20 Amazon gift card as compensation for their participation in the study.

Table 3.1 lists the demographic details of the participants.[1] Participants ranged from 19 to 62 years of age, with most participants in their early twenties. Our participants skewed mostly young, white, and male, paralleling the general demographics of Reddit users [2]. In total, 11 of our participants lean politically left, 5 are right leaning and 2 are right-leaning independents. We interviewed participants in different occupations such as software programmers, high school teacher, university administration staff, census worker, undergraduate and graduate students. P12 is also a moderator of a political subreddit. Our participants' experience on Reddit ranges from 1-15 years with a median of 5.25 years of involvement, and many spent months lurking before creating their account.

### 3.3.3 Data collection

The interviews were conducted by the first and second authors. Almost all participants were interviewed using video conferencing software (except P16 with whom we conducted a telephonic interview and narrated the designs instead). The audio was recorded after obtaining informed consent and later transcribed. The median duration of the interviews was 55 minutes. Each interview consisted of two parts: semi-structured interview (around 40 minutes) and design probe interview (around 15 minutes). From the semi-structured interviews, we obtained rich and detailed information on their motivations, positive and negative discussion experiences, and strategies they use to participate in these discussions. In the design probe part of the interview, we shared 2-3 designs based on decategorization and cross-categorization strategies on screen and after a brief explanation of the probe, we asked for their feedback and reactions to the probe. We also specifically probed for concerns they may have about using the interface and about others using this interface when interacting to them.

### 3.3.4 Data analysis

Each interview was transcribed using otter.ai before manual revisions and corrections by the first and second authors. The interviews were coded using a grounded theory approach [33] consisting of both open and axial coding using NVivo software. The first and second authors independently

[1] We report participant responses as is from a short open-ended demographic survey.

[2] https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

coded the interviews (12 and 5 interviews respectively) using open coding. These codes were then combined into higher level categories using an axial coding process. The two authors met multiple times to discuss and combine these categories at length, and identified emerging themes around (i) motivations for participating in cross-partisan discussions, (ii) qualities of political discussions, (iii) proactive and reactive strategies adopted by participants to have good discussions, (iv) folk theories of why cross-partisan discussions are difficult to sustain, and (v) humanizing effects of the design probes and concerns around misuse. Through the course of interviews, we held weekly meetings with the research team to discuss the feedback from interviews about the designs, allowing us to incorporate minor modifications to the design probes detailed in 3.3.5.

### 3.3.5 Design probes

Currently, the Reddit interface, as shown in Figure 3.1a (the interface excluding the user card), does not directly provide any information about the interlocutor. Users need to hover over the username to obtain basic profile attributes such as time since joining Reddit and total karma points. To view past comments or other subreddits their interlocutor has participated in, the user has to go to the interlocutor's profile page by clicking on the profile icon. Through our design probes [70], we explore alternate versions of the Reddit interface where the user has access to additional information which is expected to decategorize or cross-categorize their interlocutor. By visually showing designs containing this extra information, as opposed to asking participants to imagine such a possibility, we provide a realistic representation of this information on which to base their opinions. The aim of the designs are two-fold: (i) To understand how participants perceived the impact of the extra information on their conversations and (ii) To explore different designs based on participant feedback to build a functional browser extension. Below, we detail each component of the user card which is intended to show up when users click the 'reply' button to reply to another user's comment (as shown in Figure 3.1a).

#### 3.3.5.1 (A) Shared subreddits

This component shows the list of non-political subreddits that both the participant and their interlocutor have recently participated in. By explicitly highlighting shared group memberships, we alert the user to the presence of cross-cutting identities which is found to have a calming effect on partisan hostility as described in Section 3.2.2.1. The subreddits will be ordered such that smaller or less common subreddits (based on number of subscribers) will be shown first, since group size is negatively associated with affinity towards the group in online communities [111].

**Feasibility Analysis** Displaying shared subreddit memberships to interlocutors will be beneficial only if they actually share subreddit memberships. This concern is especially significant now as political science research suggests conservatives and liberals on average make different choices on even non-political decisions such as coffee choice and fast food consumption [49]. This may result in few common subreddit memberships between cross-partisans. Therefore, to evaluate the reach and thereby effectiveness, we estimate the prevalence of shared non-political group memberships among users who engage in cross-partisan discussions on Reddit using publicly available data [17].

First, using a simple heuristic from prior work on Reddit, we identify users who are left or right-leaning based on their activity in left and right leaning subreddits. First, we identify r/politics, r/Liberal, r/progressive as left-leaning and r/TheDonald, r/Conservative, r/Republican as right-leaning subreddits. Then, we classify users as left leaning if (i) they comment in more left-leaning than right-leaning subreddits (ii) the mean karma points of their comments in left-leaning subreddits is higher than their score in right-leaning ones and (iii) their mean karma score in left leaning subreddits is at least 1. Likewise, we identify right-leaning users. [3] Then, using these user classifications, we identify all distinct co-partisan and cross-partisan interlocutor pairs in 277 political subreddits (previously identified by [180]). For each pair, we identify if they both participated in a common subreddit within the last 3 months (approximately 100 days), while excluding the 277 political subreddits and the default subreddits[4] from consideration. We find that, in an average subreddit, 44.26% and 51.94% of all cross-partisan and co-partisan discussion pairs share at least one common non-political subreddit. These percentages are encouraging because (i) in an average subreddit, about half of all discussion pairs share a non-political subreddit indicating that showing shared subreddits is a viable option for a sizable population of interactions, (ii) the difference between co-partisan and cross-partisan percentages, although statistically significant, is small enough to suggest this difference may not significantly exacerbate out-party differences.

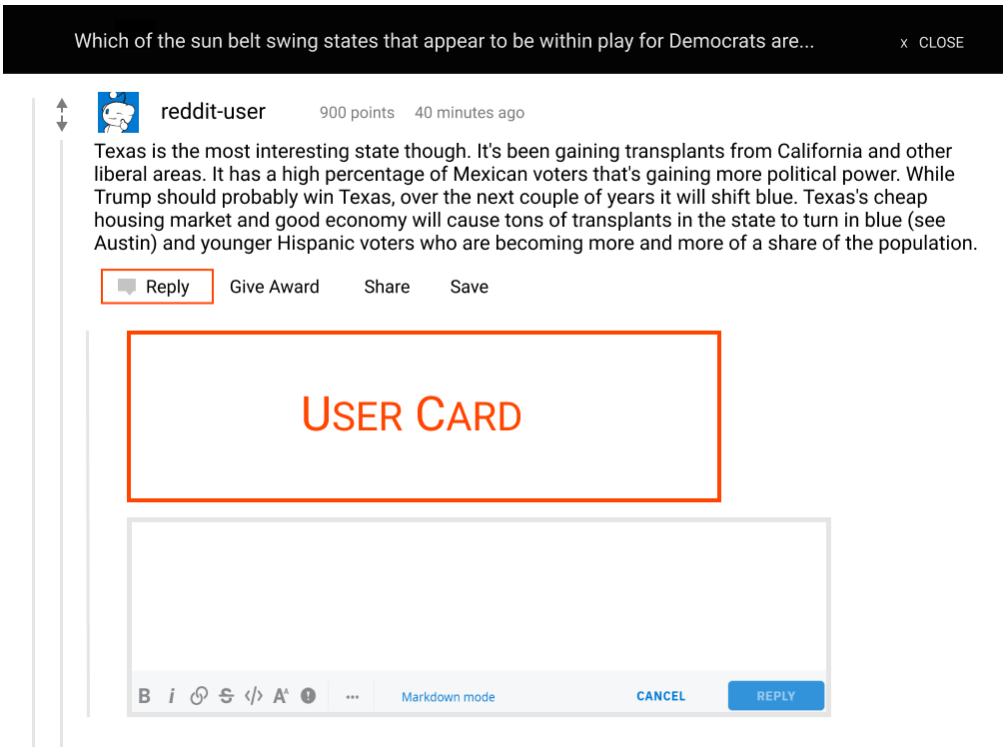
#### 3.3.5.2 (B) Active subreddits

This component shows the list of non-political subreddits that the interlocutor has recently participated in, excluding the "shared subreddits". By explicitly highlighting the interlocutor's varied interests and identities based on their activity, we aim to reduce hostility through decategorization as described in Section 3.2.2.2. Again, these subreddits will be ordered such that smaller or less common subreddits (based on number of subscribers) will be shown first.

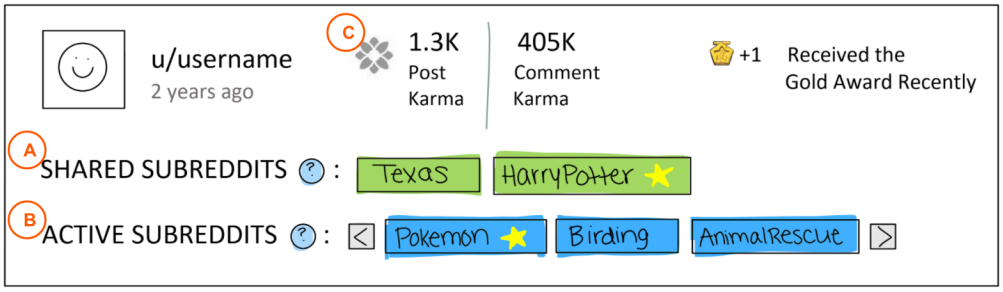
[3] We classify 1,223,229 users as left leaning and 367,363 users as right leaning. We cannot identify the political leanings of other users using this approach.

[4] Until June 2017, Reddit users were automatically subscribed to these subreddit which are amongst the largest on the site.

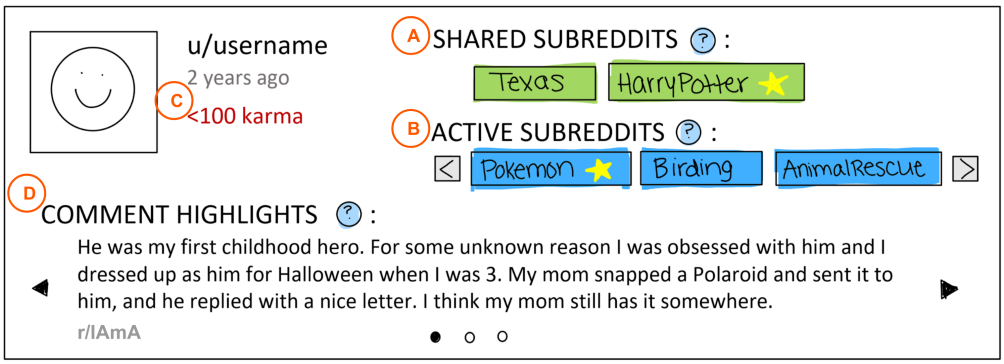
Figure 3.1: User card designs




(a) Example of a comment on the Reddit interface with our user card. The user card would appear when users click the reply button to type a reply.




(b) Design A: The user card shows active and shared subreddits as well as karma points and awards.




(c) Design B: In addition to components in design A, the user card shows comment highlights.

**Feasibility Analysis** Displaying active (and not shared) subreddit memberships will be beneficial only if they actually participate in other subreddits. Redditors are known to create multiple throwaway accounts to provide added anonymity especially when discussing contentious issues [117]. If users predominantly use throwaways when talking politics, then there may be few other subreddits to display to interlocutors. Therefore, using a similar approach as earlier, we calculate among users who participate in political subreddits in a given month, the average number of nonpolitical subreddits they participated in the prior three months. We find that the left and right leaning users active in political subreddits in 2019, on average, engage in about 23 and 20 subreddits respectively in the prior three months, providing evidence that this design is indeed feasible given current user behavior data.

#### 3.3.5.3 (C) Karma points and awards

This component shows the karma points and awards earned by the interlocutor. Though unrelated to decatagorization or cross-categorization, karma points may have a potential to improve conversations by providing an indicator of trust or reputation bestowed on the user by the Reddit community. This feature was designed based on feedback from ConsiderIt where their study participants expressed difficulty in evaluating the trustworthiness of claims put forth by other users about whom they knew nothing about [112]. Highlighting awards and karma points could present one way to highlight trust without giving away partisan cues about the user.

#### 3.3.5.4 (D) Comment highlights

This component highlights top comments posted by users in non-political subreddits based on karma points. By providing examples of top non-political comments by the interlocutor, we aim to showcase their positive behavior in other subreddits indicating that they have multiple interests apart from their politics. Along with the active and shared subreddits, comment highlights provide deeper insights into not only where they participate but also how they do so in the other subreddits. A more discrete version of comment highlights is the star shaped link in the active and shared subreddit boxes which links to a top comment (above 50 karma) posted by the user in that subreddit.

**Design evolution** First, we conducted interviews using one design (Design A, Figure 3.1b). During the first few interviews (P1-P3), participants suggested providing cues about not just *where* users were active but also *how* they behaved in those places. This feedback resulted in us evaluating additional designs that made visible more details such as "comment highlights" in Design B (Figure 3.1c). We showed Design A to all participants and Design B to P4-P18. We also devel-

oped other largely similar versions of these designs aimed at reducing the size of the user card by moving the placement of karma points, showing shared and active subreddits on the same row and linking to a highlighted comment rather than displaying full text. All designs featured shared and active subreddits, the primary focus of our research enquiry.

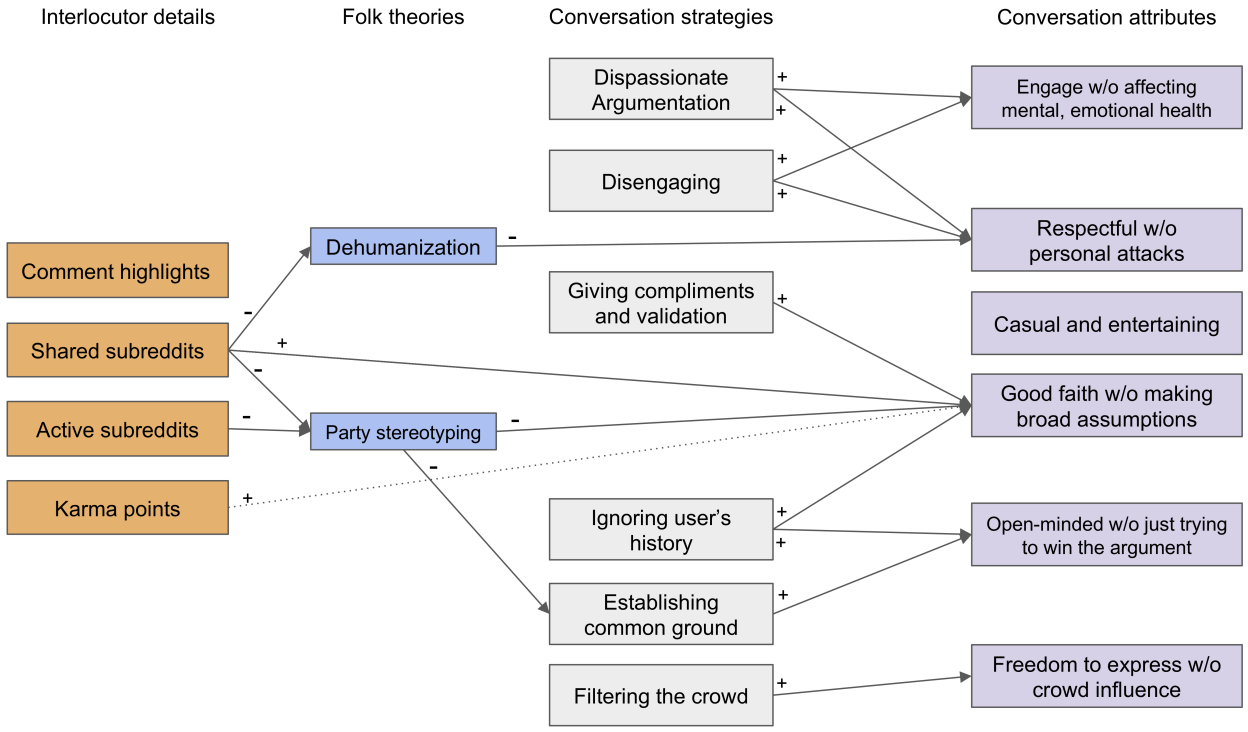
As a cautionary tale, we note that in our case, the initial mode of recruitment (private messages) resulted in mostly White/Caucasian male participants (P1-P7) who did not have major concerns about their information being made more visible in the user cards. However, as we interviewed a more diverse set of participants recruited through other channels later in the study, concerns about revealing information became clearer. We caution that designs such as ours that highlight user information need to be carefully evaluated for their effects, especially on members of disadvantaged groups early in the design process.

**Target demographics** A significant advantage of these designs is that they are not explicitly political; users would simply see the non-political activity of other users. Therefore, an extension built using these designs can be marketed as a fun tool that helps users learn more about others, which we expect will help diversify the kinds of users who install the extension. By positioning it as a general purpose fun tool, we anticipate that all users, not just the ones most motivated to improve their discussions, will use the extension. However, because of the nature of these designs, we expect it to be less effective on extreme partisans whose non-political subreddit membership is likely largely stereotypical. Further, the extension is not expected to reduce hostility expressed by individuals who are determined to be hostile, rather it is a subtle intervention aimed at users who engage in cross-partisan interactions in earnest.

## 3.4 Findings

We organize our findings as follows: First, we detail the qualities participants seek in a good cross-partisan political discussion (right most in Figure 3.2). Next, we highlight strategies that participants adopt to improve the chances of experiencing these good qualities in their discussions (center-right, in grey). Then, we detail two folk theories–dehumanization and stereotyping–that participants attribute to the many bad conversations they have in spite of following these strategies (center-left, in blue). Finally, we explore how the user information embedded in our designs may help overcome dehumanization and stereotyping but may also lead to other concerns (left most). The (+) and (-) signs in Figure 3.2 indicate positive and negative relationships between the entities. For example, establishing common ground increases the odds of having open-minded discussions while increase in dehumanization decreases the odds of having respectful interactions.

Figure 3.2: Summary of findings showing the relationships between good conversational attributes and user strategies employed during the conversation, folk theories and interlocutor details made available through design.




The (+) and (-) signs indicate positive and negative relationships between the entities. For example, establishing common ground increases the odds of having open-minded discussions while increase in dehumanization decreases the odds of having respectful interactions. The connection from karma points to good faith is a dotted line since karma is a weak/basic indicator of good faith.

### 3.4.1 What is a 'good' cross-partisan political discussion?

When asked what they considered to be a good cross-partisan interaction, participants described two kinds of interactions: (i) serious deliberative discussions on political or policy issues and (ii) casual conversations for entertainment and banter. Interestingly, many participants reported engaging in both these types of conversations depending on their mood or time constraints. We describe these conversations in detail below.

#### 3.4.1.1 Serious deliberative discussions

Most participants expressed that they were looking for some form of serious deliberative discussions. Many of the specific attributes they looked for in such conversations directly mapped to the deliberative ideals of mutual respect, reasoned arguments and the freedom to express without

coercion (such as crowd influence).

**Respectful without name calling and personal attacks** Most participants expressed that they aim for conversations to be polite and respectful without devolving into personal attacks.

*The bad conversations start off right away antagonistically, you'll have like a Trump supporter or a liberal supporter like basically just start off by saying nasty ad hominem attacks about the other side, these are already like non starters like you're not going to get anywhere.* - P01

**Listening with an open mind without simply trying to win the argument** Most participants entered cross-partisan discussions without expectations of changing others' views. Instead, they looked for conversations where their interlocutors were simply open to acknowledging at least some of the issues that they had raised.

*For instance, [pretend] you're pro Trump. But I made a point that you can't find anything to disagree with about. Can you actually say that? You know, while I support Trump, you actually have a valid point on this particular issue. So, being willing to listen is to at least consider what the other person is saying, which does require listening, is huge.* - P16

However, participants indicated that many conversations are not open-minded exchanges of ideas but rather interlocutors simply trying to one-up each other. Thus, some participants do not even look for open-minded users, instead they use the conversation to explore an issue. For example, P12 recounted instances where he argued with others "for the sake of just understanding that idea". Few of our participants actively looked to change others' views. Still, they reported that instances where they changed others' viewpoints were uncommon.

**Good faith without making assumptions** Participants look forward to having good-faith conversations with others–conversations where everyone has good intentions, engages in earnest, and refrains from making assumptions of others.

*[A good conversation needs] understanding that each person participating has experiences that you might not be able to relate to or like, language is really imperfect... understanding that like everyone is like trying to do right by their communities and families. So even if like we can't understand what those obligations look like, they are "good people"...* - P14

However, participants noted that in many of these conversations, people quickly make assumptions and judge others without giving them a chance to explain their beliefs.

**Informative without unverifiable claims and party talking points** A prime motivation for almost all participants to engage in cross-partisan political discussions is to learn about opposing viewpoints and contribute alternate perspectives. Similar to Semaan et al.'s [202] findings, most

participants in our study explicitly acknowledged the effect of filter bubbles or echo chambers on their own beliefs. They stated that they actively try to engage in cross-partisan political discussions to gain alternate perspectives.

*I enjoy talking with [conservatives], because they'll see an article and see it completely different than the way I see it. It's a curiosity for me... And it's good for me to know that they exist, and not just this little bubble that I'm in.* - P04

However, in many conversations, participants noticed that the interlocutors were simply regurgitating party lines or spreading debunked misinformation without doing research on their own to understand the issue.

*When the person is not willing to debate facts, when they start spewing basically talking points, talking points that are disputed, talking points that aren't related, talking points that don't make sense... then that's a pretty good indication it's not going anywhere.* - P17

**Freedom to express without crowd influence** Almost all of our right-leaning participants and many left leaning participants described how their comments are often heavily downvoted or heavily answered by many users (dogpiled), which overwhelms them.

*Reddit, a lot of it is primarily liberal. So it's like, if you want to come with any conservative opinion whatsoever, you're probably just going to get mobbed on and, you know, for every 100 people that mob on you, there might be five actual discussion points in there.* - P05

However, participants sometimes do take into account the feedback they receive from others, especially if they are co-partisans. As P13 described, receiving downvotes or multiple replies does prompt her to reflect and question her own positions on the issue.

*People will reply vehemently... I'm really surprised when it's more than two people... It makes me wonder whether or not my position on that topic should be that position. Do I change my mind? No, not necessarily, but the thought is there and that's important too, because you have to constantly question your own thoughts.* - P13

**Engaging without affecting mental and emotional health** Many specifically described the toll some of these conversations had on mental health. Participants report that these cross-partisan conversations often involve a lot of work and mental effort, the after-effects of which may continue to linger in their thoughts through the day. Even conversations that do not veer into name calling or character attacks sometimes leave participants frustrated and exhausted.

*The fear is that like, it'll consume the whole day. I'll be thinking about something politically... and I'll just keep talking about, thinking about like, something political and get worked up about it.* - P07

However, some participants are able to ignore views that they dislike and not let it affect them,

or quickly move on to lighter content to decompress. As P11 put it, "continue to scroll to find a cute puppy". Others, like P2, are resigned to the fact that being exposed to objectionable views is the cost of having a cross-partisan conversation.

*But it's those little prejudices that people have that bother me, but while I do feel bad reading them, I don't think I am necessarily upset because of it. Because the main reason why I'm on there is for a political discussion.* - P02

#### 3.4.1.2 Casual and entertaining conversations

Participants explained that they were on Reddit primarily to have fun and entertain themselves. They did not use Reddit for political discussion alone and most actively engaged in other relatively non-political communities such as r/DIY and r/Makeup. They saw their participation in political conversations as one among many other leisure activities in which they partake on Reddit. In fact, some participants explained that many of their political conversations were incidental and stemmed from casually browsing through their home feed. They were not actually trying to engage in the conversation deeply and would typically quickly comment and leave.

*I kind of will just comment whatever, not really trying to seek out [conversation] because also, once you do get a productive conversation going, it takes a lot of energy, it can take a lot of time... I don't know if I have the stamina for it all the time.* - P17

Sometimes, participants engage in more casual political subreddits such as r/PoliticalHumor and r/PoliticalCompassMemes. However, many find discussions in mainstream discussion subreddits entertaining as well. P09 described how he uses Reddit most heavily when he is bored at work, and primarily looks for entertainment when participating in political discussions. Below, he described one such discussion where a conversation in r/politics devolved into a conspiracy theory.

*One of the funniest ones I ever remember reading was a pizzagate-like thread of comments. It was just hilarious. Because that was a conspiracy theory and then lots of people branch off and like, I don't know, it was so entertaining... Like watching some people just put two and two together on things I could have, like, never thought twice. That's some high entertainment value.* - P09

Participants also mentioned that even within more serious discussions, someone may post a witty rejoinder, or a funny meme which makes the tone of the conversation fun and casual. They also pointed to how some political subreddits have dedicated discord chat servers and occasional free talk threads, allowing users to have casual discussions unrelated to politics. In certain instances, participants indicated that normatively anti-social behavior such as trolling and making others angry were also fun activities to take part in on political subreddits.

*If I am in a really bad mood, and I'm just out to, you know, troll up a storm, then what I think of as good [conversation] is when I get someone's goat, when I make them very viscerally angry, and they keep responding, and I can tell they don't want to respond, but they have to respond and*

*that's when I've got them!* - P10

This poses a direct contradiction with one of the most commonly cited motivations—to have respectful deliberation. It is important to note that the same participants who seek entertainment in political discussions also participate in more deliberative political discussions. The kind of conversations which participants choose to take part in depends on a wide range of factors such as mood, time constraints and current events.

**Frequency of "good" discussions** Most participants reported that they participated in at least a few political discussions that they felt were good and satisfying. However, these occurrences were rare. Many long-time participants reflected that their conversations have turned angrier and devolved more into name calling over the past two years. However, many participants characterize engaging with the other side as a form of civic duty, something that is difficult but needs to be done in order to deeply understand issues affecting the country. Therefore, to navigate these conversations and increase the odds of having a good interaction, participants have developed multiple strategies to select where, who and how to talk to cross-partisans which we detail in the following sections.

### 3.4.2 Strategies adopted prior to engaging in cross-partisan discussions

#### 3.4.2.1 Choosing where to have the conversation

Many participants reported taking part in multiple political subreddits and carefully curating the subreddits that they subscribed to, considering the quality of discussion, member composition and level of moderation in conversations in those communities. Some participants completely avoided large generic subreddits such as r/news and r/politics and instead participated in relatively smaller niche political subreddits such as r/tuesday which is a relatively small center-right subreddit whose participants are derided as RINOs (Republicans In Name Only) for not being Republican enough and r/moderatepolitics, a moderate sized non-partisan discussion subreddit where they could have more nuanced conversations. [5] Note that these niche subreddits are not homogeneous partisan groups. They were relatively much smaller, well moderated, and frequented by members who similarly value cross-partisan interactions. P05 explained why he customizes the subreddits he participates in:

*The reason why I follow some of these particular subreddits is just because people seem a little bit more reasonable in how they respond. You know, we probably both know that Reddit is primarily liberal, just in general. And you kind of have to go to specific subreddits if you want to get say like,*

[5] r/tuesday and r/moderatepolitics had about 12,000 and 50,000 subscribers respectively at the time of conducting the study.

*right leaning information or commentary. But some of the subreddits like r/Conservative and, you know, now banned, r/TheDonald, they're just so far over there. And the quality of discussion, in my opinion, is very, very low.* - P05

Through experience, some participants claimed to understand how their comments will be received depending on the type of subreddit in which they participate. A few others explained that they participated in subreddits that only partially aligned with their views which allowed them to have disagreements knowing that there was also some common ground.

*The r/Neoliberal one is a good one for me, just because I do dissent somewhat from some of the things they believe, but I also have a lot of common ground. So it's a space where I can have a lot of discussions with people who write, you know, at least they have similar moral frameworks, similar sort of ideological frameworks, even if some of the actual practices diverge a little bit there.* - P06

Even within subreddits, a few users are selective about which threads to engage in. For example, P04 explained that he recently started engaging in the open talk discussion threads in r/tuesday, rather than topic-specific ones. He reasoned that most of the subreddit "regulars" hung out there and were able to have deeper conversations since these threads do not usually get upvoted enough to show up on people's homefeeds and attract widespread attention from casual users. Generally, participants attest that identifying the right space to participate in tremendously affects all aspects of the discussion.

#### 3.4.2.2 Choosing who to talk to

All participants said that they viewed only the text of a comment by a user to decide whether to engage with them. Many participants described having an intuitive sense of how the conversation was going to unfold based on their reading of a users' initial comments. Some said that they could understand the 'personality' of the user by reading between the lines to make quick judgements about whether to talk to them.

*I can usually tell from the first comment–and I assume other people could as well–about the tone of the discussion with this person... Most of the time, it's just like, I know, that's going to be a bad convo, and that this is going to be more reasonable... I think I think it's probably about 90 or 95% of the time, the first comment generally identifies how the conversation is going to go.* - P05

Many participants also explained that they try not to enter into discussions with users who use profanity or strong emphasis words such as 'obviously' or 'clearly' when making suppositions on a topic. However, some participants also recalled times when they deliberately chose to engage with users making such statements when they were in a combative mood.

### 3.4.3 Strategies adopted during cross-partisan discussions

#### 3.4.3.1 Establishing common ground and posing questions

Participants recognize the current contentious political climate and are extra careful in how they communicate in cross-partisan discussions. Most participants reported that they typically start by signaling common ground with the user, highlighting parts of the argument that they agree with in a polite and respectful manner. Then, they detail aspects that they disagree with while remaining extremely deliberate about how they frame their critiques, often posing them as questions. Many noted that they sometimes rewrite their comments multiple times to ensure that their views are conveyed accurately but without offending the people to whom they are talking. For example, P13, a high school teacher, described how she communicates with those she disagrees with.

*I find that when I'm super careful about how I engage somebody whose opinions I differ with, the more careful I am, the better the conversation goes... rule number one when you have a parent teacher conference is "this is somebody's kid, say something nice." So for online, it's what do you agree with? What did this person say that you wholeheartedly agree with? and start from there. And then after that though, don't attack, [instead] question...* - P13

Thus, by establishing common ground and approaching conversation partners without a heavy gavel, users aim to signal that they are **open-minded and reflective**.

#### 3.4.3.2 Giving compliments and validating the interlocutor

Some participants, upon sensing that a conversation has turned for the worse, typically give one last shot at reviving the conversation by explicitly complimenting or validating the interlocutor, as a sign of good faith. P18 explained how they sometimes try to correct the course of the conversation.

*I think if someone's aggressive, but you can kind of sense that they do have the ability to have a better conversation, I think just being nice, I think not playing into their tricks, even validating them in a certain way helps... I love to say well, I agree with that. But I also have these things that I believe in and this and that, so I think actually validating them a little bit... you kind of show them your cooperation, they might actually come out to be more cooperative and I've seen that happen.* - P18

Thus, by acknowledging and validating the interlocutor, participants signal **good faith** to the interlocutors.

#### 3.4.3.3 Dispassionate argumentation

Many participants explicitly aimed to keep a calm and dispassionate demeanor when interacting with cross-partisans. Knowing that disagreements tend to engender strong emotions, participants

keep conversation on topic by focusing on the facts, providing arguments, and trying not to react emotionally to the interlocutor's arguments. For example, P04 described a particularly difficult conversation with a right-leaning user who argued that reports of police brutality in the US were overblown:

*I always try to start off with a very dispassionate response. And try to back up my claims with as much fact as possible and try to keep feelings out of it as much as possible, leave my worldview out of it as much as possible because we clearly don't share the same worldview. So I'm never going to be able to win that person over with that aspect, but just try to make it dispassionate.* - P04

However, he conceded that being dispassionate on topics like the Black Lives Matter protests is especially difficult and instead, chose to disengage altogether. In other instances, once participants sense that a user is becoming emotional in their replies, they swiftly disengage or concede the argument before it (potentially) devolves. For example, P12 said:

*There'll be times when I just stopped a conversation because someone's emotional. And I'll just concede the argument. There's no point in pushing somebody into a character attack when they're just getting emotionally invested in the argument.* - P12

Thus, users aim to remain dispassionate in their conversations in order to **maintain their own mental health and to prevent the conversation from potentially devolving into name calling and character attacks**.

#### 3.4.3.4 Avoiding looking at the user's profile unless the conversation goes stale

Many users actively avoid learning more about their conversation partners by refraining from viewing their profile details such as karma points or past comments unless the conversation goes awry. By ignoring other possibly disturbing details about their conversation partner, participants focus their attention squarely on the argument that the person presents, not biased by their past opinions in other topics.

*Normally, if I know that the other person that I talked to is a huge racist, or a sexist, or has some very, you know, skewed perspective on the world, then I would immediately want to stop talking to them... so I tend not to read the other person profile. I just try to, you know, discuss the topic with them, just that topic and nothing else... I don't want to know about that person, other than the things that are relevant for that discussion.* - P02

Others who do view past comments express doubts about whether knowing more about the user helps or hurts the conversation. For example, P04 was concerned about whether knowing a user's positions outside the topic might prejudice him in the conversation.

*I have [viewed past comments] in the past, especially with a name that I don't recognize, just to get an idea of what I'm getting myself into. But at the same time, I almost feel like that kind of*

*almost prejudices me. And I almost want to have a narrowly scoped discussion that doesn't have the baggage of previous discussions or previous outside-of-this-subreddit's discussions.* - P04

Others, like P14 felt like knowing more about a user makes having a conversation with them difficult since the distance between their worldviews becomes more apparent. Looking at user profiles, most participants viewed karma points not as a predictive indicator of whether the user would be a good person to talk to but instead more as an explanatory variable when a conversation goes awry to make sense of the user's behavior.

Thus, by not viewing the user comment history or karma points, participants essentially try not to learn more about the user, ensuring that they discuss in **good faith and with an open mind without making assumptions**.

#### 3.4.3.5 Filtering the crowd

Many participants recalled instances where their reply notifications "blow up", where multiple users angrily reply to their comments. In those instances, participants typically put on "social blinders" and focus on replying to only a specific individual.

*Another thing I might do if I had a bad argument, but I liked one of the other people in the audience, and then everyone else is a bit [much]... I might sort of put like, social blinders on, and just tag that one person over and over again, to make it clear I'm only talking to them. Or I might continue the conversation in the chat box with them, Reddit has chat now, and continue in the DMs.* - P10

Unrestrained by the topic or other users in the subreddit, P16 found that DMs allow for users to switch topics and be more open. She noted that "it just seems to me though, that direct message allows for a level of intimacy and being real." In other instances however, multiple participants reported that they have been directly targeted or harassed by others through DMs.

#### 3.4.3.6 Disengaging from the conversation

By far, the most common reaction to a conversation which regresses into a personal attack or becomes combative is to disengage and exit the conversation.

*I don't have to sit there and have somebody be ugly to me. That's not what I'm on the Internet for. I'm on the Internet to have fun and to be educated and not to be harassed.* - P11

Some participants use more stringent methods to disassociate themselves from the conversation by deleting their comment, reporting to the moderators, blocking the user and in rare cases, unsubscribing from the subreddit. It is important to note that disengaging is not a last resort action that participants take, oftentimes, disengaging is the first action that they take. Thus, to **safeguard their own mental health and to shield themselves from personal attacks**, participants simply

disengage and walk away from the conversation.

#### 3.4.3.7 Counter strategies

Not all strategies employed by the participants are conciliatory or aim to further the discussions. In some instances, participants said that they would counter by using aggressive or condescending language.

*if I'm feeling petty, it's not like the right thing to do, but I like call them out in kind of a condescending way, I don't like using insults and things like that. But if I do want to be petty, it'll be more like, yeah, condescending or rhetorical questions, lighter, but still, I know, I shouldn't be talking like that.* - P08

In other situations, recognizing that the user they are talking to is angry, some users try to make them angrier.

*Normally, when they're really mad and they go out of their way to like, target me. I normally just like, take the piss, you know, I kind of try to make them more mad... I don't confirm their prejudice. I just go, you know, oh man, look at this guy...haha... It's kind of stuff like that.* - P07

Others described using some of the tactics described by Jhaver and colleagues [98] such as identity deception and sockpuppeting to counter hostility.

**Do these strategies adopted during the conversations work?** *Sometimes.* Most participants acknowledged that while they do employ many of these strategies, the most effective approach in dealing with volatile conversations is to leave. Many recalled instances where they've tried to course correct a conversation only to make it worse. For example, P13 said:

*I tried to engage once with somebody that vehement, and they just were, they just attacked. It was like, you know, it was like getting a text that's like, all caps from your mom. And it's just, you know, who needs that? So I'll just drop it. I don't reply. I just let it go.* - P13

Most participants explained that it is best to find another conversation to participate if their current conversation became worse. As P6 put it, "when you invest a lot of energy into what is effectively an online discussion, it can sometimes feel like shoveling money into a fire."

### 3.4.4 Party stereotyping and dehumanization: folk theories on what affects their conversation

While we did not specifically ask participants for why they thought their strategies did not always bear fruit, many participants expressed unprompted explanations of their own, specifically attributing party stereotyping and dehumanization as a cause for concern in cross-partisan discussions.

#### 3.4.4.1 Party stereotyping

Some participants attributed certain conversations going awry to stereotyping along party lines. In their experience, some users were quick to judge them as an extreme liberal or conservative and project on them, what they perceive to be the typical characteristics of the group. P07 explained one such instance:

*I think the worst one is where like, they kind of view you as the representation of like the right wing or something. I'm not very conservative, but it's annoying when people are like, oh, you religious conservatives. Like, I'm not very religious. I'm not very conservative, they assume that like, you represent the whole like, you know, straw man of the entire wing* - P07

In other cases, participants were concerned about how a co-partisan user supporting a position held by the participant may speak up for them. However, in doing so, they may provide reasons that are incongruent with the participant's own reasons on why they support a certain position.

*You end up with the problem of sometimes someone will say something as if he's speaking for you. But really, it's like, No, no, don't put me in there. [co-partisan would say] "And that's the issue is Republicans, Black people. And I'm sure everyone else here [agrees with me]", please no no noooo! We are not the same, though.* - P06

Therefore, party stereotyping erases differences between individual group members (both in-group and out-group), leading users to **make broad assumptions of each other and affecting the ability to build common ground**.

#### 3.4.4.2 Dehumanization

Contrasting with face-to-face interactions or interactions with people they personally know on other social media, many participants **attribute personal attacks** to dehumanizing effects afforded by anonymity on online platforms like Reddit.

*Especially the anonymity that Reddit has, it's very easy for you to forget that that's a real person on the other side or for other people to forget that's a real person on the other side, you just start like throwing vitriol and people are just like, non-caring, like, will use any type of language to try and get their point across. And it's like, hey, I'm a human being, let's be at the very least cordial, we don't have to agree, but we should probably not try to like kill each other with words.* - P11

Either through personal experiences or subreddit rules or by reading the 'Reddiquette' [6] which urges users to 'remember the human', many participants recognize the need to view other users as human beings instead of a username on a screen.

[6] informal norms that users are urged to subscribe to, https://www.reddithelp.com/hc/en-us/articles/205926439, "Remember the human. When you communicate online, all you see is a computer screen. When talking to someone you might want to ask yourself "Would I say it to the person's face?" or "Would I get jumped if I said this to a buddy?"'

*I'm somebody who grew up with the internet evolving. I didn't start it when I was a kid. You know, I didn't have a phone in my hands until I was in my 20s. So I still go into every online conversation the way I would a real conversation. I'm constantly remembering that there's somebody on the other end. I'm consciously like this. I really pay attention to the words on the screen.* - P13

While participants understood the importance of remembering the human, they found it difficult to practice it online without other visual or auditory cues. Most participants felt that knowing more about the user and their interests would help view them as more complete human-beings rather than just as someone who has strong political opinions. For example, P06 explained that the users he talks to online are strangers and that knowing more about them would humanize them:

*It would be cool to know what kinds of stuff the other person's into, and just to maybe not put a face to it, but maybe, you know, at least see some additional humanity behind what is otherwise a username and text.* - P06

However, many of the participants who acknowledged the importance of 'seeing the human' remained deeply skeptical of knowing too much about their conversation partner for fear that extra information may distract or bias the conversation. For example, later, when asked if he would like to know more about users he talks to, P06 said:

*In a sense, I don't want to know very much about the person other than that they are a good partner or conversationalist or whatever... I wouldn't want to know anything about the person, their race, I wouldn't want to know their gender, I wouldn't want to know shit... I think beyond including resources that clue people into someone being a good debate partner, the other information becomes more so distracting or brings about expectations that will guide the conversation in a way that is not based on the substance of the argument itself.* - P06

Thus, many participants appear to navigate the following paradox: knowing too little, you risk dehumanizing them. Knowing too much, you risk the integrity of the conversation—and usually, participants lean toward minimizing the additional information they know about the user.

### 3.4.5 How do users consider the extra information provided by the designs?

#### 3.4.5.1 Shared subreddits

**Potential for humanizing users** Many participants stated, often enthusiastically, that viewing shared subreddits on the user card would remind them that there was a real person, a human being, on the other side of a conversation. For P08, shared subreddits would make them feel more connected to the user who is otherwise just a random stranger, and would likely to reduce anger and negative emotions.

*I think this would be very humanizing. I think you can see what kind of, you know, interests they have on Reddit outside of politics and the conversation that you're having... if it's happening*

*in a negative, or a politically charged conversation... you know when you're talking with someone anonymous, you can be a lot ruder, a lot more condescending, and there's not really consequences to it, but when you see this, I think for me, it would reduce my anger or my negative emotions.* - P08

**Potential for reducing stereotyping** Other participants explained that highlighting commonalities could help bridge the gap between partisans and see the person in a (relatively) more positive light.

*I think [shared subreddits] is helpful because outside of the political spectrum people do have common interests. So my feeling might be, well, okay, maybe this person is not so bad. They like technology, they share the same interests in sports.* - P17

**Potential for fostering good faith and common ground** Many participants felt that viewing the subreddits they shared with another user would help establish for themselves some common ground with the user. They explained that in conversations that get particularly heated, knowing that they share a common interest would help to build some goodwill.

*I think that could sponsor a little more goodwill among people, like, even if you have two people that are vehemently arguing with each other and calling each other–you know, flipping each other off verbally–if they find out that, oh, you have an ATV too, or a four wheeler and you'd like to go out, it could sponsor a little bit more goodwill, which I think could ultimately lead to better conversations, for sure. And it's a good idea.* - P05

**Concerns** Some participants were concerned that there might be few instances where the users actually share common subreddits, however, our data analysis (in Section 3.3.5.1) revealed that a sizable number of cross-partisan and co-partisan pairs participate in at least one common non-political subreddit. Also, a few participants explained that they would be inclined to look at the user card only if it was someone that they recognize or have spoken to earlier. They thought that this information would be less useful for one-time interactions.

#### 3.4.5.2 Active subreddits

**Potential for reducing stereotyping** As expected, some participants explained that knowing the other subreddits in which their interlocutor participates would help reassure them that the person is not fixated on politics and has other interests as well.

*[Showing active subreddits would help because] that'll tell me if you're not stuck in a particular way, that they do have other interests that could influence their thought process.* - P17

A few participants liked that they could quickly get an idea of the kinds of the subreddits in which the user participates. They explained that currently, they needed to scroll through a reverse chronological list of their past comments to get a sense of where they participated.

**Concerns** Participants expressed two significant concerns with the active subreddits component. First, some participants felt that some active subreddits could present them in a negative light. They feared that when others view that they participated in a fun subreddit such as a meme subreddit, they may not take their political arguments seriously or worse still, use their participation in those subreddits to discredit their arguments.

*[Would not like active subreddits because] I don't want it to be like, Oh, this person's trying to describe to me economics and they browse like, I don't know, but just the Jojo subreddit all day. You know, it's a very easy path for like judgment, I guess.* - P07

The other concern was that, in its current form, the active subreddits component simply displayed too much information about their activity on the site. Many participants suggested that providing a way to customize the subreddits shown on their own card or to opt out of displaying the active subreddits will allay these concerns.

When you're first reading a comment from people it's like an interaction at a bar–you want to give them enough so that they come over, but you don't want to sell them the house. - P13

However, another factor that might compound these concerns is that this design when deployed as a browser extension would result in an information asymmetry, users may not even know that their interlocutor is using the extension to view information about them. This is a serious concern that we expand on in Section 3.5.4.4.

#### 3.4.5.3 Comment highlights

Some participants liked the idea of viewing different dimensions of a person based on their top comments in other subreddits. Some even recounted past comments that became viral or were gilded (awarded Reddit gold) by other users which they would be proud to highlight on the user card. However, many participants raised two major concerns. First, participants were concerned that comment highlights based on karma points could produce a biased view of the person and cause easy judgement. They explained that most top voted comments were either extremely opinionated, controversial or partisan which might provide fodder for more conflict.

*If someone is passionate about one thing and not passionate about the other, or somebody could have an extreme opinion about one thing and not about another. So you would see the most extreme thing, maybe, as their top comment, and then now you get to judge people on their most extreme opinion. I don't think that would be a very good idea.* - P07

Others noted the comments would likely distract them from the actual conversation or may contain outdated beliefs which may color the viewer's opinion about them. Another major concern was that participants expressed feeling self-conscious about the information revealed in the comment highlights. It is telling that all four participants stating this concern were either female or genderqueer (P8, P13, P14, P18). They felt that the comment highlights made their comments more public and their profile more open to scrutiny. P14 expressed that they would likely choose how they word their comments carefully because of the increased visibility. P18 explained that she liked the way past comments were structured on the Reddit profile page, a simple list of past comments in reverse chronological order. It afforded her some amount of privacy by not being very organized or accessible.

*[I like the profile page] because I know that it's not necessarily that open and always accessible and when people are touching on touchy topics and they are expressing themselves, [they] might want to keep some sort of an anonymity, I feel like having things more presented in a way that shares more information could actually be a problem.* - P18

Further, P18 expressed concern that her views on one topic may be used against her when she is discussing other issues and said that she would likely have to make throwaway accounts to prevent users from connecting issues. Similarly, while P13 did not express specific concerns about the design, she had earlier described a prior experience where users racially abused her after finding out that she was a Black woman from a photo she had previously uploaded. This component likely exacerbates these concerns by increasing the visibility of specific comments.

#### 3.4.5.4 Karma points and awards

**Potential as a basic/weak good faith indicator** Many interview participants expressed that it did not matter if their interlocutor amassed high karma points and awards, as they used it not so much to determine if the person posted quality comments, but to simply indicate if an interlocutor was a troll.

**Concerns** Many participants explained that high karma points only indicated that the person makes good jokes or puns and it said nothing about the quality of someone's views. This somewhat lukewarm response to karma points is in line with Massarani's observation that while redditors place some value on karma scores, they are also suspicious of users with very high scores [144]. Almost all right-leaning participants pointed out that since Reddit has more liberal users, the karma points and awards usually only reveal how liberal the user is and therefore might bias the conversation. Hearing this initial feedback, we converted our karma indicator to display only if they had less than 100 karma for use as a very basic good faith indicator. Later participants told us that this information, coupled with information on the age of the account, was enough to identify

troll behavior.

## 3.5 Discussion

### 3.5.1 Education vs entertainment in cross-partisan discussions

We find that participants engage in political discussions both to educate and to entertain. Depending on external factors such as time constraints and outside news cycles, the same user may engage in relatively serious political discussions or may casually peruse the site and join in casual banter, satire and trolling. This finding is also in line with past work that shows that trolling behavior is context-dependent and not an immutable individual characteristic [35].

The two goals may be at odds with each other. The same comments may be appreciated differently by people seeking education vs. entertainment. Interventions that aim to coach users to talk to the other side (such as [240]) might help produce comments that are more effective for education than entertainment. Participants may also be more receptive to such coaching when they are primarily motivated by education rather than entertainment. On the other hand, the two goals can also be complementary. None of our participants joined Reddit for its political content and most had significant interests in other non-political subreddit topics. By hosting something for everybody, Reddit likely allows casual political observers who happen to peruse the site for other reasons to engage in political talk. Also, many participants commented that they often switched to lighter content when conversations go awry or when they simply needed a break from a heavy discussion. We speculate that this easy access to fun and entertaining content has therapeutic effects and serves to lighten the after-effects of serious political discussions. This recuperative function is particularly important given the participants' concern about the emotional toll of these conversations.

### 3.5.2 Unintended consequences of cross-partisan discourse?

The outcomes of cross-partisan interactions, both online and offline, have been typically evaluated in terms of highly valued outcomes such as political participation and political efficacy. However, little is known about the effects of cross-partisan interactions on users' emotional and mental well-being. From our interviews, we observe that most participants were weary about the discussions' repercussions on their mental health and employed multiple strategies to negate these effects. Thus, we call on researchers to attend to the psychological effects of participating in these discussions in addition to studying normative democratic outcomes, especially in these highly polarized times.

We observe that many participants, as a form of mental self-preservation, aim to have dispassionate discussions and sometimes even preemptively disengage if they feel that they or their

interlocutor is getting emotional, for fear that emotions could devolve into name-calling. They make a distinction between being "emotional" and "rational". However, this hyper rational, impersonal style of deliberation could have unintended consequences. Firstly, research suggests that taking emotions out of political discussions does not necessarily lead to more rational outcomes; while anger spurs aversion and leads to close-mindedness, when the emotional response is anxiety, people seek new perspectives and become open to compromise [135]. Anger, on the other hand, also increases political participation [224]. Secondly, as Young notes, "a norm of dispassionateness dismisses and devalues embodied forms of expression, emotion, and figurative expressions. People's contributions to a discussion tend to be excluded from serious consideration not because of what is said, but how it is said." [241] Clearly, this limits whose views are engaged with in cross-partisan interactions; users who are directly affected by the discussed issues likely passionately voice their opinions while those that are unaffected likely remain detached. Thus, the views of users who have the highest stakes may be less attended to. Finally, pure reasoning, with its emphasis on rationality as opposed to passion is known to be also exclusionary towards members of disadvantaged groups and individuals with less formal education as this form of communication is deliberately learned and developed [153]. Important questions around the outcomes of these conversations remain; does this kind of cross-partisan discourse contribute positively to building a deliberative democracy? In its current form, the prevailing hostility, toll on mental health, and the possible unintended consequences of participants' strategies to have deliberative discussions suggest otherwise–at least on Reddit.

### 3.5.3 Impacts of information about interlocutors: humanizing, stereotyping, judging, and attacking

In our interviews prior to showing the design probes, participants readily acknowledged that knowing more about others could be humanizing, allowing them to "see a little more humanity in what is otherwise a username and text" (P06). However, in current practice, participants described that they exclusively focused on the comment text and not on the author of the comment due to the fear that they may become prejudiced by viewing the author's past behavior or positions. Given that the Reddit interface does not provide means to only see humanizing information while avoiding prejudice-inducing information, participants resolve this issue by simply not viewing user profiles entirely. This reduces opportunities to build common ground and trust. We aimed to address this issue by showing potentially humanizing information about the user through our designs.

Upon viewing the user card, as expected, participants indicated that it would alter how they participate in conversations, making them consider both the comment and the comment's author when responding. In the case of shared subreddits, many predicted that this shift could humanize the in-

terlocutors and promote goodwill, with participants expressing that they would be more mindful of their own behavior and more charitable of their interlocutors' potential transgressions. However, participants also expressed many significant concerns about other ways that the information might be used. With active subreddits, they worried that other users might judge/stereotype them negatively for participating in casual subreddits such as meme subreddits and also felt this component would disclose too much information about them. For comment highlights, our female and minority participants were especially concerned about how these comments could provide more fodder to attack them. Some of these concerns can be addressed by providing users with more control over what information is shown about them on the user card.

However, more broadly, focusing attention on the user profiles, while potentially humanizing especially when users share group memberships, could have major negative implications especially for female and minority users by increasing visibility and inviting increased scrutiny on their profiles. While participants expressed frustration over Reddit's anonymity providing a safe harbor for trolls, they also appreciated how this anonymity allowed them to express opinions without being targeted.

### 3.5.4 Challenges and opportunities for future designs to improve cross-partisan discourse

Given the concerning feedback that we received, we decided not to proceed with building the proposed browser extension. Instead, we detail challenges in designing to improve cross-partisan discourse and opportunities we see to move forward in this space.

#### 3.5.4.1 Countering the different forms of hostility in cross-partisan discussions

Personal attacks and name-calling during cross-partisan discussions are commonplace online. Our designs aimed to minimise such occurrences by highlighting non-political group memberships to offset the effects of the out-group categorizations. However, it is important to consider other forms of hostility. For example, determined users can search through interlocutors' activity history to find material to disrupt the discussion and attack them. These concerns are particularly significant for many of our female and minority participants for whom partisan hostility often interacts with sexism and racism in cross-partisan interactions. The culture of harassment based on toxic conceptions of race, gender and sexual identities supported by Reddit's design and governance, which Massanari terms as "toxic technocultures" [143], negatively influence and exacerbate the hostility already prevalent in these political discussions. Future work on designs to improve cross-partisan discourse should attend to the multiple forms of hostility prevalent and how the designs to reduce hostility may differentially impact members of disadvantaged and marginalized social groups.

#### 3.5.4.2 Countering party-stereotypes without revealing user information

Our designs aimed to cross-categorize and decategorize at an individual level by making certain user activity more visible. While all participants expressed strong support for the shared subreddits design, as we showed in Section 3.3.5.1, not all user interacting pairs have common group memberships. Further, in the previous section, we highlighted some concerns about revealing user information, and one user went so far as to say she would make throwaway accounts to disrupt that. Thus, alternate approaches to de-stereotype without calling attention to individual profiles may be more effective. For example, Alher and Sood [3] found that people consistently overestimate the extent to which party supporters belong to party-stereotypical groups, sometimes by over 300% (for example, Atheists for Democrats and Evangelical Christians for Republicans) and correcting these misconceptions led to significant reductions in out-party hostility. Similarly, we could surface subreddit memberships in aggregate to counter some of these extreme stereotypes, for example, by showing that only (a surprisingly low) 5.34% of Reddit users who participate in r/Conservative also participate in r/Christianity (based on 2019 Reddit comment data). [7]

#### 3.5.4.3 Intervening in cross-partisan discussions

In recent years, researchers have developed algorithms to detect when a conversation is likely to go awry to encourage either the moderators or the conversation participants to possibly take course correction measures [32, 134]. However, from our interviews, participants' own attempts at de-escalating often either have little/no effect or cause more harm (Section 3.4.3). Designers aiming to intervene in individual conversations must critically evaluate if and when to intervene, taking into account possible adverse effects when those interventions do not work. It is likely that the potential harm caused by prolonging a negative conversation may outweigh the potential benefit of continuing that conversation. Given the low rates of success for turning around a conversation and the possibility of unintended harm, we recommend that designers explore preventive measures such as improving community norms around deliberation rather than corrective measures to improve individual discussions. Not every conversation can be "fixed", nor do they need to be. Given users' educational and entertainment motivations to continue to participate, a more productive aim may be to facilitate more of the kinds of conversations (Section 3.4.1) that people want to participate in.

#### 3.5.4.4 Information asymmetry

Our designs aimed to provide more information about interlocutors to facilitate better discussions. As outside researchers do not have access to the site, these designs are typically deployed as

[7] 69,343 users commented in r/Conservative, 55,342 users commented in r/Christianity and 3,705 users commented in both.

browser extensions or external apps. However, such a deployment would result in some users (who download the extension) having easy access to information about others, while other users may not even know that their interlocutors have access to their information. Further, even if the extension allowed users to customize or remove content on their user cards, users first would need to know that such an extension exists and download it. This will likely compound concerns that users already have about revealing more information about them. Given that cross-partisan interactions often turn into adversarial situations, one approach could be to apply an affirmative consent lens to design, centering individual agency with interactions structured around consent that voluntary, informed, revertible, specific, and unburdensome [92]. Thus, a possible modification could be that users be able to view subreddit participation details of only other extension users who consent to information sharing. This change may necessitate a user recruitment strategy where extension users have high likelihood of interacting with each other. To maximize the chances of such interaction, the deployment could be targeted to users participating in a particular subreddit, with consent from moderators and community members.

## 3.6 Limitations

Our study focuses on cross-partisan discussions on Reddit only; future work on other platforms will surely improve our understanding. Given our participants' strong support for showing shared group memberships between Reddit users who are essentially strangers, we expect that showing such connections on Facebook, especially between weak ties, will have a similar impact. However, we expect that showing other active group memberships will have little impact on Facebook as users already have access to some individuating information about others in the form of a real name, profile picture and cover photo, unlike on Reddit where users typically only identify themselves using a username.

Our study is US-centric and was conducted in highly polarized times, during the lead up to one of the most contentious presidential elections in history. In less polarized counties/settings, given that partisan identities will be less salient then, we speculate that these designs would actually have smaller effects on reducing hostility in interactions, as partisan group dynamics is unlikely to be the cause for the hostility. Alternately, approaches aimed at establishing more deliberative discussion norms through example setting [215] may be more effective as these norms would face little resistance from the hostile partisan norms that we observe today.

While in this work, we have focused on cross-partisan political discussions online, we do not contend that cross-partisan interactions are more important or should take primacy over other forms of political discourse that in some cases specifically exclude dissenting voices. As, Mansbridge et. al [137] note:

*Activist interactions in social movement enclaves are often highly partisan, closed to opposing ideas, and disrespectful of opponents. Yet the intensity of interaction and even the exclusion of opposing ideas in such enclaves create the fertile, protected hothouses sometimes necessary to generate counter-hegemonic ideas. These ideas then may play powerful roles in the broader deliberative system, substantively improving an eventual democratic decision.*

## 3.7 Conclusion

In this work, we have explored how users navigate the contentious political climate to engage in cross-partisan discussions. We find that participants have different, multiple motivations for engaging in these interactions, sometimes they prefer serious deliberative discussions and other times, they look for entertainment and banter. These different motivations coupled with the hyper partisan environment presents challenges to participants seeking to engage with "the other side". Through experience, participants have developed multiple strategies to have good conversations. From our design probes, we observe that participants find shared non-political subreddit memberships of their interlocutors humanizing, however, sharing other details such as other group subreddit memberships and past top comments raise significant concerns around privacy and misuse.

# CHAPTER 4

# GuesSync!: An Online Party Game To Reduce Outparty Hostility

## 4.1 Introduction

Over the past decade, social scientists have recorded a significant rise in affective polarization in the US – "the tendency of people identifying as Republicans or Democrats to view opposing partisans negatively and copartisans positively" [96]. Although many countries have deep political fissures, compared to other longstanding democracies, the US is exceptional in the levels of affective polarization observed today [24]. Increasingly, partisans ascribe negative stereotypes to the other side, calling them closed-minded, unpatriotic and immoral [51]. By the start of the Biden presidency, 72% of Americans reported believing that the opposing party is "a serious threat to the United States and its people" and 59% reported somewhat or strongly believing that the opposing party is "downright evil" [99]. This increase in affective polarization has important social, economic and political ramifications that threaten to tear the fabric of American democracy. Americans are more reluctant to talk to opposing partisans, even about nonpolitical topics [204]. Affectively polarized partisans are significantly less likely to be comfortable with outpartisans as friends or neighbors [94]. Affective polarization also influences economic decisions such as where people buy and how much they are willing to pay for goods and services [147]. In the political realm, affective polarization reduces trust in an outparty government and reduces support to compromise with outparty elites increasing partisan gridlock [84]. Further, a recent study highlights a link between affective polarization and specific policy positions. Researchers found that as partisan animus increases, Republicans are less concerned about COVID-19 and are less supportive of mitigation policies, though their opposition is tempered by the level of infections in their county [55]. Given its wide-ranging consequences, the high levels of affective polarization we observe today in US politics are extremely concerning. Therefore, social scientists have explored numerous approaches to reduce affective polarization, particularly outparty hostility.

A promising approach to reduce affective polarization is to correct misperceptions about outpartisans [62]. While Republicans and Democrats have deep differences, perceived differences between these groups have been exacerbated over the past few decades [129] because of a range of factors such as selective mass media coverage, the rise of partisan outlets, and social media [94]. Mass media coverage typically focuses on polarization [125], and the most extreme politicians are extensively covered. Partisan outlets show both elites and ordinary outpartisans as extreme [69]. Further, exposure to political discussions on social media, which are usually between strong partisans, also adds to the illusion that most outpartisans are extreme and have little common ground with the other party [13]. Thus, social scientists have explored correcting different misperceptions about the outparty such as ideological extremity [52], political engagement [52], party composition [3] and group meta-perceptions [118] to reduce outparty hostility. Although these approaches are effective in survey settings, how these interventions can be scaled up and applied beyond survey respondents to wider audiences is still a largely unexplored question.

In this work, we make two major contributions. First, expanding on prior work correcting misperceptions about the ideological extremity of outparty supporters, we examined whether correcting misperceptions about specific policy positions held by ordinary Republicans and Democrats reduces outparty animosity and increases willingness to engage with outparty supporters. Second, unlike prior studies, we tested the effectiveness of the intervention by incorporating misperception correcting information in an online party game that can be scaled up to potentially reach a large audience. In GuesSync! (a portmanteau of guess and sync), a game we designed and developed, two players form a team, and in each game round, they are both shown a question, but only one player is shown the correct answer. That player (clue-giver) must convey the correct answer to their teammate (guesser) by providing clues within certain constraints on how they can communicate. The team scores points based on how close the guesser's answer is to the correct answer.

To study the effects of playing the game, we performed a pre-registered between-subjects online experiment with 665 participants. Participants played one of three versions of the game: a control version where no questions about political views held by Republicans and Democrats were included, a mixed version where two out of seven rounds contained questions on political views and a fully political version where all seven rounds contained questions on political views [1]. After the game, participants answered a survey containing standard affective polarization outcome measures (feelings thermometer ratings and social distance) and a behavioral intention outcome (willingness to talk to an outparty supporter) along with measures of potential mediators and moderators. We also collected game experience-related measures to compare across the three game versions.

We summarize a few key results. We did not detect a statistically significant reduction in hostility towards outpartisans between the control version of the game and the two treatment versions

[1] As a short hand, we refer to the mixed and fully political versions of the game as treatment versions

(no main treatment effect). However, performing a pre-registered moderation analysis, we found that Democrats playing the treatment versions of the game expressed warmer feelings towards Republicans than Democrats playing the control version. In line with prior research, Democrats over-estimated Republicans' support for conservative political views and correcting them through the game resulted in warmer feelings towards Republicans. We did not observe a similar effect for Republican players, likely because our choice of game questions on Democrats' political views. Playing the mixed game version also increased the willingness to talk about political issues with outparty supporters compared to control. We also identified psychological reactance as a potential mechanism that might affect the effectiveness of depolarization interventions. Interestingly, we found no difference between game favorability ratings given by players playing the control version of the game and the two treatment versions, suggesting that adding more political questions to the game did not appreciably impact how fun and enjoyable the game was.

The rest of the paper is organized as follows. First, we detail related work on reducing affective polarization, correcting misperceptions and design work on pro-social games. Second, we introduce and substantiate our hypotheses. Third, we introduce the game, our design choices and how we developed the game content. Then, we describe our experiment, measures and analysis. Finally, we discuss the results of our study and its implications.

## 4.2 Background

### 4.2.1 Correcting misperceptions about Republicans and Democrats

A wide array of social science research has established that partisans perceive wider differences between the two parties than the actual difference that exists [129, 60]. More recently, there has been a growing consensus that correcting these misperceptions about the outparty is a promising approach to reducing affective polarization [62]. These interventions are aimed at addressing misperceptions related to two primary drivers of affective polarization [50]: (i) policy disagreements, suggesting that outparty animosity is a reflection of the extent of disagreement about salient policy issues [227, 169] and (ii) partisan identity, suggesting that outparty animosity stems from people's tendency to dislike and discriminate against outgroups to elevate their own group status [95].

For example, Ahler and Sood [3] found that people significantly overestimated the extent to which outpartisans belong to party-stereotypical groups (Democrats who are union members and Republicans who are Evangelical) and correcting for these misperceptions reduced outparty animus. Lees and Cikara [118] also demonstrated that people overestimated outgroup negativity towards the ingroup (group meta-perceptions) and correcting the inaccuracy reduced negative outgroup attributions. Similar group meta-perception corrections have been shown to be effective

across over 25 countries [190]. Druckman et al. [52] showed that, without any additional information, people imagine the typical outpartisan to be more ideological (liberal Democrat and conservative Republican) and more politically engaged than is the reality, resulting in outparty hostility. When outpartisans are described as moderate (and modal Republicans and Democrats are, in fact, moderate), people exhibit reduced hostility towards outpartisans. Thus, correcting perceptions of the ideological extremity of the outparty can reduce affective polarization. In this study, we expand on Druckman et al.'s study by correcting misperceptions about specific political views held by ordinary Republican and Democratic supporters rather than misperceptions about the ideological makeup of the parties' supporters. Focusing on supporters' views on specific political topics instead of broadly how liberal or conservative they are, allows us to present a more precise picture of the supporters' views and convey the complexity of their issue positions. This is especially important as analyses on ANES survey data [169] (appendix section 6) shows that only about 8% of partisans hold ideologically consistent positions across multiple issues such as abortion, gun control and welfare despite their prominence in electoral politics. Further, highlighting views on specific political topics provides more opportunities to establish common ground on a wide range of topics. Thus, we expect that playing a game providing corrective information about party supporter views can reduce affective polarization. By correcting misperceptions about outparty views, we also expect that participants will be more open to having conversations with outparty members.

One concern with correcting misperceptions about outparty members is the potential for a "backfire effect" where the correction entrenches people's belief in the misperception, especially in cases where the issue is salient or identity relevant [165]. However, as Nyhan notes in a recent survey article [164], these backfire effects are extremely rare. The aforementioned studies on misperception correction reducing affective polarization did not result in such effects. But even in the absence of backlash, Nyhan summarises that the effects of the misperception correction are only moderately effective owing to a range of factors: motivated reasoning towards claims that are more congenial, continuous elite and partisan messaging that reinforces misperceptions, lack of targeting fact-checking towards people with the most exposure to misinformation and low levels of cognitive ability and processing effort among the public. Although not addressing all these factors, our game design did not provide additional partisan cues that encourage partisanship-motivated reasoning. Further, by designing the game such that more accurate answers are incentivized, the in-built accuracy motivation likely makes individuals more receptive to the corrective information than default or motivated reasoning, as has been observed in survey experiments [225]. Also, since we correct misperceptions about party supporters' views and not factual beliefs (such as Obama being born in the US), we likely encounter less resistance to corrective information on these topics.

### 4.2.2 Pro-social Games

Over the past decade, designers and researchers have aimed to change attitudes and behaviors through game mechanics and gameplay. These persuasive games focus on a wide-ranging set of topics such as promoting personal behavior change such as exercise routines [132], increasing pro-social attitudes towards marginalized groups such as the homeless [191], and encouraging critical thinking about social issues such as poverty and AIDS [65]. These games have been designed to either be direct in their issue goals and game mechanics or be implicit and obfuscate the true intentions of the game.

By far, the most common approach is the direct one. These games are designed such that the characters and scenarios modeled in the game overtly promote the desired outcome. The assumption is that this design will "encourage and enable players to internalize, and transfer, the game's modeled beliefs and behaviors to real-life contexts." [101] For example, in *Darfur is Dying*, players take on the role of a refugee trying to find water in a desert to bring back to the refugee camp while evading being killed by the militia. Playing the game elicited greater role-taking and increased willingness to help Darfurian refugees than simply reading a text containing the same information [175]. *Spent* is another game that simulates a scenario where the players are single parents without a job or a home and need to survive on $1000 for the month. Testing on middle and high school students, playing *Spent* was found to have significantly increased affective learning scores, a measure of the internalization of positive attitudes towards homeless populations, even three weeks after playing the game [192]. However, these overt persuasion approaches may not always be effective and sometimes even backfire, causing more harm to the target populations. For example, researchers found that playing *Spent* led to players believing that poverty is personally controllable and did not promote positive attitudes towards the homeless among online adult and undergraduate study participants [189]. *Papers, please*, a popular game where players take on the role of an immigration inspector charged with restricting entry to a fictitious country decreased intention, subjective norms and self-efficacy to help immigrants [174]. Kaufman et al. [101] suggest that such explicit efforts may fail as they might trigger a psychological reactance [25] among players who may perceive that their freedom to think freely is threatened by an external agent. Such a state makes players more resistant to persuasion. Further, making persuasion attempts direct and on-message may hamper the players' ability to fully immerse themselves into the transformative experience of the game.

A recent alternate approach is to incorporate stealth interventions within the game. Popularized by Kaufman et al. [101], this 'embedded design' approach aims to effect change by incorporating the persuasive mechanism in an implicit and subtle way within the game mechanics or game context rather than making the persuasive message be the focal point of the game. They outline three embedding strategies: (i) intermixing, which interweaves and balances on-message and off-

message content to make the persuasion non-threatening and palatable. (ii) obfuscating, which uses framing or genre to hide the game's persuasive intent (iii) distancing, which employs fiction or metaphors to introduce a psychological distance between players' prior associations and the game's persuasive content. These strategies have proved to be successful in reducing stereotyping and prejudice. In *Buffalo*, marketed as a party trivia game, players flip a person card (such as scientist) and a descriptor card (such as female) and need to name a real or fictional person that fits the descriptions in the cards ('female scientist') as fast as possible. The game employs intermixing by mixing on-message (stereotype-breaking) descriptor cards with off-message ones. The game also obfuscates its persuasive intentions by presenting as simply a party game without the de-stereotyping framing. Experiments [100] suggest that the game reduces prejudice and stereotyping by encouraging "greater inclusiveness in players' representations of social identity groups."

In this work, we experiment with two versions of the game, a fully political direct persuasion version where all questions are about political views held by Republicans and Democrats that explicitly aims to correct political misperceptions and an indirect persuasion version that includes some political misperception corrections but is still largely nonpolitical. We compare the effects of these two versions on affective polarization measures and willingness to engage with outpartisans against a control version of the game containing no misperception correcting information.

We believe games can be especially effective in reducing affective polarization for the following reasons. First, they provide a ready-made "magic circle" [218], a separate social and psychological space that players enter into when deciding to play a game, where the rules and norms of the game are activated, which likely supersede at least for the duration of the game, the hostile partisan norms we observe today. Second, they likely attract a larger audience than other misperception-reducing interventions such as engaging in political discussions [130], which are most likely to be attended by the most politically engaged. Thus, the games have the potential to reach a large majority of the US population who do not engage in politics [115]. Finally, games also present varied and creative modes of interaction between players outside of back-and-forth text-based discussions, allowing designers to interweave interventions within game mechanics. Thus, we believe that games could offer an alternate venue to develop interventions to reduce affective polarization among ordinary citizens.

Related to correcting misperceptions, more recently, researchers have also experimented with designing games aimed at training users to identify fake news and misinformation. These games rely on inoculation theory, the idea that by pre-emptively exposing individuals to a small amount of misinformation or information techniques in a controlled environment, individuals develop psychological resistance to being persuaded by "real" misinformation attempts later on [131]. These games take a direct approach; players take on the role of an unscrupulous protagonist who aims to cause chaos through different manipulation techniques. Through gameplay, the players learn

about and build resistance to such techniques. Of particular relevance to our study is *Harmony Square* [186], a game where players play the role of a Chief Disinformation Officer whose job is to "ruin the square's idyllic state by fomenting internal divisions and pitting its residents against each other, all while gathering as many "likes" as they can." Among the five manipulation techniques, players learn to fend against is how nefarious actors focus on "polarizing audiences by deliberately emphasizing and magnifying inter-group differences." Post-game tests suggest that players develop immunity against posts that employ such manipulations and find such posts less credible. However, they do not test how playing the game might affect outparty hostility, the focus of our study.

## 4.3 Hypotheses and Research Questions

As detailed in Section 4.2.1, we expect that games delivering misperception-correcting information will result in lowering outparty hostility and social distance (standard measures of affective polarization [94]) and increasing willingness to engage with outparty supporters.

We do not have a prediction about whether the mixed or fully political versions might have larger treatment effects on the desired outcomes and do not test for them. Given the polarized current political climate and the ordinary Americans' disdain for partisan politics [109], an overt attempt to correct perceptions about party supporters may result in psychological reactance as described earlier (Section 4.2.2), which may result in reduced effectiveness of the intervention. At the same time, the mixed version of the game contains little corrective information. Participants could be distracted by other more interesting aspects of the game, resulting in smaller treatment effects. From a practical standpoint, we powered our study to detect a difference in measures of outparty hostility between the control and treatment game versions (more in Section 4.5.1). We do not expect the difference in treatment effects between the two treatment game versions to be large enough to be able to detect them. The primary purpose of this study is to compare both treatment game versions to the control version. Therefore, we formulate the following hypothesis:

**H1: Players playing the mixed and fully political game will exhibit lower outparty hostility, lower social distance and higher willingness to engage in conversations with outparty supporters than those playing the control version.**

We also do not have a prediction about which game versions the players will like more. However, knowing which game version players like more can inform future iterations of the game. Therefore, we test for differences between the control version of the game and the two treatment

versions with respect to game favorability ratings[2].

**RQ1: Are there differences between game favorability ratings provided by two treatment version game players and the control game players?**

### 4.3.1 Underlying mechanisms

We examine three potential mechanisms described in Section 4.2.1 that might mediate reducing outparty hostility: perceived commonality, party stereotyping and psychological reactance. By correcting misperceptions of outpartisans' political views, we expect the games to increase perceived commonality between the two political groups, which in turn would reduce outparty hostility. Similarly, by learning that party supporters do not hold ideologically consistent positions on every issue, we expect the games to reduce party stereotyping, which in turn would reduce outparty hostility. Finally, we test if the games induce psychological reactance that results in increasing outparty hostility. Players may feel that the game forces them to temper their opinions about outpartisans, and this perceived lack of freedom to think freely may result in the intervention backfiring. Therefore, we test the following hypotheses about potential mediators.

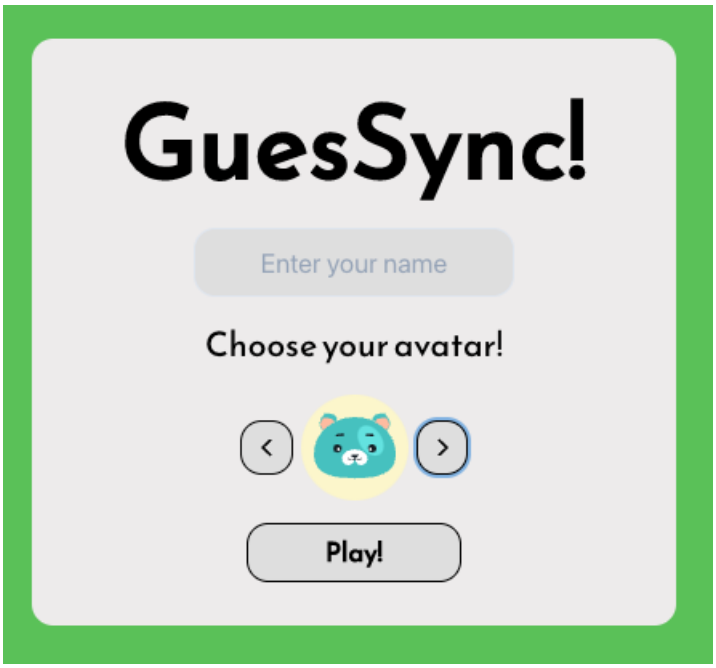
**H2: Perceived commonality, party stereotyping and psychological reactance mediate the effect of playing the games on feelings of outparty hostility.**

### 4.3.2 Subgroups of interest

We analyze the effects of playing the games on four key subgroup classifications: party identification, party strength, size of misperception and political knowledge. Given the significant differences between Republicans and Democrats, and especially considering that Republicans are becoming radicalized at a much faster rate [99], we examine if the game has heterogeneous effects on the supporters of the two parties. Also, research suggests that strong partisans, as a consequence of having a more ingrained partisan identity and stronger motivated reasoning, would be less inclined to moderate feelings of outparty hostility than weak partisans [110]. Thus, we examine potential differential effects for strong and weak partisans. Past research also suggests that higher political knowledge is correlated with stronger affective polarization [214]. Therefore, we compare the effects of playing the games on the high and low political knowledge groups. Finally, given that games aim to reduce misperceptions about party supporters' policy views, we examine if the

[2]Note that we had pre-registered to detect the difference in ratings between the two treatment versions. Instead, we compared the treatment versions to the control version as these comparisons can provide more direct insights on how adding more political content to the game affects favorability ratings.

Figure 4.1: Game landing page




games have differential effects on participants with high and low initial levels of misperceptions. Overall we examine the following research question:

**RQ2: Are there heterogeneous treatment effects of playing the game by party identification, the strength of partisanship, political knowledge and size of misperceptions?**

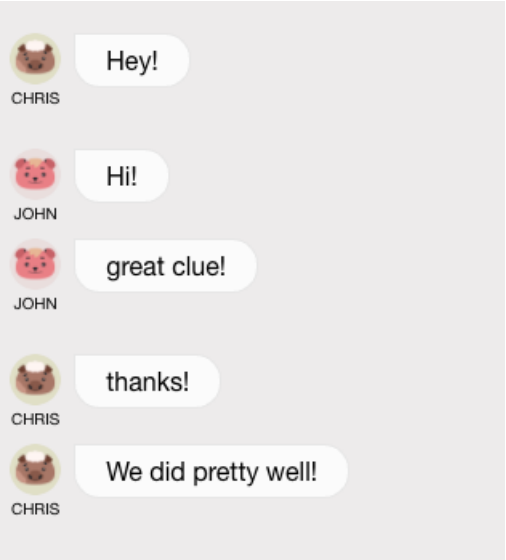
## 4.4 GuesSync!: A game designed to reduce affective polarization

### 4.4.1 Game details

GuesSync! is an online two-player cooperative party game. In the game, each player is randomly matched with another player. The game consists of multiple rounds. In each round, the two players are shown a question. They work together as a team, provide clues and guess the answer. The game design was inspired by two popular games: Family Feud[3], a popular cable network game where players work as a team to guess survey answers and Wavelength[4], a social guessing party game where teams try to read each other's minds using clues.

[3] https://www.familyfeud.com/
[4] https://boardgamegeek.com/boardgame/262543/wavelength

Figure 4.2: Chat window

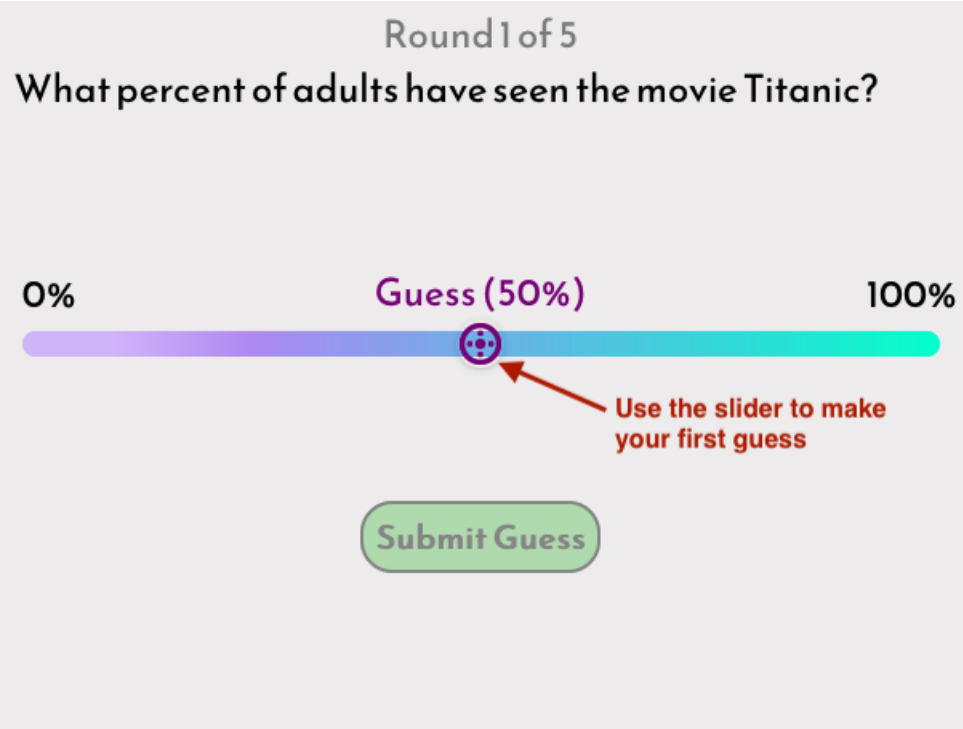



When a player lands on the game homepage (www.guessync.com), they select their game avatar and input a player name (Figure 4.1). Then, the player is shown a tutorial on how to play the game. After the tutorial, the player enters the matching lobby, where they are randomly matched with another player. Once matched, players can use the in-game chat to talk to their partner and start the game (Figure 4.2). Each game consists of seven rounds. Players play two trial rounds followed by five game rounds. The trial rounds are nearly identical to the game rounds except that they provide helpful tips on using the game UI and that no points are awarded. After the five game rounds, players view a game summary listing the total points they scored, and for each round, the question, the correct answer, the team's answer and the points scored. Each round consists of four phases: the initial guess phase, the clue-giving phase, the final guess phase and the grand reveal phase. We describe the four phases in detail:

#### 4.4.1.1 Initial guess phase

At the beginning of each round, both players are asked to independently provide their best guess answer for a question using a slider (Figure 4.3). All questions require players to guess a percentage amount, for example, 'what percent of adults have seen the movie Titanic?' Depending on the game version, some or all of these questions may be about party supporters' political views. For example, what percent of Republicans (Democrats) think that high-income individuals pay too little in taxes? For the mixed version, we randomly select one Democrat-related question and one Republican-related question, randomize the order of the two questions and place them in the second and fifth (final) game rounds. For the fully political version, we randomly choose between two

Figure 4.3: Initial guess phase



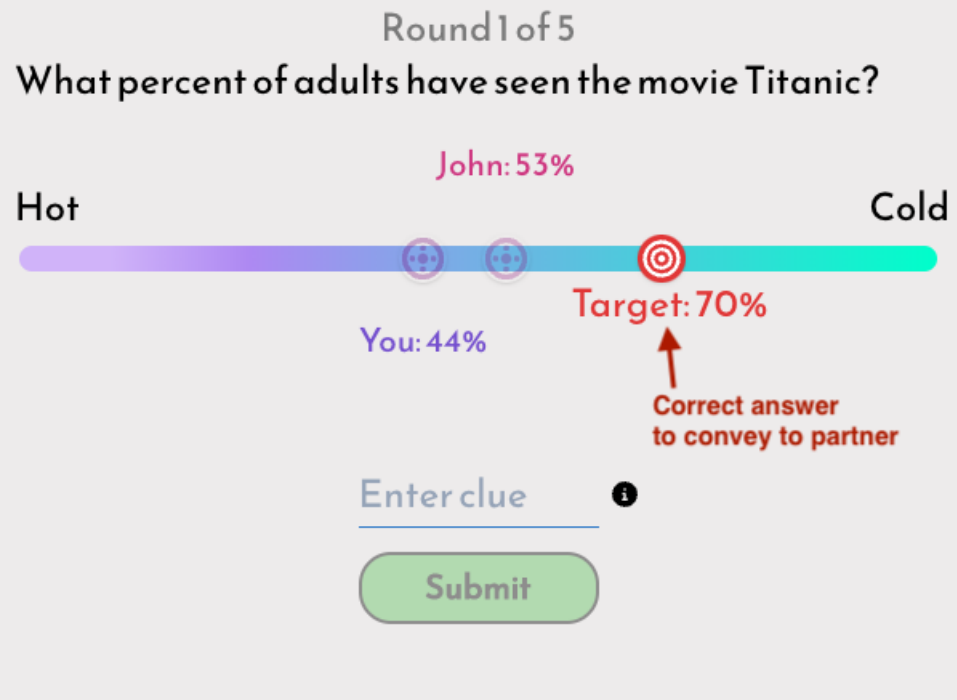

question configurations, three Democrat-related questions and four Republican-related questions or four Democrat-related questions and three Republican-related questions. Then, we select the questions at random based on the selected configuration, randomize their order and place them in the trial and game rounds. In the control version, no political questions are included in the game. Players are given 60 seconds to come up with their best guess for the answer.

#### 4.4.1.2 Clue giving phase

Then, the game assigns one player as the clue-giver, and the other player is the guesser. The game reveals the correct answer only to the clue-giver. The clue-giver must convey the correct percent to the guesser using a scale provided by the game, for example, a hot-cold scale. The clue-giver needs to develop a clue using the hot-cold scale to help their partner guess correctly (see Figure 4.4). Here, a good clue would be something the partner can identify as being more cold than hot as the target is closer to the cold end of the scale. 'Lemonade' might be a good clue for this example since it's usually consumed cold. If the correct answer was 5% (close to hot), 'sun' might be a good clue. If the answer was 95% (close to cold), 'arctic' might be a good clue. The scales change in each round, and the players take turns being the clue-giver and guesser.

The clues must be only one or two words long, cannot have more than 20 letters, and cannot include numbers nor quantifier words such as lot and little or direction-related words such as left

Figure 4.4: Clue giving phase




and right. Guesses also cannot include words like same and correct that convey the answer without using the provided scales. We maintained a blocklist of such words to ensure that players used clue words that were conceptually on the scales provided. The clue-giver is given two minutes to enter their clue. While the guesser waits for the clue, they are also provided the scale and the clue-giver's initial guess. They can use this time to think of potential clues the clue-giver might give and what percentage the clues might correspond to.

#### 4.4.1.3 Final guess phase

Once the clue-giver inputs the clue, the guesser must interpret the clue according to the scale and input their team's final answer (Figure 4.5). The guesser is given 60 seconds to make their final guess.

#### 4.4.1.4 Grand reveal phase

After the final answer has been submitted, the correct answer is revealed to the guesser, and the final guess is revealed to the clue-giver (Figure 4.8). Points are awarded based on how close the final guess is to the correct answer. Teams get 5 points if their final guess is within 5% of the correct answer, around the margin of error in these surveys. Teams get 2 points if their final guess is within 10% of the correct answer. Players can talk to each other in this phase through the in-

Figure 4.5: Final guess phase

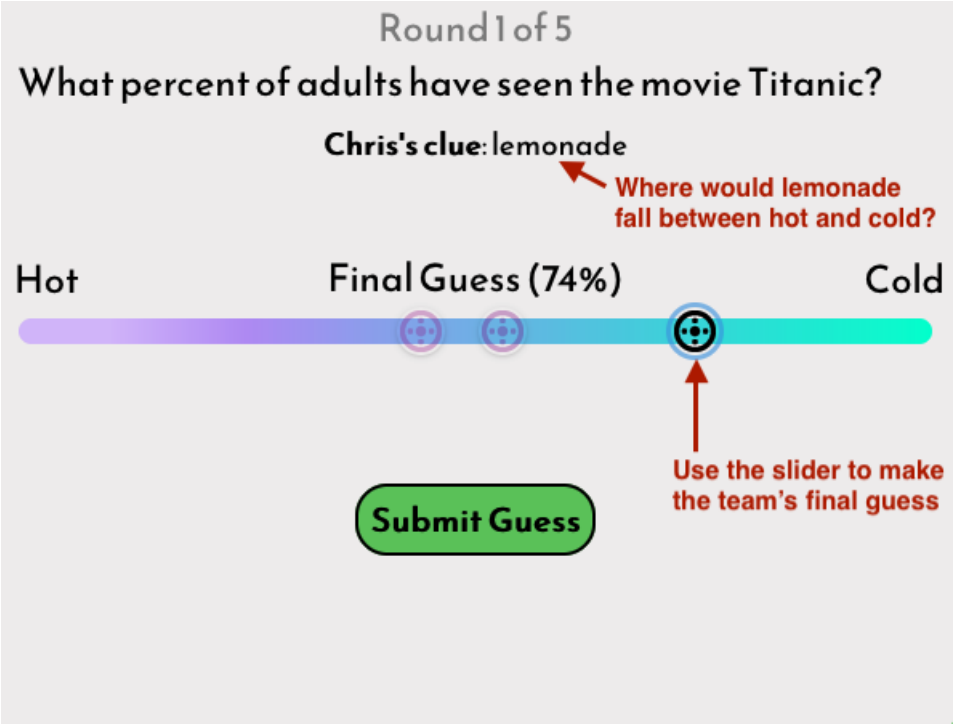
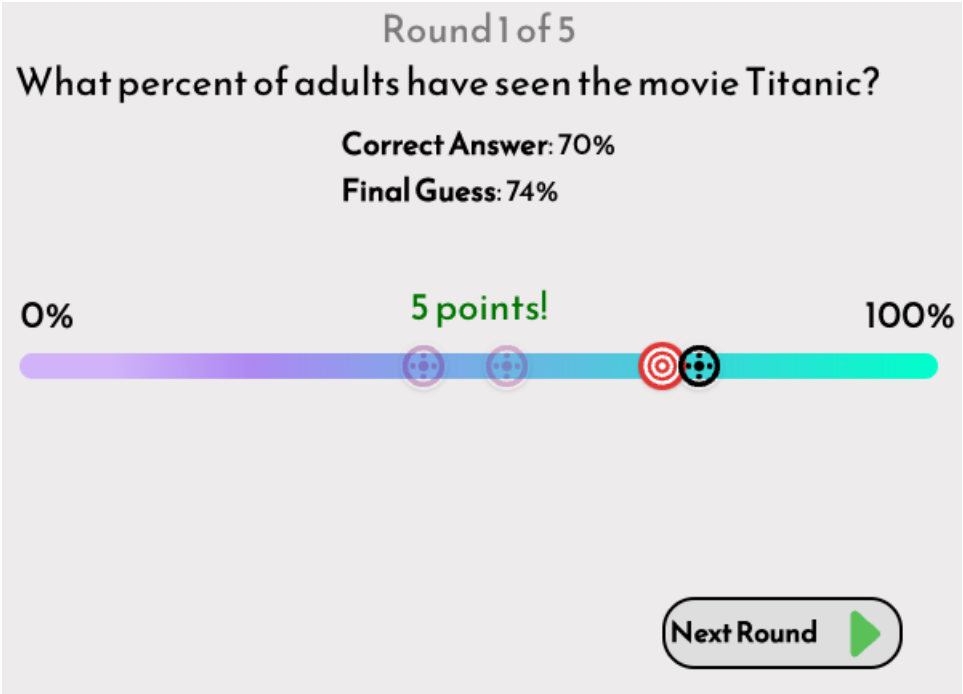



Figure 4.6: Grand reveal phase



game chat window. Players can either type into the chat or choose one of the game-suggested text input prompts (for example, 'great job!', 'good clue'). The chat is disabled during the other three phases of the game.

### 4.4.2 Key game design decisions

In developing GuesSync!, we made several key design decisions to maximize the game's effect on outparty hostility. We discuss those decisions below:

1. No prior political knowledge needed: The game was designed so that players do not need to know the answers to questions to enjoy the game. We deliberately avoided designing it as a political trivia game as such games likely attract only individuals with high political knowledge. When answering political questions in the game, while knowledge of politics may help, the game primarily revolves around players being able to provide clues based on the scale provided and their partners being able to interpret the clues accurately.

2. Minimal partisan cues: The game was designed to avoid presenting partisan cues which may cue partisan motivated reasoning and bias. We do not ask about the players' political leanings at any time during the game. Further, the avatars that the players can choose for themselves are cute animals [5] instead of humans as demographic details may also cue partisan identities since the two parties are also increasingly sorted along racial lines.

3. Ask then reveal design: In each round, the game first asks both players to independently answer a question with the correct answer revealed through the course of the round. This approach provides dedicated time at the beginning of the round for players to reflect and input their best guess answer. By being asked a question and having to work for the answer (in the clue-giver's case, having to translate the correct answer to a concept), players likely engage more deeply with the question than when they are directly provided the answer, as is the case of news reports.

4. Interactive design: Players provide all their percentage answers using sliders. Studies [166] have shown that the physical act of clicking and dragging sliders as opposed to simply clicking or hovering creates an immersive experience resulting in cognitive absorption, a state where the person is "consciously involved in an interaction with almost complete attentional focus", which in turn is associated with being more receptive to persuasion.

5. Slow thinking: We allocated fairly liberal time limits for each game phase. We provided one minute for players to provide their initial guess, two minutes for the clue-giver to come

[5] derived from https://github.com/roma-lukashik/animal-avatar-generator

up with a clue and one minute for the guesser to provide the final guess. We did not want the game rounds to have rapid-fire style interactions that likely resulted in top-of-the-head responses. By providing adequate time to think through, we allow for slow thinking and more considered responses which experiment data suggest result in smaller degrees of misperceptions [2].

6. Credibility: To increase the credibility of the game answers shown, we state that the answers are from reputed nonpartisan sources such as Gallup and YouGov both at the beginning of the tutorial and at the end when players view a summary of the game.

7. Team interactions: Players could optionally chat with their teammates before and after each round. The feature allows players to interact and connect with their teammates. Critically, it also provides opportunities for in-game discussion and reflection of especially surprising answers, which can aid retention [152]. Though chatting was optional, the median number of comments made by players in the experiment was 6.

#### 4.4.2.1 Playtesting

To refine the design, we playtested the game in two phases. First, we recruited eight players through TurkerNation, a collective of crowdworkers on the Amazon Mechanical Turk platform. After providing informed consent, they played the game online, completed the post-game survey and were interviewed by the first author to obtain their feedback on ways to improve the game. Together, the game and interviews took about 30 minutes. Participants were paid $7 to complete both the game and the interview. In the second phase, we playtested the game directly on the MTurk platform. Eighteen workers completed the game and the post-game survey. We paid $3.75 for their participation. In both playtesting phases, we collected all inputs that players used in the game, including their initial and final guesses to questions, the clues they provided and their chat messages. Through playtesting, we refined the game in the following ways:

1. Based on interview feedback that it took a few rounds initially for players to learn how to play the game, we added two practice trial rounds to the game. While functionally identical to the other game rounds, these rounds were not scored and included helpful tooltips and instructions on how to use the game interface. They were also helpful for players to get in sync with their partners.

2. From the list of clues provided during the game, we inferred that some players did not use the scale to provide clues and instead used quantifier words such as lot and little and direction-related words like left/right and higher/lower to convey the correct percent. We added these words to our existing blocklist of clues.

3. In an initial version of the game, we did not have any time limits to guess the correct percent or to provide clues. However, to keep the game moving and to detect when a player left the game midway, we had to institute time limits. As discussed earlier, it was important to provide enough time for the players to think through their answers instead of responding on the fly. We settled on providing a minute for players to input their guesses and two minutes for them to provide clues. One concern with providing a lot of time is that while the clue-giver takes time to come up with a clue, the guesser will have to wait and might lose interest. However, through our interviews, we found that the waiting period increased anticipation and added to the excitement. As one participant put it, "it was like waiting to open a Christmas day present... If [the clue] took a while, it must be a doozy!" During the waiting period, we also included a nudge "use this time to think of possible clues that [your partner] might come up with." to keep the guesser focused on the game.

4. We also made other minor UI changes to the game such as updating the game instructions with clearer directions, providing information on the number of rounds completed and how many more rounds to go to finish the game, and adding a game summary page after completing the game containing all the game questions, answers and points scored.

### 4.4.3 Selecting game questions and scales

The game requires three main components: questions on party supporters' political views, nonpolitical questions and scales. We used crowdsourcing and publicly available surveys to select these components. We describe the process below:

#### 4.4.3.1 Selecting questions on party supporters' political views

To obtain an initial set of questions on party supporters' political views, we used nationally representative survey data from the 2020 American National Election Studies (ANES) Time Series Study, 2020 Cooperative Election Study (CES) and the 2021 General Social Survey (GSS). We also considered selecting questions from Pew Research. However, since they have a robust website reporting on their surveys, the answers to their questions were indexed by Google, making it very easy to find using Google search. So we opted not to use their survey questions in the game. However, note that searching for answers on the Internet to play the game, while possibly reducing the fun of playing the game, would still result in players being exposed to the corrective information. We manually selected all questions on political views from these sources and, using the survey data, obtained the percentage of Republicans and Democrats who held those views. In

Table 4.1: Game questions on Republican party supporters' political views

| What percentage of | Survey estimate | Median Dem. answer | Median Rep. answer |
|---|---|---|---|
| Republicans say they would be pleased if the supreme court reduced abortion rights? | 43 | 85 | 75 |
| Republicans say that abortion should never be permitted? | 19 | 80 | 50 |
| Republicans are willing to open up protected nature areas for economic development? | 16 | 58 | 40 |
| Republicans say that the US spends too much on alternative energy sources? | 23 | 79 | 47 |
| Republicans support laws that protect gays and lesbians against job discrimination? | 81 | 27.5 | 40 |
| Republicans support requiring background checks for gun purchases at gun shows or private sales? | 82 | 26 | 40 |
| Republicans say that the government should make it easier to buy a gun? | 11 | 70 | 57 |
| Republicans say that the US spends too much on the nation's health? | 16 | 60 | 25 |
| Republicans support making all unauthorized immigrants felons and sending them back? | 24 | 70 | 37 |
| Republicans support sending back children who were brought to the US illegally and have lived here for 10+ years? | 21 | 70 | 25 |
| Republicans say that the federal minimum wage should be decreased? | 4 | 60 | 12 |
| Republicans oppose requiring employers to offer paid leave to parents of new children? | 13 | 60 | 10 |
| Republicans say that the police officers never use more force than necessary? | 3 | 50 | 50 |
| Republicans support requiring police officers to wear body cameras while on duty? | 88 | 40 | 75 |
| Republicans say that blacks face no discrimination at all in the US? | 5 | 45 | 56 |
| Republicans believe that the legacy of slavery affects the position of black people in society today? | 68 | 26.5 | 27 |
| Republicans think that high-income individuals pay the right amount in taxes? | 29 | 70 | 53 |
| Republicans say that eligible voters are never denied the right to vote? | 23 | 90 | 70 |

Table 4.2: Game questions on Democratic party supporters' political views

| What percentage of | Survey estimate | Median Dem. answer | Median Rep. answer |
|---|---|---|---|
| Democrats believe that climate change has been mostly due to human activity? | 69 | 90 | 90 |
| Democrats are unwilling to pay much higher prices in order to protect the environment? | 17 | 40 | 40 |
| Democrats support the death penalty for convicted murderers? | 44 | 32 | 20 |
| Democrats oppose making free trade agreements with other countries? | 7 | 41 | 30 |
| Democrats support lowering the eligibility age for Medicare from 65 to 50? | 77 | 60 | 30 |
| Democrats feel that courts deal too harshly with criminals? | 40 | 60 | 70 |
| Democrats say that the US spends too much on reducing crime rates? | 8 | 29.5 | 45 |
| Democrats believe that the legacy of slavery affects the position of black people in society today? | 97 | 82 | 70 |
| Democrats think that high-income individuals pay too little in taxes? | 75 | 90 | 30 |
| Democrats say that transgender people face no discrimination at all in the US? | 1 | 11 | 30 |
| Democrats support requiring showing a government photo ID when voting? | 48 | 50 | 25 |
| Democrats say that eligible voters are never denied the right to vote? | 7 | 32 | 36 |
| Democrats say that the US spends too little on assistance to the poor? | 44 | 66.5 | 75 |

total, we formulated 556 questions related to the political views of the two political groups. We used the following procedure to select a subset of the 556 questions to be used in the game:

1. Using an MTurk survey, for each question, we obtained the workers' best guess answer and their ratings on a 5-point scale on how important the question topic is for them (importance rating). We obtained at least five responses for each question from Republican and Democrat workers who were identified from a prior qualification survey. We paid $0.10 per worker per question.

2. Then, for each question, we calculated separately for Republican and Democrat workers, the median importance rating, median quality rating, median guess percent and the absolute difference between the correct percent (from the surveys) and the median guess (size of misperception).

3. Since each question in the game would be viewed by both Republicans and Democrats, we needed to select questions that were important to both groups. However, as we were interested in influencing feelings towards outparty, we weight the importance rank provided by outparty workers more than inparty workers in the selection process. Thus, we selected

questions that the party supporters had the highest levels of misperception on issues they considered the most important.

4. For questions about Republicans' political views, we select the top 20 questions based on the size of Democratic workers' misperceptions that were assigned at least a median importance rating of 4 points by Democrats and at least a median importance rating of 3 points by Republicans. We similarly obtained questions on Democrats' political views. In total, we have 40 questions, 20 Republican-related and 20-Democrat-related.

5. The 40 questions were reviewed manually and were lightly reworded for clarity and brevity. Questions that were too long, confusing or contained double negatives were removed. Then, we manually assigned each question to a topic. To reduce variance in treatment outcomes, we selected a maximum of two questions per topic in the final pool of questions. In each game, the game algorithm randomly selected the topics first (two topics in the mixed version and five topics in the fully political version), and from each topic selected one question.

6. In total, 31 questions were selected for the game (Tables 4.1 and 4.2). We also include the survey estimate and the median answers provided by Republican and Democrat workers to those questions in the survey. Note that these political views largely correspond to policy or indicate the kind of policies the two political groups support.

Table 4.3: Non political game questions

| What percentage of | Survey answer |
|---|---|
| Adults say they would like to bring back dinosaurs? | 12 |
| Adults say that chocolate glazed donuts are their favorite donuts? | 12 |
| Adults in a relationship met their partner online? | 12 |
| Adults have at least one tattoo? | 26 |
| Adults are single? | 31 |
| Adults consider a hotdog to be a sandwich? | 33 |
| Adults believe in ghosts? | 36 |
| Adults like their eggs scrambled? | 37 |
| Adults believe in UFOs? | 39 |
| Dog owners got their dogs from a shelter? | 40 |
| Adults set an alarm but do not snooze when waking up? | 40 |
| Pet owners dress up their pets for halloween? | 45 |
| Adults say they drink coffee everyday? | 62 |
| TV-owning adults watched Neil Armstrong set foot on the moon?" | 94 |
| Adults say they have had a teacher who changed their life for the better? | 51 |
| Households are dog owners? | 54 |
| Adults in a relationship say they are satisfied with their relationship? | 94 |

#### 4.4.3.2 Selecting nonpolitical questions

We selected nonpolitical questions from publicly available nationally representative surveys conducted by YouGov and Ipsos available on their websites. We manually identified 54 questions from seven broad, largely nonpolitical categories: pets, relationships, supernatural, entertainment, hobbies, food and lifestyle. We selected a subset of questions using the procedure detailed below.

1. We again ran an MTurk task showing workers a question and asking them to provide their best guess answer, along with ratings on a 5-point scale on how curious they were about the answer to the question, how difficult they found the question and how they would rate the quality of the question if they saw it in a party game. We obtained ratings from 5 workers per question, paying \$0.10 per worker per question.

2. We selected all questions that received a median curiosity rating of 4 or above. Since the curiosity rating and quality rating were heavily correlated ($r = 0.83$), we used only the curiosity rating as a threshold. We decided against using a difficulty threshold as the players in the game would also be provided with clues to help answer the questions.

3. To ensure that the game included questions for which the correct answers balanced (some answers below 50% and some above 50%), we removed and replaced some questions that had answers below 50%.

4. In total, we selected 17 questions to be used in the game (Table 4.3). In the control version of the game, all seven questions (two practice questions and five game questions) were randomly selected from these 17 questions and for the mixed version of the game, five questions were randomly selected (two practice questions and three game questions). In the fully political version of the game, no nonpolitical questions were included.

#### 4.4.3.3 Selecting scales

To select the scales to be used in the game, we followed the following procedure:

1. Drawing on word lists and Wavelength game cards, we constructed 30 scales which largely consist of two words that are antonyms, for example, tall-short.

2. We ran a short MTurk task asking workers to come up with clues to identify different positions on a given scale. After using the game scale, workers were asked to rate on a 5-point scale the difficulty in coming up with the clues (difficulty rating), their confidence that someone looking at their clues would be able to identify the original positions on the scale

Table 4.4: Game scales

| Scales |
|---|
| Old - New |
| Mild - Spicy |
| Skill Luck |
| Nature - Nurture |
| For kids - For adults |
| Need - Want |
| Deserted - Crowded |
| Safe - Dangerous |
| Sport - Game |
| Flashy - Modest |
| Formal - Casual |
| Dog name - Cat name |

(confidence rating), and the overall quality of the scale if they saw it in a party game (quality rating). We obtained ratings from 5 workers per scale, paying $0.30 per worker per scale.

3. We selected all scales that workers gave a median difficulty rating of more than two and a median quality rating of 4 or 5. Since the confidence rating and quality rating were quite correlated (r=$-0.6$), we did not use the confidence rating.

4. In total, 12 scales satisfied these thresholds (Table 4.4). Of the 12 scales, we used two relatively easy scales, old-new and mild-spicy, during the trial rounds for all teams to allow them to ease into the game. Five scales were randomly selected from the other ten scales for the five game rounds.

In all the above three tasks, we limited the MTurk participant pool to only US-based MTurk workers who had at least a 98% task acceptance rate and had completed at least 1000 tasks.

### 4.4.4 Game Development

The game was developed entirely using Javascript and React, building on the codebase of an open-source version of the Wavelength game [6]. The game was hosted using the Google Firebase platform: we used the Realtime and Firestore Databases to store game data and Cloud Tasks to manage matching users in real-time. We used StreamChat[7] library to facilitate in-game chatting.

[6]https://github.com/cynicaloptimist/longwave
[7]https://getstream.io/chat/

## 4.5 Experiment

To test the hypotheses and research questions described in Section 4.3, we performed a pre-registered[8] between-subjects experiment on Amazon Mechanical Turk where participants were assigned to one of three game versions. The University of Michigan IRB reviewed the study and determined that it is exempt based on federal exemptions 3(i)(A) and 3(i)(B).

### 4.5.1 Power analysis

#### 4.5.1.1 Contextualizing the difficulty of reducing outparty hostility

Reducing outparty hostility in this highly contentious political climate is a hard task. Further, this experiment was conducted in the month of May, 2022, a week after a leaked Supreme Court draft opinion signaled that the Court was ready to overturn *Roe v Wade*, striking down Americans' right to have an abortion. Releasing the draft opinion increased mobilization on both sides around abortion rights, likely causing partisans to double down on their beliefs about the other party, making reducing outparty hostility harder. Even before this event, many efforts to reduce outparty hostility through survey experiments have had modest effects [243, 127, 232]. In conducting affective polarization research, scholars typically measure outparty hostility using multiple survey instruments such as a feelings thermometer, social distance and trait battery. In this work, we use two measures, a feelings thermometer that measures, on a 101-point scale, how warm or cold respondents feel about Republicans and Democrats, and a social distance measure that gauges how comfortable respondents are with having outpartisans as their friends, neighbors and son/daughter-in-law. We turn to prior studies to determine a reasonable effect size to detect. Priming American identity by reading and writing about America's strengths improved outparty feelings by about 5 degrees [126]. Similar effect sizes are observable when correcting misperceptions about group composition [3], highlighting warm interactions between party elites [89] and intergroup contact [188, 232]. These interventions are somewhat lightweight and administered through an online survey. In contrast, an elaborate experiment intervention, one of the most successful to date, which facilitated in-person cross-partisan discussions for about 15 minutes, improved outparty feelings by 10 degrees [130]. Given that our intervention through a game is much less intense and subtle, we expect it to reduce outparty hostility by 5 degrees, much like the aforementioned online survey interventions.

[8] https://aspredicted.org/blind.php?x=SZK_F1H

#### 4.5.1.2 Sample size considerations

Therefore, we powered our experiment to detect a 5-degree increase in feelings thermometer ratings towards outparty supporters at 80% power and an Alpha level of .05 based on simulations using the ANES 2020 dataset. A higher feelings thermometer rating indicates more warmth and less hostility. The power analysis indicated requiring 225 participants per experiment condition. Given the potential for dropoffs, we decided on recruiting about 250 participants per condition, accounting for a 10% dropoff. We note that our study is underpowered to detect subgroup effects, and all moderation analyses must be considered exploratory.

### 4.5.2 Recruitment and experiment procedure

This experiment was performed in 27 batches (May 12, 2022 - May 31, 2022) as it required players to be present online at the same time. The median number of participants per batch was 21. In each batch, participants were randomly matched and assigned to play the control, mixed or fully political game versions.

Approximately one hour before the start of each batch, we published a task (Human Intelligence Task, HIT in MTurk parlance) where workers indicated if they were available to play the game at a proposed time and if they could use a laptop or desktop to play the game (as we did not support playing the game on mobile devices.). Participants were also informed that the game would close for new players within five minutes of the scheduled time. This was done to ensure that most participants would start the game simultaneously and be matched with another participant. In the scheduling task, we collected demographic details such as age, gender and 7-point party identification scale[9] and how often they played party games. From the 16th batch onwards, we included a simple captcha-type question to ensure that the players were real people (not bots) and could follow English instructions[10]. Participants satisfied the following conditions were invited to play the game: (i) they correctly completed the captcha question (if shown to them), (ii) they indicated being available at the said time and could use a laptop or desktop, and (iii) they were not political Independents[11]. We limited this scheduling task to only US-based MTurk workers who had at least a 98% HIT acceptance rate and had completed at least 1000 HITs. We also excluded workers who playtested the game or participated in any game content creation tasks described in Section 4.4.3. We also excluded workers who had previously completed the scheduling HIT in an

[9]Note that asking for their party identification could increase the salience of partisan identities. Since we asked this question along with other demographic questions about an hour before the actual game, we expect it to have minimal impact on the game.

[10]We included this question after a few workers through free-text feedback in the post-game survey said they weren't sure if their partners understood English.

[11]We did not invite political independents as we do not have a clear hypothesis on how they would engage with the game.

earlier batch. All workers, regardless of whether they were invited to play the game, were paid $0.10 for completing this scheduling task.

Then, 10 minutes before launching the game task, we sent a notification through the MTurk platform reminding them of the game start time and providing instructions on finding the game on the platform. At the said time, we launched the game and again sent a reminder that the game was launched. The game closed for new participants 7 minutes after the game was launched (2 minutes more than the 5 minutes in the scheduling instructions to allow for stragglers). After providing informed consent, participants land on the game home screen where they choose an avatar and provide a game name. Then, participants were provided a tutorial on playing the game and provided multiple examples. From batch 10th, we changed the tutorial such that participants had to spend at least a minute on the tutorial before moving on to the matching screen[12]. In the matching phase, participants were matched with another participant (if available) to form a team, and the team was randomly assigned to one of the three game versions. If participants were not matched with another person in three attempts, they were provided $0.50 as compensation for their time. Only 50 participants were not assigned a partner and had to leave.

Once matched, participants played the game as described in Section 4.4.1. In the experiment, if a player did not provide a valid input for more than 90 seconds in the two guessing phases or did not provide a valid clue or hit the pass button after 150 seconds in the clue-giving phase, we assume that the player has left the game. In that case, we redirect their partner to the post-game survey to complete the HIT. When players completed the game, they filled out a post-game survey to complete the HIT. We paid $3.75 to all participants who completed the HIT. In total, 777 participants completed the post-game survey. Of the 777, 103 participants completed the survey after their partner left the game mid-way. There was no major difference in dropoffs across the three conditions. 31 control, 36 mixed and 36 fully political version players dropped off the game mid-way. Among the 674 participants, nine indicated that they were political Independents in the post-game survey and were removed from the analysis. In total, for our analysis, we have data from 665 participants: 224 control version players, 225 mixed version players and 216 full version players.

### 4.5.3 Measures

#### 4.5.3.1 Outparty feelings

We measured feelings towards outparty supporters using the 0–100 feeling thermometer ratings (lower ratings represent colder/unfavorable ratings and higher ratings represent warmer/favorable

[12]We included this stipulation as some participants complained that they or their partners did not fully understand the game instructions. We also updated the instructions with more examples.

feelings). We collected feelings thermometer ratings towards Republicans ($M = 45.05$, $SD = 24.60$) and Democrats ($M = 37.59$, $SD = 24.43$), and used the participant's party affiliation to determine outparty feelings (overall, $M = 39.72$, $SD = 24.70$).

#### 4.5.3.2 Social distance

To measure social distance, we used a standard scale measuring how comfortable/upset the participant would be with having an outparty supporter as a close friend, neighbor or relative ($\alpha = 0.83, M = 2.93, SD = 0.76$).

#### 4.5.3.3 Willingness to talk to outpartisans

We used two items to measure willingness to engage with outparty supporters on 5-point scales. We asked how willing participants were to have political conversations with outparty supporters ($M = 3.43, SD = 1.30$) and how willing they were to have nonpolitical conversations with outparty supporters ($M = 4.28, SD = 0.95$). Since the Cronbach $\alpha$ was low (0.55) for these measures, we did not combine them.

#### 4.5.3.4 Perceived commonality

To gauge perceived commonality, we used Levendusky et al's [130] two-item measure asking participants how much they agree on the two statements on a 5-point scale: "There are many policy areas where Democrats and Republicans agree and can find common ground to work together." and "Democrats and Republicans agree on many more issues than the media says that they do." Since the two items were highly correlated, we combined them by taking their mean ($\alpha = 0.80$, $M = 3.39$, $SD = 0.96$).

#### 4.5.3.5 Outparty stereotyping

To measure outparty stereotyping, we used a two-item measure asking participants how much they can tell about a person's political policy preferences by knowing that they are an outparty supporter and how much they can tell about a person's other values and goals by knowing that they are an outparty supporter. Since the two items were highly correlated, we combined them by taking their mean (higher means more stereotyping, $\alpha = 0.80$, $M = 3.31$, $SD = 0.90$).

#### 4.5.3.6 Psychological reactance

We modeled our psychological reactance measure using Moyer-Gusé et al.'s cognitive reactance scale [155] on measuring reactance to persuasive messages. Using a three-item measure, we asked

participants how pressured, manipulated and forced they felt to form certain viewpoints about Republicans and Democrats. Since the three items were highly correlated, we combined them by taking their mean (higher means more reactance, $\alpha = 0.94$, $M = 2.18$, $SD = 1.20$).

#### 4.5.3.7 Demographics

We collected participants age (18-24: 3.61%, 25-34: 36.99%, 35-50: 41.95%, 51-65: 15.19%, 65+: 2.25%), gender (male: 54.74%, female: 44.51%, 3 participants non-binary and 2 preferred not to disclose), race (Caucasian/White: 80.75%, Asian/Pacific Islander: 6.91%, Hispanic/Latino: 5.86%, African-American/Black: 5.56%) and political party identification in the pre-game scheduling survey taken approximately a hour before the start of the game. We also collected participants' party identification and political ideology ratings in the post-game survey. For all our analyses, we used the post-game party identification (Strong Democrat:38.34%, Weak Democrat:16.54%, Lean Democrat: 15.94%, Lean Republican: 11.12%, Weak Republican: 6.01%, Strong Republican: 12.03%). For the moderation analysis, we classify strong Democrats and Republicans as strong partisans and other (Weak/Lean) Democrats and Republicans as weak partisans. For 56 participants, because of a glitch in the post-game survey, we recorded their post-game party identification (Republican/Democrat) but not their party strength (Strong, Weak, Lean Republican/Democrat) was not recorded, we used the party strength that they provided in pre-game scheduling survey instead.

#### 4.5.3.8 Political knowledge

To gauge the political knowledge of the participants, we asked four factual multiple-choice political questions: Do you happen to know who the majority leader in the U.S. Senate is? Do you happen to know which political party has a majority in the U.S. House of Representatives? In the case of a tied vote in the U.S. Senate, who casts the deciding vote? What is the U.S. Electoral College?. We aggregated the number of correct answers that they provided ($M = 2.74, SD = 1.16$). For the moderation analysis, we classified participants who correctly answered at least three of the four questions as high political knowledge participants (402 participants) and the rest as low political knowledge participants (263 participants).

#### 4.5.3.9 Game-related measures

To measure players' game playing experience, we asked "how often do you play party/card/board games?" on a 4-point scale ($M = 3.25, SD = 0.70$). To measure players' perception of the game, we collected their responses on a 10-point scale about how they would rate the game ($M = 7.74, SD = 1.82$), their partner's efforts ($M = 6.04, SD = 2.51$) and their own efforts playing

the game ($M = 5.78, SD = 3.56$). We also asked on a 5-point scale how likely were they to play the game again with another set of questions ($M = 3.73, SD = 1.05$) and how likely they were to recommend the game to their friends ($M = 3.88, SD = 0.98$). Finally, we also asked questions around how much they found the game to be fun ($M = 4.11, SD = 0.78$), confusing ($M = 1.96, SD = 1.12$), informative ($M = 3.85, SD = 0.89$), surprising ($M = 3.83, SD = 0.91$) and difficult ($M = 3.24, SD = 1.24$) on a 5-point scale.

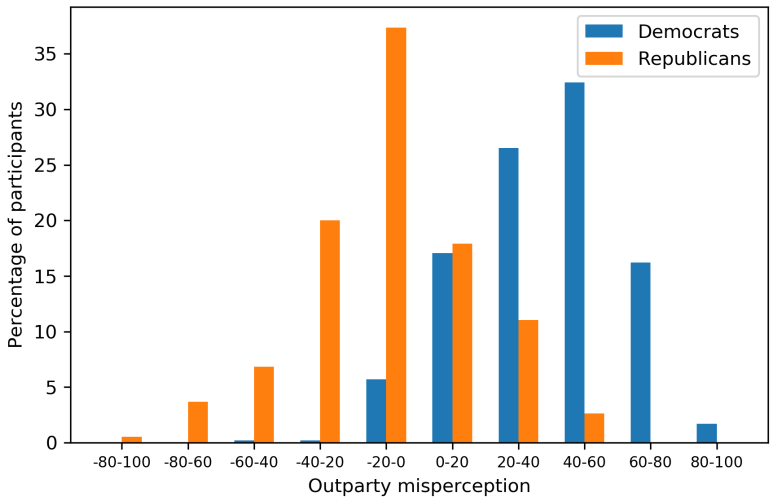
#### 4.5.3.10 Manipulation and attention checks

As a manipulation check, we asked participants on a 5-point scale how political participants thought the game was. As expected, the control version was perceived as the least political ($M = 1.67, SD = 1.18$), followed by the mixed version ($M = 2.40, SD = 1.03$), followed by the fully political version ($M = 4.53, SD = 1.08$) . We asked two instructional manipulation checks [168] to test whether participants paid attention to the questions and followed the written instructions. 98% of participants passed both checks, and no participant failed both checks. No participant completed the survey in less than 45 seconds, the pre-registered threshold to remove them from the analysis.

#### 4.5.3.11 Post-game misperception

After answering the outcome measurement and mediator measurement variables, participants who played the mixed and full game versions were asked two political questions that they were previously asked during the game. To make answers comparable across the two conditions, we asked the fully political game players the questions they answered in the second and fifth rounds. Participants playing the control game were asked two questions at random from the political questions (1 Democrat and 1 Republican-related question) in Tables 4.1 and 4.2. For each experiment condition, we calculated the size of misperception measured as the average difference between the answers provided by the participants to the two questions and the correct answers (survey estimates). If our game was effective in correcting misperceptions and participants could recall them, we would find that the players who played the mixed and fully political conditions, on average, provided answers closer to the correct answer than players who played the control version of the game. We found that mixed ($M = 15.88, SD = 14.86$) and fully political version ($M = 21.87, SD = 16.85$) players indeed on average supplied answers with lower levels of misperception than the control version ($M = 30.30, SD = 14.99$). Interestingly, we found that mixed version players were significantly more accurate than fully political version players ($t = 2.14, p = 0.03$), perhaps because political questions in the mixed version are rare and salient, thereby improving recall.

Figure 4.7: In-game outparty misperceptions




*Note:* For Democrats, bars to the right indicate higher misperceptions of Republicans as being more conservative. For Republicans, bars to the right indicate higher misperceptions of Democrats as being more liberal.

#### 4.5.3.12 In-game outparty misperception

For players playing the treatment game versions, we measure outparty misperceptions as the difference between the answer to questions about outparty supporters' views provided by players during the initial guessing phase and the correct answers to those questions. We determine the direction of misperception, that is, whether the misperception is more liberal or conservative based on how the given answer compares to the survey estimate. For example, from Table 4.1, consider this question: "What percent of Republicans say they would be pleased if the supreme court reduced abortion rights?" The survey estimate is 43%. From survey data, we also obtained the percentage of Democrats who say the same (that they would be pleased if the supreme court reduced abortion rights.) This answer is 5%. Thus, for this question, we would determine that an answer higher than 43% would indicate that the participant perceives Republicans to be more conservative than they actually are. An answer lower than 43% would indicate that the participant perceives Republicans as more liberal than they actually are.

We incorporate the direction of the misperception as follows: if the participant misperceives Republicans' views to be more conservative than their actual views, then we assign a positive sign to the misperception magnitude, else we assign a negative sign to the misperception magnitude. If the participant misperceives Democrats' views to be more liberal than their actual views, then we assign a positive sign to the misperception magnitude, else we assign a nega-

tive sign to the misperception magnitude. Thus, the misperception is positive when participants misperceive Democrats or Republicans to be more extreme in the conventional direction (more liberal or conservative respectively), else the misperception is negative. Figure 4.7 shows the distribution of outparty misperceptions by party (Democrats: $M = 37.64, SD = 23.03$, Republicans: $M = -9.93, SD = 26.36$). For the control version players, we use the outparty question asked in the post-game survey (Section 4.5.3.11) to gauge the size of misperception (Democrats: $M = 36.25, SD = 26.77$, Republicans: $M = -10.88, SD = 26.10$). Therefore, while Democrats (as expected) misperceive Republicans to be more conservative, Republicans misperceive Democrats to be more conservative. Based on Figure 4.7, Republicans appear to harbor fewer misperceptions that Democrats are more liberal. We detail potential reasons for this phenomenon in Section 4.5.7. Note that this measure determines the pre-correction misperception size, whereas the measure in Section 4.5.3.11 measures the post-correction misperception size.

### 4.5.4 Pre-registered analysis plan and deviations

We ran an OLS regression with random effects for teams and experiment batches to estimate the main effects of playing the treatment games on outparty feelings thermometer ratings. We control for relevant socio-political variables such as age, race, gender and party identity and game-related variables such as past gaming experience and ratings of the game, the partner and self. To estimate the main effects on social distance, we performed a similar regression analysis with the same controls with the social distance measure as the dependent variable. To estimate the main effects on willingness to have political conversations with outpartisans and willingness to have nonpolitical conversations with outpartisans, we ran two separate ordinal regression analyses using the same control variables and random effects as above. As per our pre-registration plan, we did not combine the two measures as the Cronbach Alpha was $0.55$. Finally, we ran an ordinal regression analysis to compare the game favorability ratings across conditions using the same control variables and random effects as above. We used the $lme4$ package [16] to run the random effects OLS models and the $ordinal$ package [36] to run the ordinal regressions.

We deviated from our pre-registered plan in a few ways. First, we planned to control for the experiment batch by adding a fixed effect. However, since there were 27 batches and each batch had 10-46 participants, we decided to control for the experiment batch as a random effect. The random effect allows for partial pooling of individual batch effects, reducing overfitting. Second, we planned to control for educational attainment, but because of a coding error, the measure was not collected and not included in the analysis. Similarly, because of a coding error, we did not collect party stereotyping, social distance and willingness to engage with outpartisan measures for the first 56 participants. We removed those participants from any analysis of the aforementioned

measures.

To examine whether perceived commonality, outparty stereotyping and psychological reactance mediate these outcomes, we ran mediation models with controls for demographics and game-related variables using PROCESS package in R [82]. We also examined how the main effects vary by party identification, party strength and political knowledge. To evaluate moderation effects, we used a single random-effects OLS regression modeling outparty feelings with controls for demographics and game-related variables and an interaction term between the treatment condition and each moderator variable. Although we pre-registered to examine treatment effects on participants with low and high outparty misperception, we did not perform the analysis as the size of misperception in our study is heavily correlated with party id as observed in Figure 4.7. We use the $emmeans$ R package [121] to estimate the contrasts between playing the treatment game versions and the control version for the subgroups. We note that we did not pre-register to analyze moderation by political knowledge; however, there is strong evidence to suggest that individuals who have higher political knowledge have more polarized attitudes [102, 214]. Further, unlike other interventions, we expect the game to be played by people who are not necessarily politically engaged and knowledgeable, so we examine how political knowledge might moderate the main effects. However, we note that the experimental setup is powered to detect main effects only and that the mediation and moderation analyses are exploratory.

### 4.5.5 A note about the experiment sample

Before we describing the experiment's results, we provide insights into our participant pool to better interpret the results. As noted above, we recruited participants using MTurk; our sample is not nationally representative. Specifically, our participant sample was overwhelmingly Democratic (71.42%), Caucasian/White (80.75%) and slightly more male (54.74%). Also, our sample has fewer younger (18-24 years:3.61% vs 12%) and older (65+ years: 2.25% vs 22%) participants compared to census estimates. While the MTurk sample population is clearly not representative, prior research suggests that they are more representative than college-based convenience samples [19]. Further, research using MTurk population has successfully replicated results from canonical political science and political psychology experiments [19]. Other research also suggests that the responses obtained from MTurk samples are of high quality and comparable to those obtained from national surveys [42, 43].

However, Krupnikov and Levine (KL), in their study comparing MTurk, YouGov and undergraduate samples, find that MTurk "may not produce generalizable results for all but the simplest experimental designs" [114]. Further, the COVID-19 pandemic has also resulted in an influx of new workers who are more diverse and representative but are less attentive [8]. Thus, we follow

Table 4.5: OLS regression co-efficients modeling outparty feelings

| | *Dependent variable: feelings towards outparty* | | |
|---|---|---|---|
| | Base model | Base model + demographics | Base model + demographics + game experience |
| Mixed game (vs control) | 2.097 | 2.243 | 2.609 |
| | (2.330) | (2.374) | (2.263) |
| Full game (vs control) | 2.145 | 2.089 | 2.169 |
| | (2.351) | (2.379) | (2.269) |
| Republican (vs Democrat) | | 7.672*** | 8.360*** |
| | | (2.100) | (2.001) |
| Age (vs 18-24) | | | |
| 25-34 | | −6.129 | −5.743 |
| | | (5.189) | (4.930) |
| 35-50 | | −8.725* | −7.340 |
| | | (5.156) | (4.902) |
| 51-65 | | −9.920* | −7.642 |
| | | (5.561) | (5.292) |
| 65 | | −0.354 | 7.289 |
| | | (7.982) | (7.674) |
| Woman | | 4.131** | 2.962 |
| | | (1.883) | (1.801) |
| White | | 5.264** | 3.468 |
| | | (2.424) | (2.326) |
| Prior game experience | | | 5.401*** |
| | | | (1.368) |
| Self rating | | | 0.790 |
| | | | (0.554) |
| Partner rating | | | 1.318** |
| | | | (0.516) |
| Game rating | | | 0.983* |
| | | | (0.567) |
| Constant | 38.793*** | 37.980*** | 5.944 |
| | (2.003) | (5.678) | (6.801) |
| Observations | 665 | 665 | 665 |
| Log Likelihood | −3,063.040 | −3,035.406 | −2,996.641 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 4.6: OLS regression co-efficients modeling social distance

| | *Dependent variable: Social distance* | | |
|---|---|---|---|
| | Base model | Base model + demographics | Base model + demographics + game experience |
| Mixed game (vs control) | 0.014 | −0.018 | −0.024 |
| | (0.076) | (0.075) | (0.074) |
| Full game (vs control) | 0.029 | 0.021 | 0.024 |
| | (0.075) | (0.074) | (0.074) |
| Republican (vs Democrat) | | −0.364*** | −0.377*** |
| | | (0.068) | (0.068) |
| Age (vs 18-24) | | | |
| 25-34 | | −0.163 | −0.150 |
| | | (0.168) | (0.167) |
| 35-50 | | −0.203 | −0.203 |
| | | (0.167) | (0.167) |
| 51-65 | | −0.173 | −0.175 |
| | | (0.180) | (0.179) |
| 65 | | −0.303 | −0.380 |
| | | (0.251) | (0.253) |
| Woman | | −0.099 | −0.088 |
| | | (0.061) | (0.061) |
| White | | −0.099 | −0.088 |
| | | (0.061) | (0.061) |
| Prior game experience | | | −0.022 |
| | | | (0.047) |
| Self rating | | | −0.015 |
| | | | (0.019) |
| Partner rating | | | −0.016 |
| | | | (0.018) |
| Game rating | | | −0.020 |
| | | | (0.020) |
| Constant | 1.062*** | 1.498*** | 1.871*** |
| | (0.055) | (0.181) | (0.228) |
| Observations | 609 | 609 | 609 |
| Log Likelihood | −703.496 | −694.048 | −700.553 |

*Note:* *p<0.1; **p<0.05; ***p<0.01

KL's advice in documenting potential reasons our results may not be generalizable [114]. First, the recruitment process is a unique two-step process where participants first accept an invitation to participate and then show up at the said time. This process might skew the participant pool towards more attentive individuals who spend long hours on MTurk. Second, since the game was framed as a party game, the participants accepting the task on MTurk may have differed from those who participated in more traditional political science experiments. For example, the game might not have attracted many from the small subset of the population that Krupnikov and Barry [115] call the "deeply involved", who are heavily vested in politics and are most affectively polarized. This might explain why the average outparty feelings are warmer ($M = 39.72$, Section 4.5.3.1) than other prior studies[233, 126] which typically have outparty feelings in the range of 20-30s. Third, relatedly, we find that Republicans in our participant pool exhibit, on average, 7 degrees warmer feelings towards Democrats than vice-versa (Section 4.5.3.1). This is somewhat unusual as most measures of outparty hostility suggest that either Republicans and Democrats express largely similar levels of hostility or Republicans exhibit even higher hostility towards Democrats than vice-versa [99]. We explore potential reasons for these phenomena in Section 4.5.7.

## 4.5.6 Results

### 4.5.6.1 Main effects of playing the mixed and political versions of GuesSync!

The H1 set of hypotheses posited a positive effect of playing the two versions of the game on our affective polarization outcome measures of outparty feelings and social distance. Table 4.5 shows the coefficients of the outparty feelings OLS regression analyses. We include coefficients from three models: a model with only random effects for experiment batch and game (left-most column), a random-effects model that also controls for demographic variables (center column), and the pre-registered random-effects model that controls for both demographic variables and game experience. Across the three models, we do not find reliable evidence of an increase in outparty warmth when playing either treatment versions versus playing the control version.

Table 4.6 shows the coefficients of the social distance feelings OLS regression analyses. Similar to the above, we include coefficients from three models: a model with only random effects for experiment batch and game (left-most column), a random-effects model that also controls for demographic variables (center column), and the pre-registered random-effects model that controls for both demographic variables and game experience. Again, we find no reliable evidence of a reduction in social distance when playing the mixed or fully political version of the game compared to the control version.

Given these consistent null results, we can conclude that there is no evidence of the main effects of playing the games on affective polarization. However, from the pre-registered model, we find

surprisingly consistent evidence that Republicans in our sample are less affectively polarized than Democrats, exhibiting about 8.36 degrees more warmth towards Democrats than vice-versa and about 0.38 points less socially distant than Democrats.

The H1 set of hypotheses also posited a positive effect of playing the mixed and fully political games on willingness to engage in political and nonpolitical talk. We report the coefficients of the pre-registered ordinal regression analyses with game and experiment batch random effects controlling for demographic and game-related variables in Table 4.7 [13]. We find that participants playing the mixed version of the game exhibited 44% higher odds of willingness to engage in political discussions than players playing the control version ($b = 0.359, OR = 1.43, p = 0.027$ using a one-tailed test as per pre-registration). Playing the fully political game did not result in a statistically significant increase in willingness to talk politics with outparty supporters, but its effects were directionally similar to that of the mixed game version ($b = 0.301, OR = 1.35, p = 0.051$ using a one-tailed test as per pre-registration). However, neither playing the mixed or the fully political version resulted in a reliable increase in willingness to have nonpolitical conversations with outparty, at least partly due to ceiling effects ($M = 4.28$ out of 5, Section 4.5.3.3), meaning people are already quite open to engaging in nonpolitical topics with outparty supporters. Thus, our analyses indicate partial support for our hypotheses about engaging with outparty supporters.

Answering our research question on game ratings using an ordinal regression with the demographic controls [14], from Table 4.7, we find no significant difference between ratings given to mixed and political game versions compared to control[15]. Given the reasonable number of observations, the lack of evidence of a main effect is unlikely due to small sample sizes. It appears that adding political questions to the game does not significantly change how people rate the game.

#### 4.5.6.2 Mediator analyses

Next, we examine the indirect effect of playing the treatment game versions on the outcome measures through perceived commonality, outparty stereotyping and psychological reactance (H2 set of hypotheses). We find that neither mixed nor full versions significantly affected perceived commonality and outparty stereotyping. But along expected lines, higher perceived commonality ($b = 8.61, p < 0.01$) and lower outparty stereotyping ($b = -3.64, p < 0.01$) increased outparty warmth. Similarly, higher perceived commonality ($b = -0.26, p < 0.01$) and lower outparty stereotyping ($b = 0.20, p < 0.01$) reduced outparty social distance. Higher perceived commonality increased willingness to discuss political ($b = 0.39, p < 0.01$) and nonpolitical

[13] We also report the coefficients of the same analyses with no control variables and with only demographic control variables in the Appendix Tables 4.13 and 4.12. The results were largely similar to the pre-registered analysis.

[14] We did not include game-related variables as control they correlate with the outcome variable.

[15] We also report the coefficients of the same analyses with no control variables and with only demographic control variables in the Appendix Table 4.14. The results were largely similar to the pre-registered analysis

Table 4.7: Ordinal regression co-efficients modeling behavioral intent and game ratings. Full model co-efficients in Appendix Tables 4.12, 4.13 and 4.14.

| | *Dependent variable:* | | |
|---|---|---|---|
| | Willingness political talk | Willingness nonpolitical talk | Game rating |
| Mixed (vs control) | 0.359** | 0.208 | 0.082 |
| | (0.185) | (0.200) | (0.171) |
| Fully political (vs control) | 0.301* | 0.078 | 0.214 |
| | (0.184) | (0.196) | (0.173) |
| Observations | 609 | 609 | 665 |
| Log Likelihood | −886.09 | −675.07 | −1236.20 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | |

($b = 0.22, p < 0.01$) topics with outparty while the effect of outparty stereotyping on willingness to discuss with outparty members was not significant.

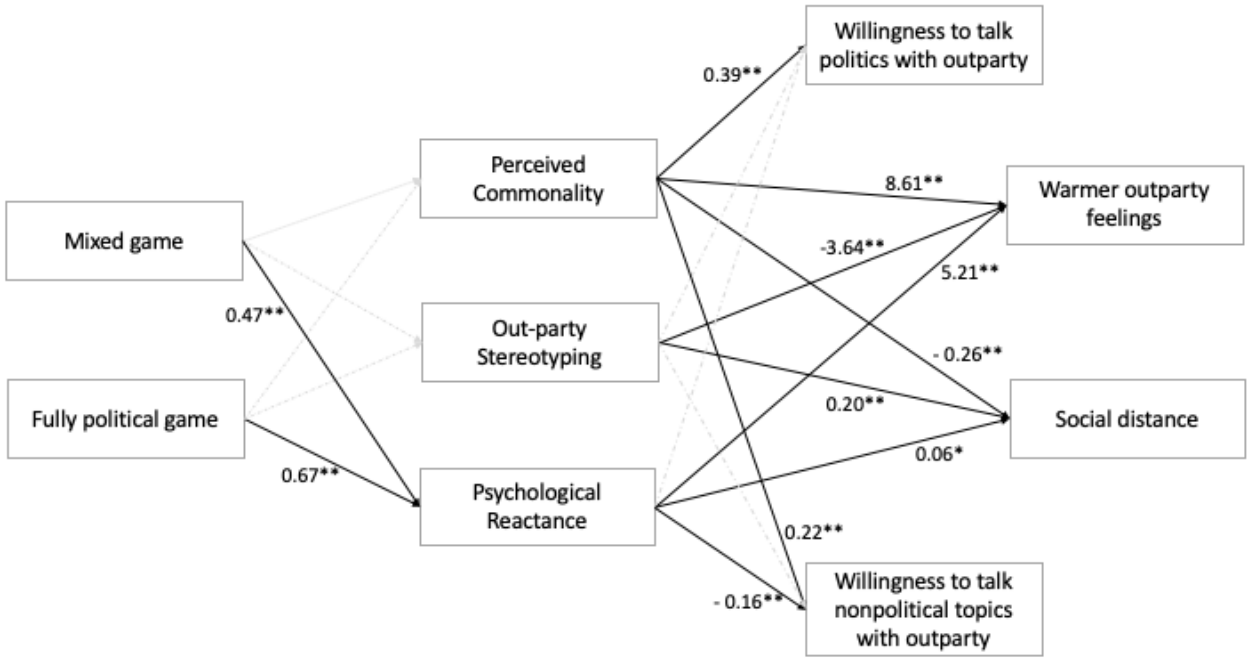
Interestingly, we find that both treatment game versions significantly increase psychological reactance (mixed game $b = 0.47, p < 0.01$, fully political game $b = 0.67, p < 0.01$). However, its effect on the outcome variables were inconsistent. As expected, psychological reactance increased outparty social distance ($b = 0.06, p < 0.05$), decreased willingness to discuss nonpolitical topics with outparty members ($b = -0.16, p < 0.01$), but it also increased outparty warmth ($b = 5.21, p < 0.01$). We discuss potential reasons for this unexpected phenomenon in the discussion section.

#### 4.5.6.3 Moderator analyses

Finally, we analyze how party identification, party strength, and political knowledge moderate outparty feelings. We report on results from the pre-registered OLS regression model controlling for demographic and game-related variables in Figure 4.9[16]. Figure 4.9 plots each moderator's mean treatment effect and confidence intervals of the treatment game versions. Analyzing moderation by party identification, we find that Democrats playing the mixed and fully political versions exhibited outparty feelings that were, on average, respectively 6.58 degrees ($p = 0.01$) and 5.26 degrees ($p = 0.04$) warmer than Democrats playing the control version. Republicans playing either treatment version did not reliably exhibit changes to outparty feelings compared to the control version.

[16]The results without controls variables and with only demographic controls are available in the Appendix figures 4.10 and 4.11 respectively. The results were largely similar to the pre-registered analysis

Figure 4.8: Mediation analyses




*Note:* The mixed game denotes the difference between the mixed game and control condition. The fully political game denotes the difference between the fully political game and control condition. All numbers are regression co-efficients. Solid lines represent denote statistically significant relationships ($*$ indicates $p < 0.05$, $**$ indicates $p < 0.01$), gray dotted lines denote non-significant relationships.

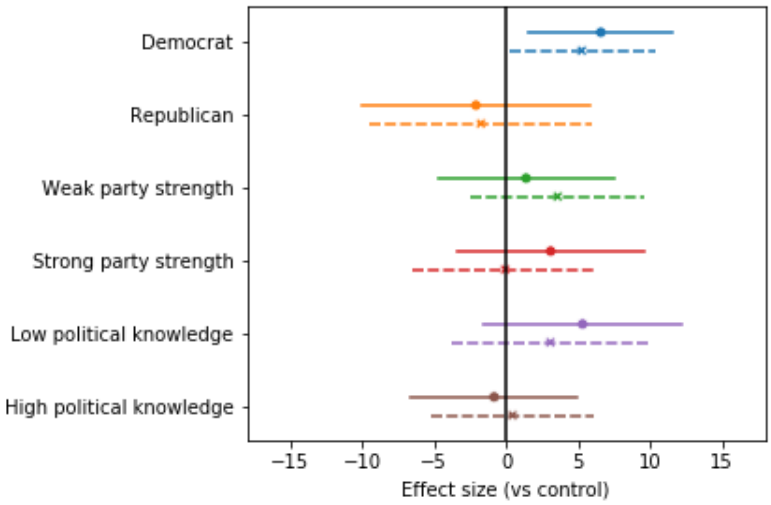
Comparing the effects of playing the treatment games on strong and weak partisans, none of the differences were statistically significant. Comparing the effects of playing the treatment games on low and high political knowledge players, none of the differences were statistically significant.

### 4.5.7 Exploratory analyses

The above analyses raise important questions around Republican participants expressing significantly warmer outparty feelings than Democrats and the game having differential treatment effects on Republicans and Democrats. To better understand these phenomena, we conduct exploratory analyses focusing on the sample of study participants, misperceptions they held, and how the game questions posed may affect our reading of participants' misperception.

First, we compare the study sample demographics with data from the nationally-representative ANES survey. We find that 41.57% of sample Republicans indicate that they "Lean Republican"

Figure 4.9: Moderator analyses




*Note:* The solid line represents the 95% confidence interval of the effect size of the mixed game compared to control, and the dotted line represents the 95% confidence intervals of effect size of the fully political game compared to control. The dots and crosses represent the mean effect size estimates.

compared to 24.95% of Republicans nationally estimated from the ANES survey. In comparison, 23.57% of sample Democrats indicate that they "Lean Democrat" compared to an estimated 24.76% of Democrats from the ANES survey. Thus, Republicans in the study sample are more moderate than the typical Republican in the broader electorate. Also, in the study sample, the average Republican is more moderate than the average Democrat. As weak partisans typically exhibit less outparty hostility, this could be one reason why Republicans in our sample, on average, exhibit warmer feelings towards Democrats than vice-versa. However, Republicans being more moderate still does not fully explain why the game has differential effects on Republicans and Democrats. To understand this phenomenon, we examine the misperceptions participants held during the game.

Following the approach in Section 4.5.3.12 we analyze participants' own party and outparty misperceptions based on the initial guesses provided by those playing the mixed and fully political game versions. We summarize how Republican and Democratic participants answered questions on party supporters' views in Table 4.9. We find that Democratic participants overestimated how conservative Republicans were in about 92% of their answers about Republicans. In contrast, they underestimated how liberal Democrats were in about 57% of their answers about Democrats. On the other hand, Republican participants overestimated how conservative Republicans were in about 92% of their answers about Republicans. At the same time, they underestimated how liberal Democrats were in about 64% of their answers about Democrats. Thus, in ample cases (92%),

Table 4.8: Relationship between outparty misperception and outparty hostility

| | *Dependent variable:* |
|---|---|
| | Feelings towards outparty |
| Full game (vs control) | 2.277 |
| | (2.326) |
| Mixed game (vs control) | 2.764 |
| | (2.310) |
| Avg. outparty misperception | $-0.154^{***}$ |
| | (0.029) |
| Constant | $42.252^{***}$ |
| | (2.078) |
| Observations | 665 |
| Log Likelihood | $-3{,}052.072$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

the game could correct Democrats' misperception that Republicans were extreme conservatives, whereas only in a minority of cases (36%), could the game correct the Republicans' misperception that Democrats were extreme liberals. This difference between game experiences of Republicans' and Democrats' could have contributed to the differential effects. However, for this to be a possibility, assuming the game was effective in lowering misperceptions, the lower levels of outparty misperception should be related to higher outparty warmth.

To examine how outparty misperception correlates with outparty hostility, we run a regression modeling the feelings thermometer ratings using random effects for teams and experiment batches, a fixed effect for game type (base model), and additionally, the average misperception of outparty supporters' views as an independent variable. The coefficients of the models are shown in Table 4.8. As expected, we find that higher levels of outparty misperception is inversely correlated with outparty warmth. We find that, on average, for every percentage that participants' misperceive outpartisans to be more extreme, the feelings thermometer ratings reduce by 0.15 degrees. Since Democrats exhibit more outparty misperceptions than Republicans in the study, correcting for them, likely results in more reduction in hostility for Democrats than Republicans who exhibit fewer outparty misperceptions. To more stringently test this reasoning, ideally, we would like to compare the games' effect on Democratic and Republican participants exhibiting varying levels

Table 4.9: Misperceptions gauged based on guesses in initial guess phase

| Player Party ID | Misperception about Democrats | | Misperception about Republicans | |
|---|---|---|---|---|
| | More Liberal | More Conservative | More Liberal | More Conservative |
| Democrat | 40.88% | 57.36% | 7.02% | 92.43% |
| Republican | 35.69% | 63.71% | 7.93% | 92.07% |

of outparty hostility, however, because of high collinearity between outparty misperception levels and party id, we are unable to perform that analysis.

Now, we turn to the question of why Republicans exhibit fewer misperceptions about the extent to which Democrats are liberal. To answer this, we reviewed the game questions in detail. While Republicans being more moderate in our sample than is typical might have resulted in more tempered perceptions of Democrats, we believe that a significant cause was the game questions we had selected. Take for example, the question "what percentage of Democrats say that eligible voters are never denied the right to vote?". The survey estimate for this question was 7%. Any guess above 7% would imply that the player thought that Democrats were more conservative than they actually are, while only a guess below 7% would imply that the player thought Democrats were more liberal than they actually are. Consider another question "what percentage of Democrats say that transgender people face no discrimination at all in the US?". The survey estimate was 1%. For this question, because of the extremity of the survey estimate, only a guess of 0% would imply that the player thought Democrats were more liberal than they actually are. We found 4 such questions out of 13 Democrat-related questions. These questions likely skewed our misperception estimates of Democrats' views towards being more conservative. In contrast, the Republican-related questions with extreme survey estimates likely resulted in exaggerated misperception estimates of Republicans being more conservative. Take for example, the question "What percentage of Republicans say that the police officers never use more force than necessary?". The survey estimate was 3%. Here, any guess from 3%-100% would imply that the player thought Republicans were more conservative than they actually where. There were 3 such questions. Given that individuals were less likely to guess extremely low or high numbers, these questions likely exaggerated our misperception estimates of Democrats and Republicans being more conservative. Another issue that compounded this problem was that in the question selection process, we only selected questions that participants had the most misperception on, however, we did not consider the direction of the misperception. In hindsight, for this particular intervention, we ought to have selected questions for which the survey estimates are not extreme values and questions for which Democrats' views are misperceived to be more liberal and Republican views' are misperceived to be more conservative.

Table 4.10: Game experience measures by game type

| Game Type | Star ratings | Play again | Recommend to friends | Fun | Informative | Surprising |
|---|---|---|---|---|---|---|
| Control | 7.68 (1.79) | 3.62 (1.03) | 3.85 (0.96) | 4.09 (0.82) | 3.71 (0.93) | 3.75 (0.91) |
| Mixed | 7.71 (1.8) | 3.74 (1.05) | 3.86 (1.01) | 4.14 (0.77) | 3.92 (0.84) | 3.92 (0.86) |
| Fully Political | 7.82 (1.88) | 3.83 (1.07) | 3.94 (0.98) | 4.12 (0.77) | 3.92 (0.88) | 3.81 (0.96) |

Showing the mean and standard deviation (in parenthesis).

Table 4.11: Game chat messages

| Game Type | Number of messages |
|---|---|
| Control | 6.72 (5.06) |
| Mixed | 5.97 (4.46) |
| Fully Political | 6.43 (4.25) |

Showing the mean and standard deviation (in parenthesis).

## 4.6 Additional exploratory analyses

We also include additional analyses that we performed to understand how players engaged with the game. Table 4.10 shows by game version the mean and standard deviations of key game perception metrics we collected. We find that on all measured metrics: 10-point game ratings, likely to play again, likely to recommend to friends, how fun, informative and surprising the game was, the ratings for the treatment versions were comparable to the nonpolitical control version of the game. Similarly, comparing the number of chat messages during the game, we find that the three game versions have a similar mean number of game messages sent (Table 4.11). Overall, only 7.22% of players did not send a message to their partner.

## 4.7 Discussion

### 4.7.1 Engaging in politics through games

GuesSync! is likely one of the first games aimed at reducing affective polarization. Though we did not find a main effect on outparty feelings (a 5 degree increase in the 101-point feelings thermometer scale), the moderation analyses suggest that the games might be particularly effective among Democrats. Playing the mixed version of the game increased the willingness to engage in political

discussions with the outparty. These findings take on greater importance as prior research suggests that partisans exhibit a strong reluctance to engage with the other side [204]. Further, these results suggest that this game could be used as a potential ice-breaker activity in local community meetings, participatory planning meetings and citizen forums before participants engage with opposing partisans on substantive issues.

Importantly, based on game experience measures (Table 4.10), participants appear to enjoy playing the treatment game versions at least as much as the nonpolitical control version. This suggests that corrective political information can be incorporated within game settings without negatively impacting the game's fun quotient. Such games can be scaled up relatively easily to a broader audience by embedding them in social media platforms such as Facebook. Further, these games could complement (or be a precursor) to other interventions that require a deeper engagement with outparty individuals, such as having one-on-one [13] or group discussions [130]. As more people show little appetite for politics [115], these games could provide a small dose of politically relevant information packaged in a casual, fun way.

### 4.7.2 Role of psychological reactance in attempts to reduce outparty hostility

Through the mediation analyses (Section 4.5.6.2), we find that playing either treatment version resulted in higher ratings on the psychological reactance scale. Note that the psychological reactance scale measures feelings of being pressured/manipulated/forced to form certain views about Republicans and Democrats. Thus, feelings that could potentially result in psychological reactance were created by playing the game. However, we found its effects on the outcome measures were mixed, reducing the willingness to talk politics with outpartisans and increasing social distance, but increasing outparty warmth. It is unclear why there are opposite effects for the different outcomes. One potential reason could be that the feelings thermometer ratings measure a somewhat abstract concept of feelings towards outparty, whereas social distance and willingness to talk politics measure attitudes towards specific scenarios and behaviors. Thus, the feelings of being pressured/manipulated/forced do not translate into psychological reactance when asked about abstract attitudes, but they likely do when asked about engaging with an outpartisan which is perhaps a bridge too far.

Psychological reactance has not been previously tested as a potential mechanism in the context of affective polarization. However, it is a plausible explanation of why efforts to reduce outparty hostility have yielded largely modest effects. Relatedly, in another study [127], Levendusky tested if inducing partisan-ambivalence by asking people what they dislike about their own party and like about the other party could reduce affective polarization. Many participants resisted the

task with responses such as 'nothing' and 'are you kidding me?'. While psychological reactance was not formally measured, the responses suggest that it could have been induced as this was a somewhat direct manipulation. The fact that even the mixed version of the game containing little political information triggered a measurable increase in psychological reactance suggests that other approaches could also trigger the same. More research is needed to better understand when psychological reactance is triggered and ways to mitigate it.

### 4.7.3 Heterogeneous effects of correcting misperceptions about party supporters' political views

Moderation analysis (Figure 4.9) suggests that Democrats playing the treatment games generally exhibited more warmth towards Republican supporters, whereas Republicans playing the treatment games did not reliably exhibit a change in their feelings towards Democrats compared to those playing the control version. Exploratory analyses in Section 4.5.7 suggests that a major reason for this might be that many of our Democrat-related game questions did not result in correcting misperceptions of Democrats being extremely liberal. Instead, the survey estimates for those questions only reaffirmed that Democrats were extremely liberal in their views. In hindsight, we ought to have selected questions for which the survey estimates were not extreme values.

Note that in Section 4.4.3.1, we selected political game questions based on the size of misperception and importance rating. However, we did not factor in the direction of the misperception. For this game, we ought to have considered the direction of misperception and selected questions for which Republicans overestimate how liberal Democrats' views are. However, this surfaces an important conundrum. Do we correct misperceptions about the outparty only on certain views where we know that outparty extremity is exaggerated? One concern is that by focusing on only issues that are misperceived to be contentious, we might reduce outparty hostility. However, we run the risk of players perceiving more common ground than there is, which may dampen political mobilization efforts (I discuss more about potential negative consequences in Section 5.2.1). Moreover, long term, if the game is perceived only correct certain kinds of misperception, players might consider it to be overly manipulative and not return to play again or the game's effectiveness in reducing outparty hostility might be reduced.

## 4.8 Limitations

We acknowledge that our study has some limitations in the game's design and the experiment. As discussed earlier, the game questions suppressed misperception of Democrats' being more liberal which likely reduced the effectiveness of the game intervention on Republicans. Also, the game

always corrects perceptions about both Republican and Democratic supporters' political views, which does not allow us to distinguish between effects due to corrections about inparty and outparty political views. Also, we do not measure how confident players are about their perceptions of party supporters' views during the game. In our question selection process, we selected only the questions on topics that partisans claimed were most important to them, so participants were likely, on average, more misinformed than uninformed about these topics. Nevertheless, we cannot distinguish between the uninformed and the misinformed in this game design. Finally, it is possible that the original misperceptions that participants had about the opposing party supporters continue to shape their attitudes about them even after the in-game correction. This phenomenon, called belief echos, is observed in misperception corrections of factual news [219]. More research is needed to see if such belief echoes also exist when correcting others' perceptions.

Our experiment participants were recruited from the Amazon MTurk platform and appear to be disproportionately young, male and white compared to census statistics. Thus, it is unclear how the results might differ when the larger public plays the game. Also, the post-game survey was administered immediately after the game, so we do not know how long treatment effects might last. However, participants signaled that if given the opportunity, they would play the game again (Table 4.10) with different questions. It is likely that if the game were repeatedly played, these effects might hold long-term. Another related limitation would be the number of unique political questions available in the game if the game were to be played long-term. Naturally, we are limited by the topics that partisans misperception of party supporters' views. Unfortunately (or fortunately for the game), because of factors such as partisan media, selective media exposure and motivated reasoning, these misperceptions are likely to remain if not grow in the foreseeable future. Thus, we do not expect to run out of questions to ask in the game. In this work, because of resource constraints, we used existing survey data to formulate questions which limited the questions we could ask. If we had more resources, we could run our own nationally representative surveys to create game material for the game.

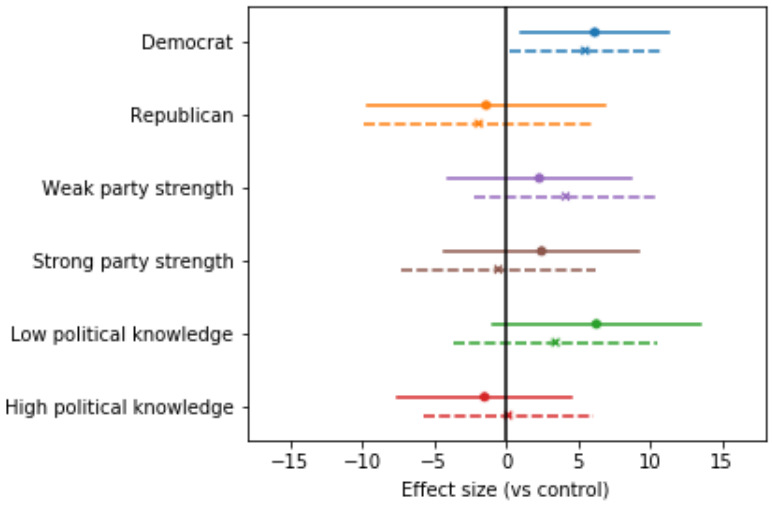
## 4.9 Conclusion

In this work, we present a fun and engaging party game GuesSync!, which we designed to help reduce affective polarization and increase engagement with outparty supporters. From experimenting with three game versions, we did not find evidence that GuesSync reliably reduces affective polarization based on standard measures of outparty feelings and social distance. However, the treatment versions of the game were effective in improving outparty feelings among Democrats. The mixed version was also effective in improving willingness to talk politics with outpartisans. We also identified psychological reactance as a potential mechanism that might affect the effec-

tiveness of depolarization interventions. Finally, our game experience measures show that the two political games were just as fun to play as the nonpolitical game version suggesting that, contrary to popular belief, people do, in fact, like to mix politics and play.

## 4.10 Appendix

Figure 4.10: Moderator analyses with random-effects OLS modeling outparty feelings without control variables




*Note:* The solid line represents the 95% confidence interval of the effect size of the mixed game compared to control, and the dotted line represents the 95% confidence intervals of effect size of the fully political game compared to control. The dots and crosses represent the mean effect size estimates.

Table 4.12: Ordinal regression co-efficients modeling willingness to talk politics with outparty

| | *Dependent variable: Willingness to talk politics* | | |
|---|---|---|---|
| | Base model | Base model + demographics | Base model + demographics + game experience |
| Mixed game (vs control) | 0.325** | 0.363** | 0.358** |
| | (0.183) | (0.183) | (0.185) |
| Full game (vs control) | 0.283* | 0.258 | 0.301* |
| | (0.183) | (0.182) | (0.184) |
| Republican (vs Democrat) | | 0.636*** | 0.765*** |
| | | (0.169) | (0.171) |
| Age (vs 18-24) | | | |
| 25-34 | | 0.536 | 0.470 |
| | | (0.396) | (0.403) |
| 35-50 | | 0.258 | 0.294 |
| | | (0.396) | (0.402) |
| 51-65 | | 0.377 | 0.462 |
| | | (0.424) | (0.431) |
| 65 | | −0.222 | 0.034 |
| | | (0.613) | (0.633) |
| Woman | | −0.208 | 0.148 |
| | | (0.149) | (0.150) |
| White | | 0.142 | 0.014 |
| | | (0.199) | (0.201) |
| Self rating | | | 0.055 |
| | | | (0.045) |
| Partner rating | | | 0.012 |
| | | | (0.044) |
| Game rating | | | 0.175*** |
| | | | (0.049) |
| Prior game experience | | | 0.385*** |
| | | | (0.115) |
| Observations | 609 | 609 | 609 |
| Log Likelihood | −933.02 | −922.03 | −895.23 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 4.13: Ordinal regression co-efficients modeling willingness to talk nonpolitical topics with outparty

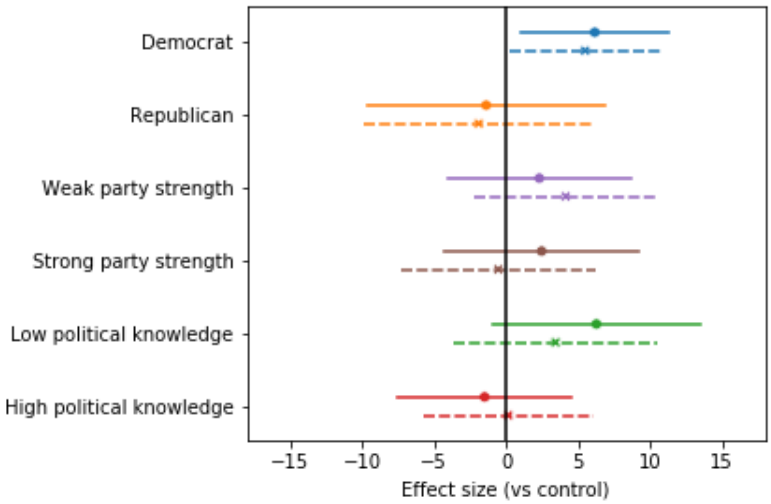
| | *Dependent variable: Willingness to talk nonpolitical topics* | | |
|---|---|---|---|
| | Base model | Base model + demographics | Base model + demographics + game experience |
| Mixed game (vs control) | 0.137 | 0.228 | 0.208 |
| | (0.197) | (0.201) | (0.199) |
| Full game (vs control) | 0.077 | 0.107 | 0.078 |
| | (0.195) | (0.198) | (0.195) |
| Republican (vs Democrat) | | 0.732*** | 0.742*** |
| | | (0.191) | (0.191) |
| Age (vs 18-24) | | | |
| 25-34 | | 0.856** | 0.777* |
| | | (0.421) | (0.417) |
| 35-50 | | 1.121*** | 1.038** |
| | | (0.419) | (0.416) |
| 51-65 | | 1.259*** | 1.155** |
| | | (0.457) | (0.455) |
| 65 | | 1.441** | 1.331* |
| | | (0.699) | (0.702) |
| Woman | | 0.091 | 0.080 |
| | | (0.163) | (0.163) |
| White | | 0.019 | 0.074 |
| | | (0.209) | (0.209) |
| Self rating | | | −0.001 |
| | | | (0.049) |
| Partner rating | | | −0.010 |
| | | | (0.046) |
| Game rating | | | 0.134** |
| | | | (0.052) |
| Prior game experience | | | −0.189 |
| | | | (0.126) |
| Observations | 609 | 609 | 609 |
| Log Likelihood | −696.85 | −682.19 | −677.86 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 4.14: OLS regression co-efficients modeling game ratings

| | *Dependent variable: Game ratings* | |
|---|---|---|
| | Base model | Base model + demographics |
| Mixed game (vs control) | 0.058 | 0.083 |
| | (0.167) | (0.171) |
| Full game (vs control) | 0.205 | 0.214 |
| | (0.170) | (0.173) |
| Republican (vs Democrat) | | −0.123 |
| | | (0.158) |
| Age (vs 18-24) | | |
| 25-34 | | −0.269 |
| | | (0.375) |
| 35-50 | | −0.082 |
| | | (0.373) |
| 51-65 | | −0.144 |
| | | (0.406) |
| 65 | | −0.231 |
| | | (0.590) |
| Woman | | −0.275* |
| | | (0.142) |
| White | | −0.011 |
| | | (0.181) |
| Observations | 665 | 665 |
| Log Likelihood | −1239.60 | −1236.20 |

*Note:*
*p<0.1; **p<0.05; ***p<0.01

Figure 4.11: Moderator analyses with random-effects OLS modeling outparty feelings controlling for demographics




*Note:* The solid line represents the 95% confidence interval of the effect size of the mixed game compared to control, and the dotted line represents the 95% confidence intervals of effect size of the fully political game compared to control. The dots and crosses represent the mean effect size estimates.

# CHAPTER 5

# Conclusion

The past decade in the US has been one of the most politically polarizing in recent memory. With rising affective polarization fueling distrust of fellow partisans [94], political elites [56], democratic institutions such as the Supreme Court [9], and undermining support for democratic norms [107][1], many scholars are concerned about the pernicious effects of polarization on the US's democracy. What, if anything, can Internet scholars do to stem the tide? Truth be told, not as much as one would wish. As discussed in the introduction, many structural factors contribute to the extreme levels of outparty hostility we observe today, and addressing them requires large-scale structural changes and political will. Moreover, it is not certain if the Internet, particularly social media, causes, exacerbates, or even reduces outparty hostility [23]. However, there are some levers that we can pull to make a meaningful difference. In this chapter, I start with a summary of the three studies that reduce hostility in outparty attitudes and interactions, and then I work through some thorny questions about this dissertation, and finally, I describe future directions that I look to explore.

## 5.1 A recap

In this dissertation, I take a mixed-methods approach leveraging nonpolitical spaces and identities to design online interventions that reduce hostility in outparty attitudes and interactions. In the first study, I perform an extensive data analysis of billions of Reddit comments to develop insights into the kinds of political discussions that online nonpolitical spaces foster. Finding that discussions in these spaces are less hostile, in the second study, I identify ways to leverage shared membership in nonpolitical spaces to humanize outpartisans and improve political discussions. Finding that people look for informative *yet fun* political interactions online, in the final study, I experiment with an online game that incorporates a hostility-reducing intervention through gameplay. Significantly,

[1] Although see [28] for a working paper that shows otherwise.

with this game, I move forward an alternate kind of political interaction that is fun and engaging and can complement more serious political deliberations. I discuss in detail the three studies below.

In the first study, through a large-scale computational analysis on Reddit, I explore the potential of nonpolitical communities in hosting political discussions and how those discussions may differ from political discussions in explicitly political communities. I find that nearly half of all political talk takes place in nonpolitical communities highlighting how nonpolitical settings are often conducive to political interactions. Importantly, cross-partisan political interactions are less toxic in nonpolitical communities than in explicitly political communities. I speculate that political discussions in nonpolitical communities are less hostile as common interests and relaxed conversational norms in these communities temper hostile partisan norms that usually affect discussions in political communities.

Building on these results of the first study, I investigate how to incorporate information about users' shared nonpolitical group memberships within political discussions in the second study. I demonstrate that users are comfortable knowing (and revealing) shared memberships in nonpolitical communities with outpartisan discussion partners and expect that information to humanize outpartisans and reduce hostile interactions. In the study, I also developed insights about how users look for both serious, informative content as well as fun, casual entertainment in political interactions.

In the final study, I design an online two-player party game that incorporates correcting misperceptions about views held by ordinary Republicans and Democrats in a casual game setting. I find that playing the game reduces hostile outparty attitudes among Democrats. Notably, the game versions containing political content were just as well rated as a version of the game that did not contain political content, suggesting that people were open to and enjoyed mixing politics and play. These results indicate that carefully designed games could significantly scale up depolarizing interventions to broad segments of the population. Together, these studies stack up to detail ways to reduce hostile outparty attitudes and interactions online.

## 5.2 Thorny questions about this dissertation

### 5.2.1 Are there unintended negative consequences of reducing hostility between ordinary Republicans and Democrats?

A significant concern for me in building out solutions to reduce outparty hostility is ensuring that the deep differences in policy preferences and social values that the two groups hold are not minimized or brushed aside. This concern stems from relatively recent research that suggests that prejudice reduction strategies have an unintentional negative effect of reducing recognition

of structural inequality and undermining support for collective action and social justice among members of disadvantaged groups [81, 193]. One pathway for this phenomenon is that strategies for reducing hostility emphasize common identities and actively aim to reduce the salience of ingroup identities. In contrast, attachment to the ingroup drives collective action among disadvantaged groups [91]. While neither Republicans nor Democrats are disadvantaged groups *per se*, since most racial, religious and sexual minorities are now aligned with the Democratic party, depolarization efforts may potentially affect these subgroups in organizing against laws enacted by Republican state legislatures that infringe on their civil rights (such as racist voting rights laws)[2]. Similarly, research suggests that even existing depolarization techniques such as priming American identity have the unintended effect of promoting affective polarization toward undocumented immigrants [230]. Thus, more research is needed to identify potential negative consequences of depolarization efforts and mitigate them.

### 5.2.2 Does mixing politics and games trivialize how we engage with politics?

Another concern is that by mixing politics and games, I both trivialize politics and sap the fun out of games. Thankfully, the results from the game experiment suggest that games with political content are still fun. Yet, do we really want to encourage this form of light engagement with politics through games? In some ways, the game could be precisely the kind of political engagement that might actually be beneficial to most people. Krupnikov and Barry in their book [115] identify the "other divide" in the US based on political involvement between a small minority of citizens who are "deeply involved" [3] and all others (who are simply in the know about politics or do not follow politics entirely). Most people encounter political interactions casually, at workplaces, social gatherings and social media. These encounters are likely with the deeply involved partisans who are the most vocal. While these encounters provide a conduit for (biased but nonetheless) political information, they also elicit negative internal comparisons with the deeply involved, resulting in even disengagement from politics altogether. Instead, these games could have the opposite effect of building curiosity and creating a positive association with politics which may have positive downstream consequences on political participation. Indeed, Lerner, in his book [122] on making democracy fun, makes a convincing case for how games and game-like processes, when designed right, can increase involvement in the democratic process by making public hearings and community meetings more fun and engaging.

However, as Lerner cautions, games could "trivialize serious political issues or manipulate cit-

[2] https://www.justice.gov/opa/pr/justice-department-files-lawsuit-against-state-georgia-stop-racially-discriminatory

[3] The deeply involved are people who (i) spend much time on politics at the cost of other activities, (ii) perceive even mundane political events as significantly important and (iii) are extremely vocal about their political thoughts and opinions. These people also harbor high levels of animosity towards outpartisans.

izen participation. Or they might lead to unfair outcomes, or simply not be any fun." Games risk overly simplifying intricate issues to facilitate gameplay which may desensitize and trivialize serious political issues [194]. Although players typically have some agency over their actions in a game, game designers wield considerable power to set the agenda and influence preferences. While this power could be used to design pro-social games, it can also be harnessed to promote suspect values such as nationalistic jingoism [209]. Thus, as the game experiment demonstrated, mixing politics and games might be a promising endeavor, but future work in this space must be attuned to the values built into such games and the impact they create on people's perceptions of politics.

### 5.2.3 Reducing hostility on both sides in the face of right-wing radicalization. Are we barking at the wrong tree?

The GuesSync! game aims to reduce hostile outparty attitudes held by *both* Democrats and Republicans. However, scholars note that the hostility towards each other is not symmetrical and that the right is being radicalized at a significantly higher level [18]. As Kalmoe and Mason (KM) note in their book on partisan violence [99], we cannot pretend that "both parties are equally culpable, that their actions are morally equivalent, or that they pose equal dangers to the democratic project." After the January 6th, 2020 attacks on the Capitol, they found that one in five Republicans endorsed political violence "today", while one in eight Democrats did. For Republicans, radicalism is strongly predicted by hostility towards Black Americans and women, whereas the opposite is true for Democrats, that is, radicalism is predicted by lower levels of racial resentment and hostile sexism (though with smaller effect sizes). How, then, do our efforts in this dissertation to reduce outparty hostility *on both sides* square with alarming levels of right-wing radicalization?

KM's work on partisan violence implies that we must focus our efforts to reduce hostility primarily on Republicans. Yet, affective polarization as measured by feelings thermometer ratings are largely symmetrical for Republicans and Democrats. In nationally representative surveys after the Capitol attack, there were no significant difference between the feelings thermometer ratings of Democrats and Republicans towards each other [195]. Surprisingly, KM found little correlation between measures on their partisan violence index and feelings thermometer ratings ($r = -0.02$). This suggests the two measures quantify different concepts. Although, feelings thermometer ratings are typically interpreted to measure hostility, participants are only asked to rate how warm/favorable or cold/unfavorable they feel about the party supporters. The partisan violence index asks when it is okay to use violence to meet political goals, send threatening and intimidating messages to party leaders, and harass party supporters in a way that makes them feel unsafe or frightened. Thus, although seemingly adjacent, these two scales measure different man-

ifestations of hostility. Specifically, Lelkes and Westwood [120] found that affective polarization, measured using the feelings thermometer predicts suppressing hostile rhetoric towards own party, avoidance of outparty members and own party preferential treatment but not intentionally causing harm to the opposition. Thus, I believe that there is value in focusing on both de-polarization and de-radicalization, and that these two initiatives complement each other. I expect that the game's relatively subtle interventions will have little effect on partisans' propensity for partisan violence, yet it can have meaningful effects on how ordinary partisans feel about the other side which have important downstream consequences [94]. Whereas, interventions to de-radicalize may have effects on improving outparty feelings as well. Moving forward, a useful line of research might be to draw out the connections between how the Internet provides a pathway to radicalization on the right (such as [138]) and how online interactions might polarize ordinary partisans.

## 5.3 Charting the path forward

Building on this dissertation, I look to work on the following topics.

### 5.3.1 (Re-)Designing online political interactions

As briefly described in the introduction chapter, online interaction interfaces such as the comments sections and discussion forums that facilitate political discussions have not significantly evolved with the changing political landscape. No wonder then are many news sites shutting down their comments section as they become too hostile[4]. Platforms have responded by employing both human moderators [211] (though not nearly enough) and sophisticated machine learning models to detect and remove such content [236]. While these approaches certainly reduce hostility, there are also other design changes that we can better facilitate political interactions. Recent social science research has identified many viable strategies to reduce outparty hostility. Yet, little work has been done to translate those findings from survey and experiment environments to real-world designs. For example, in Chapter 3, I explore how we can use *cross-categorization* and *de-categorization* approaches by highlighting nonpolitical identities to offset partisan identities in online political discussions. In the future, I hope to incorporate such evidence-based interventions into online interaction designs. I believe that these enhancements can meaningfully impact not only individual interactions but also shape outparty attitudes more broadly.

[4]https://www.kqed.org/lowdown/29720/no-comment-why-a-growing-number-of-news-sites-are-dumping-their-comment-sections

### 5.3.2 Designing fun and engaging depolarizing interventions

Results from experiments in Chapter 4 suggest that people enjoy playing fun and engaging games that may be political. This presents more opportunities to mix politics and play. A significant advantage of designing these games over other online interactions, such as political discussions, is the relatively low levels of self-selection. While political discussions disproportionately attract extreme partisans, nonpolitical games may tap into a wider audience looking for socially distant entertainment, especially during the COVID-19 pandemic. But the fact that people do enjoy politics when presented in this format is still surprising given Americans' apathy towards politics and disdain towards partisans [109]. Perhaps the game's design did not allow for deep engagement with the information presented which helped make politics palatable, or maybe the players were simply buoyed to learn that Americans were not as divided as they thought. More research is needed to understand what aspects of the game were fun to players so we can iterate and learn from it. Further, in GuesSync!, I designed the game to reduce misperceptions about party supporters, an approach known to reduce hostility. This is just one of many viable strategies, such as priming a superordinate identity that could be incorporated within game contexts. In the future, I hope to explore alternate strategies, game mechanics and storylines that can reduce hostile attitudes and behavior. Also, given the asymmetry in polarization, I expect to focus on interventions that are effective for ordinary Republicans.

### 5.3.3 Beyond hostility in online political interactions

Given the contentious partisan environment we currently inhabit, this dissertation focused primarily on countering hostility towards the outparty. Yet, a singular focus on hostility obscures the complexity of public discourse and privileges the status quo by focusing on the tone rather than substance [145]. I am keen to explore how other important qualities of political discussions, such as inclusivity, are affected in this contentious political environment. For example, are less hostile spaces more inclusive because they nurture more cordial conversational norms or do they exclude dissenting voices to maintain decorum? To answer such questions, we need to build more computational models to detect other ideals of deliberation and apply them to online political discussions.

Further, hostility in this dissertation is conceptualized as engaging in personal attacks and harassment, a definition that emphasizes the tone of political talk. More recently, researchers have challenged this emphasis on tone and instead highlighted the need to focus on the substance of political talk, for example, expressions of intolerance against minority groups [187]. In the future, I look to expand my research to consider the substance of political interactions together with the tone and how they are consequential for developing an inclusive democracy that works for all.

# BIBLIOGRAPHY


[1] Alan Abramowitz and Jennifer McCoy. United states: Racial resentment, negative partisanship, and polarization in trump's america. *The ANNALS of the American Academy of Political and Social Science*, 681(1):137–156, 2019.

[2] Douglas J Ahler and Gaurav Sood. Measuring perceptions of shares of groups. *Misinformation and Mass Audiences*, pages 71–90, 2018.

[3] Douglas J Ahler and Gaurav Sood. The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, 80(3):964–981, 2018.

[4] Kimberley Allison and Kay Bussey. Communal quirks and circlejerks: A taxonomy of processes contributing to insularity in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2020.

[5] Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. Political discussions in homogeneous and cross-cutting communication spaces. In *International AAAI Conference on Web and Social Media*, volume 13, 2019.

[6] Monica Anderson and Brooke Auxier. 55% of u.s. social media users say they are 'worn out' by political posts and discussions. *Pew Research Center https://www.pewresearch.org/fact-tank/2020/08/19/55-of-u-s-social-media-users-say-they-are-worn-out-by-political-posts-and-discussions/*, 2020.

[7] Monica Anderson and Brooke Auxier. 55% of us social media users say they are 'worn out'by political posts and discussions. *Pew Research Center https://www. pewresearch. org/fact-tank/2020/08/19/55-of-us-social-media-userssay-they-are-worn-out-by-political-posts-and-discussions*, 2020.

[8] Antonio A Arechar and David G Rand. Turking in the time of covid. *Behavior research methods*, 53(6):2591–2595, 2021.

[9] Miles T Armaly and Adam M Enders. Affective polarization and support for the us supreme court. *Political Research Quarterly*, page 10659129211006196, 2021.

[10] André Bächtiger, John S Dryzek, Jane Mansbridge, and M Warren. Deliberative democracy. *The Oxford handbook of deliberative democracy*, page 1, 2018.

[11] André Bächtiger, John S Dryzek, Jane Mansbridge, and Mark E Warren. *The Oxford handbook of deliberative democracy*. Oxford University Press, 2018.

[12] André Bächtiger, Maija Setälä, and Kimmo Grönlund. Towards a new era of deliberative mini-publics. *Deliberative mini-publics: Involving citizens in the democratic process*, pages 225–246, 2014.

[13] Chris Bail. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press, 2021.

[14] Stefano Balietti, Lise Getoor, Daniel G Goldstein, and Duncan J Watts. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, 118(52), 2021.

[15] Kevin K Banda and John Cluverius. Elite polarization, party extremity, and affective polarization. *Electoral Studies*, 56:90–101, 2018.

[16] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1), 2015.

[17] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *International AAAI Conference on Web and Social Media*, 2020.

[18] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[19] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political analysis*, 20(3):351–368, 2012.

[20] Christopher Birchall. Trying not to fall out: the importance of non-political social ties in online political conversation. *Information, Communication & Society*, 23(7):963–979, 2020.

[21] Toby Bolsen and James N Druckman. Counteracting the politicization of science. *Journal of Communication*, 65(5), 2015.

[22] Lorenzo Bosi, Marco Giugni, and Katrin Uba. *The consequences of social movements*. Cambridge University Press, 2016.

[23] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617, 2017.

[24] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Cross-country trends in affective polarization. Technical report, National Bureau of Economic Research, 2020.

[25] Jack W Brehm. A theory of psychological reactance. 1966.

[26] Marilynn B Brewer. Beyond the contact hypothesis: Theoretical perspectives on desegregation. *Groups in contact: The psychology of desegregation*, pages 281–302, 1984.

[27] Marilynn B Brewer. Reducing prejudice through cross-categorization: Effects. *Reducing prejudice and discrimination*, pages 165–185, 2000.

[28] David Broockman, Joshua Kalla, and Sean Westwood. Does affective polarization undermine democratic norms or accountability? maybe not. *OSF Preprints. December*, 22, 2020.

[29] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

[30] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[31] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[32] Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4745–4756, 2019.

[33] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[34] M Keith Chen and Ryne Rohla. The effect of partisanship and political advertising on close family ties. *Science*, 360(6392):1020–1024, 2018.

[35] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230, 2017.

[36] Rune Haubo Bojesen Christensen. ordinal—regression models for ordinal data. *R package version*, 28:2015, 2015.

[37] William G. Cochran. *Sampling Techniques, 3rd Edition*. John Wiley & Sons, 3rd edition edition, Jan 1977.

[38] Pamela Johnston Conover and Patrick R Miller. Taking everyday political talk seriously. *The Oxford handbook of deliberative democracy*, 2018.

[39] Pamela Johnston Conover and Donald D. Searing. Studying 'everyday political talk'in the deliberative system. *Acta Politica*, 40(3):269–283, 2005.

[40] Pamela Johnston Conover, Donald D. Searing, and Ivor M. Crewe. The deliberative potential of political discussion. *British journal of political science*, 32(1):21–62, 2002.

[41] Pamela Johnston Conover, Donald D Searing, and Ivor M Crewe. The deliberative potential of political discussion. *British journal of political science*, pages 21–62, 2002.

[42] Alexander Coppock. Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods*, 7(3):613–628, 2019.

[43] Alexander Coppock, Thomas J Leeper, and Kevin J Mullinix. Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446, 2018.

[44] Irene Costera Meijer and Tim Groot Kormelink. Checking, sharing, clicking and linking: Changing patterns of news use between 2004 and 2014. *Digital journalism*, 3(5):664–679, 2015.

[45] Richard J Crisp and Miles Hewstone. Multiple social categorization. *Advances in experimental social psychology*, 39:163–254, 2007.

[46] Richard J Crisp, Judi Walsh, and Miles Hewstone. Crossed categorization in common ingroup contexts. *Personality and Social Psychology Bulletin*, 32(9):1204–1218, 2006.

[47] Alison Dagnes. Us vs. them: Political polarization and the politicization of everything. In *Super Mad at Everything All the Time*, pages 119–165. Springer, 2019.

[48] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[49] Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 2015.

[50] Nicholas Dias and Yphtach Lelkes. The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science*, 2021.

[51] Carroll Doherty, Jocelyn Kiley, and Nida Asheer. Partisan antipathy: More intense, more personal. *Pew Research Center*, 2019.

[52] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. (mis-) estimating affective polarization. 2020.

[53] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. Affective polarization, local contexts and public opinion in america. *Nature human behaviour*, 5(1):28–38, 2021.

[54] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. How affective polarization shapes americans' political beliefs: A study of response to the covid-19 pandemic. *Journal of Experimental Political Science*, 8(3):223–234, 2021.

[55] James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. How affective polarization shapes americans' political beliefs: A study of response to the covid-19 pandemic. *Journal of Experimental Political Science*, 8(3):223–234, 2021.

[56] James N Druckman and Matthew S Levendusky. What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122, 2019.

[57] M. Duggan and A. Smith. The tone of social media discussions around politics. *Pew Research Center http://www. pewinternet. org/2016/10/25/the-tone-of-social-media-discussions-around-politics*, 2016.

[58] William P Eveland Jr and Myiah Hutchens Hively. Political discussion frequency, network size, and "heterogeneity" of discussion as predictors of political knowledge and participation. *Journal of communication*, 59(2):205–224, 2009.

[59] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1175–1184, 2010.

[60] Philip M Fernbach and Leaf Van Boven. False polarization: Cognitive mechanisms and potential solutions. *Current Opinion in Psychology*, 43:1–6, 2022.

[61] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[62] Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. Political sectarianism in america. *Science*, 370(6516):533–536, 2020.

[63] James Fishkin, Nikhil Garg, Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, Sukolsak Sakshuwong, Alice Siu, and Sravya Yandamuri. Deliberative democracy with the online deliberation platform. In *The 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019). HCOMP*, 2018.

[64] Jennifer Fitzgerald. What does "political" mean to you? *Political Behavior*, 35(3):453–479, 2013.

[65] Mary Flanagan. *Critical play: Radical game design*. MIT press, 2009.

[66] George Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 157–166, 2006.

[67] Samuel L Gaertner and John F Dovidio. Common ingroup identity model. *The encyclopedia of peace psychology*, 2011.

[68] Kristin N Garrett and Alexa Bankert. The moral roots of partisan division: How moral conviction heightens affective polarization. *British Journal of Political Science*, 50(2):621–640, 2020.

[69] R Kelly Garrett, Jacob A Long, and Min Seon Jeong. From partisan media to misperception: Affective polarization as mediator. *Journal of Communication*, 69(5):490–512, 2019.

[70] Bill Gaver, Tony Dunne, and Elena Pacenti. Design: cultural probes. *interactions*, 6(1):21–29, 1999.

[71] Bryan T Gervais. Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, 12(2):167–185, 2015.

[72] Noam Gidron, James Adams, and Will Horne. How ideology, economics and institutions shape affective polarization in democratic polities. In *Annual conference of the American political science association*, 2018.

[73] Pablo González, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz. A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40, 2017.

[74] Jeffrey Gottfried and Jacob Liedke. Partisan divides in media trust widen, driven by a decline among republicans. *Pew Research Center*, 2021.

[75] Todd Graham. Beyond "political" communicative spaces: Talking politics on the wife swap discussion forum. *Journal of Information Technology & Politics*, 9(1):31–45, 2012.

[76] Todd Graham, Daniel Jackson, and Scott Wright. From everyday conversation to political action: Talking austerity in online 'third spaces'. *European Journal of Communication*, 2015.

[77] Catherine Grevet, Loren G Terveen, and Eric Gilbert. Managing political differences in social media. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1400–1408, 2014.

[78] Andrew M Guess, Dominique Lockett, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, and Jason Reifler. "fake news" may have limited effects beyond increasing beliefs in false claims. 2020.

[79] Jurgen Habermas and Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991.

[80] Jeffrey T Hancock, Catalina L Toma, and Kate Fenner. I know something you don't: the use of asymmetric personal information for interpersonal advantage. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 413–416, 2008.

[81] Tabea Hässler, Johannes Ullrich, Michelle Bernardino, Nurit Shnabel, Colette Van Laar, Daniel Valdenegro, Simone Sebben, Linda R Tropp, Emilio Paolo Visintin, Roberto González, et al. A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour*, 4(4):380–386, 2020.

[82] Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.

[83] Carolyn M Hendriks, John S Dryzek, and Christian Hunold. Turning up the heat: Partisanship in deliberative innovation. *Political studies*, 55(2):362–383, 2007.

[84] Marc J Hetherington and Thomas J Rudolph. *Why Washington won't work*. University of Chicago Press, 2015.

[85] Itai Himelboim, Eric Gleave, and Marc Smith. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of computer-mediated communication*, 14(4):771–789, 2009.

[86] Michael A Hogg and John C Turner. Intergroup behaviour, self-stereotyping and the salience of social categories. *British Journal of Social Psychology*, 26(4):325–340, 1987.

[87] Leonie Huddy and Alexa Bankert. Political partisanship as a social identity. In *Oxford research encyclopedia of politics*. 2017.

[88] Leonie Huddy, Lilliana Mason, and Lene Aarøe. Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, 109(1), 2015.

[89] Leonie Huddy and Omer Yair. Reducing affective polarization: Warm group relations or policy compromise? *Political Psychology*, 42(2):291–309, 2021.

[90] Myiah J Hutchens, Jay D Hmielowski, and Michael A Beam. Reinforcing spirals of political discussion and affective polarization. *Communication Monographs*, 86(3):357–376, 2019.

[91] Collective Identity. The struggle for social equality: Collective action versus prejudice reduction. *Intergroup misunderstandings: Impact of divergent social realities*, page 291, 2009.

[92] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark Ackerman, and Eric Gilbert. Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms.

[93] Shanto Iyengar and Masha Krupenkin. The strengthening of partisan affect. *Political Psychology*, 39:201–218, 2018.

[94] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146, 2019.

[95] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431, 2012.

[96] Shanto Iyengar and Sean J Westwood. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707, 2015.

[97] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. " did you suspect the post would be removed?" understanding user reactions to content removals on reddit. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–33, 2019.

[98] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[99] Nathan P Kalmoe and Lilliana Mason. *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. University of Chicago Press, 2022.

[100] Geoff Kaufman and Mary Flanagan. A psychologically "embedded" approach to designing games for prosocial causes. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(33), Oct 2015.

[101] Geoff Kaufman, Mary Flanagan, and Max Seidman. Creating stealth game interventions for attitude and behavior change: An "embedded design" model. *Persuasive Gaming in Context*, page 73, 2015.

[102] Bumsoo Kim, Ryan Broussard, and Matthew Barnidge. Testing political knowledge as a mediator of the relationship between news use and affective polarization. *The Social Science Journal*, pages 1–13, 2020.

[103] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. Starrythoughts: Facilitating diverse opinion exploration on social issues. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–29, 2021.

[104] Joohan Kim and Eun Joo Kim. Theorizing dialogic deliberation: Everyday political talk as communicative action and dialogue. *Communication Theory*, 18(1):51–70, 2008.

[105] Joohan Kim and Eun Joo Kim. Theorizing dialogic deliberation: Everyday political talk as communicative action and dialogue. *Communication Theory*, 18(1):51–70, 2008.

[106] Yonghwan Kim, Hsuan-Ting Chen, and Homero Gil De Zúñiga. Stumbling upon news on the internet: Effects of incidental news exposure and relative entertainment use on political engagement. *Computers in human behavior*, 29(6):2607–2614, 2013.

[107] Jon Kingzette, James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, 85(2):663–677, 2021.

[108] Samara Klar. When common identities decrease trust: An experimental study of partisan women. *American Journal of Political Science*, 62(3):610–622, 2018.

[109] Samara Klar, Yanna Krupnikov, and John Barry Ryan. Affective polarization or partisan disdain? untangling a dislike for the opposing party from a dislike of partisanship. *Public Opinion Quarterly*, 82(2):379–390, 2018.

[110] Samara Klar, Yanna Krupnikov, and John Barry Ryan. Affective polarization or partisan disdain?untangling a dislike for the opposing party from a dislike of partisanship. *Public Opinion Quarterly*, 82(2):379–390, Jun 2018.

[111] Robert E Kraut, John M Levine, Marisol Martinez Escobar, and Amaç Herdağdelen. What makes people feel close to online groups? the roles of group attributes and group types. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 382–392, 2020.

[112] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012.

[113] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1559–1568. ACM, 2012.

[114] Yanna Krupnikov and Adam Seth Levine. Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, 1(1):59–80, 2014.

[115] Yanna Krupnikov and John Barry Ryan. *The Other Divide: Polarization and Disengagement in American Politics*. Cambridge University Press, 2022.

[116] Anurag Kumar and Bhiksha Raj. Classifier risk estimation under limited labeling resources. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–15. Springer, 2018.

[117] Alex Leavitt. " this is a throwaway account" temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 317–327, 2015.

[118] Jeffrey Lees and Mina Cikara. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3):279–286, 2020.

[119] Jeffrey Lees and Mina Cikara. Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B*, 376(1822):20200143, 2021.

[120] Yphtach Lelkes and Sean J Westwood. The limits of partisan prejudice. *The Journal of Politics*, 79(2):485–501, 2017.

[121] Russell Lenth. Package 'emmeans'.

[122] Josh A Lerner. *Making democracy fun: How game design can empower citizens and transform politics*. MIT Press, 2014.

[123] Matthew Levendusky. *The partisan sort: How liberals became Democrats and conservatives became Republicans*. University of Chicago Press, 2009.

[124] Matthew Levendusky. *How partisan media polarize America*. University of Chicago Press, 2013.

[125] Matthew Levendusky and Neil Malhotra. Does media coverage of partisan polarization affect political attitudes? *Political Communication*, 33(2):283–301, 2016.

[126] Matthew S Levendusky. Americans, not partisans: Can priming american national identity reduce affective polarization? *The Journal of Politics*, 80(1):59–70, 2018.

[127] Matthew S Levendusky. When efforts to depolarize the electorate fail. *Public Opinion Quarterly*, 82(3):583–592, 2018.

[128] Matthew S Levendusky. Our common bonds: Using what americans share to help bridge the partisan divide. *Unpublished Manuscript, University of Pennsylvania*, 2020.

[129] Matthew S Levendusky and Neil Malhotra. (mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*, 80(S1):378–391, 2016.

[130] Matthew S Levendusky and Dominik A Stecula. *We Need to Talk: How Cross-Party Dialogue Reduces Affective Polarization*. Cambridge University Press, 2021.

[131] Stephan Lewandowsky and Sander van der Linden. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384, Jul 2021.

[132] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *International conference on ubiquitous computing*, pages 261–278. Springer, 2006.

[133] Fabienne Lind, Maria Gruber, and Hajo G Boomgaarden. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures*, 11(3):191–209, 2017.

[134] P Liu, J Guberman, L Hemphill, and A Culotta. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Proceedings of the 12th International Conference on Web and Social Media*, 2018.

[135] Michael MacKuen, Jennifer Wolak, Luke Keele, and George E Marcus. Civic engagements: Resolute partisanship or reflective deliberation. *American Journal of Political Science*, 54(2):440–458, 2010.

[136] Jane Mansbridge. Everyday talk in the deliberative system. In *Deliberative Politics: Essays on Democracy and Disagreement*, pages 1–211. Oxford University Press, 1999.

[137] Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. A systemic approach to deliberative democracy. *Deliberative systems: Deliberative democracy at the large scale*, page 1–26, 2012.

[138] Alice Marwick, Benjamin Clancy, and Katherine Furl. Far-right online radicalization: A review of the literature. *The Bulletin of Technology & Public Life*, 2022.

[139] Alice E Marwick and Danah Boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

[140] Lilliana Mason. A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, 80(S1):351–377, 2016.

[141] Lilliana Mason. *Uncivil agreement: How politics became our identity*. University of Chicago Press, 2018.

[142] Lilliana Mason and Julie Wronski. One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology*, 39:257–277, 2018.

[143] Adrienne Massanari. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[144] Adrienne Lynne Massanari. Participatory culture, community, and play. *Learning from*, 2015.

[145] Gina Masullo Chen, Ashley Muddiman, Tamar Wilner, Eli Pariser, and Natalie Jomini Stroud. We should not get rid of incivility online. *Social Media+ Society*, 5(3):2056305119862641, 2019.

[146] Colleen McClain. 70% of u.s. social media users never or rarely post or share about political, social issues. *Pew Research Center*, 2021.

[147] Christopher McConnell, Yotam Margalit, Neil Malhotra, and Matthew Levendusky. The economic consequences of partisanship in a polarized era. *American Journal of Political Science*, 62(1):5–18, 2018.

[148] Joseph S Mernyk, Sophia L Pink, James N Druckman, and Robb Willer. Correcting inaccurate metaperceptions reduces americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, 119(16):e2116851119, 2022.

[149] Patrick R Miller and Pamela Johnston Conover. Red and blue states of mind: Partisan hostility and voting in the united states. *Political Research Quarterly*, 68(2):225–239, 2015.

[150] Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. Analyzing genetic testing discourse on the web through the lens of twitter, reddit, and 4chan. *ACM Transactions on the Web (TWEB)*, 14(4):1–38, 2020.

[151] Samantha L Moore-Berg, Boaz Hameiri, and Emile Bruneau. The prime psychological suspects of toxic political polarization. *Current Opinion in Behavioral Sciences*, 34:199–204, 2020.

[152] Robert L Morgan, James E Whorton, and Cynthia Gunsalus. A comparison of short term and long term retention: Lecture combines with discussion versus cooperative learning. *Journal of instructional psychology*, 27(1):53, 2000.

[153] Chantal Mouffe. Politics and passions. *Ethical Perspectives*, 7(2-3):146–150, 2000.

[154] Patricia Moy and John Gastil. Predicting deliberative conversation: The impact of discussion networks, media use, and political cognitions. *Political Communication*, 23(4), 2006.

[155] Emily Moyer-Gusé and Robin L Nabi. Explaining the effects of narrative in an entertainment television program: Overcoming resistance to persuasion. *Human communication research*, 36(1):26–52, 2010.

[156] Ashley Muddiman and Natalie Jomini Stroud. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4):586–609, 2017.

[157] Luke Munn. Alt-right pipeline: Individual journeys to extremism online. *First Monday*, 2019.

[158] Sean Munson and Paul Resnick. The prevalence of political discourse in non-political blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[159] Sean A Munson, Stephanie Y Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Seventh international aaai conference on weblogs and social media*, 2013.

[160] Sean A Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466, 2010.

[161] Diana C Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.

[162] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. (re) design to mitigate political polarization: Reflecting habermas' ideal communication space in the united states of america and finland. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[163] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. (re) design to mitigate political polarization: Reflecting habermas' ideal communication space in the united states of america and finland. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):141, 2019.

[164] Brendan Nyhan. Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[165] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[166] Jeeyun Oh and S Shyam Sundar. How does interactivity persuade? an experimental test of interactivity on cognitive absorption, elaboration, and attitudes. *Journal of Communication*, 65(2):213–236, 2015.

[167] Ray Oldenburg. *Celebrating the third place: Inspiring stories about the great good places at the heart of our communities*. Da Capo Press, 2001.

[168] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872, 2009.

[169] Lilla V Orr and Gregory A Huber. The policy basis of measured partisan animosity in the united states. *American Journal of Political Science*, 64(3):569–586, 2020.

[170] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015, 2021.

[171] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[172] John Parkinson and Jane Mansbridge. *Deliberative systems: Deliberative democracy at the large scale*. Cambridge University Press, 2012.

[173] Charles J Pattie and Ron J Johnston. It's good to talk: Talk, disagreement and tolerance. *British Journal of Political Science*, 38(4):677–698, 2008.

[174] Jorge Pena, Juan Francisco Hernández Pérez, Subuhi Khan, and Ángel Pablo Cano Gómez. Game perspective-taking effects on players' behavioral intention, attitudes, subjective norms, and self-efficacy to help immigrants: the case of "papers, please". *Cyberpsychology, Behavior, and Social Networking*, 21(11):687–693, 2018.

[175] Wei Peng, Mira Lee, and Carrie Heeter. The effects of a serious game on role-taking and willingness to help. *Journal of communication*, 60(4):723–742, 2010.

[176] Thomas F Pettigrew. Intergroup contact theory. *Annual review of psychology*, 49(1):65–85, 1998.

[177] Tom Postmes and Nancy Baym. Intergroup dimensions of the internet. *Intergroup communication: Multiple perspectives*, 2:213–240, 2005.

[178] Tom Postmes, Russell Spears, and Martin Lea. Intergroup differentiation in computer-mediated communication: Effects of depersonalization. *Group Dynamics: Theory, Research, and Practice*, 6(1):3, 2002.

[179] Stephen A Rains, Kate Kenski, Kevin Coe, and Jake Harwood. Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, 22(4):163–178, 2017.

[180] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *International AAAI Conference on Web and Social Media*, 2020.

[181] Stig Hebbelstrup Rye Rasmussen, Mathias Osmundsen, and Michael Bang Petersen. Political resources and online political hostility how and why hostility is more prevalent among the resourceful. 2022.

[182] Steve Rathje, Jay J Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), 2021.

[183] Stephen D. Reicher, Russell Spears, and Tom Postmes. A social identity model of deindividuation phenomena. *European review of social psychology*, 6(1):161–198, 1995.

[184] John P Robinson and Mark R Levy. Interpersonal communication and news comprehension. *Public Opinion Quarterly*, 50(2):160–175, 1986.

[185] Jon C Rogowski and Joseph L Sutherland. How ideology fuels affective polarization. *Political behavior*, 38(2):485–508, 2016.

[186] Jon Roozenbeek and Sander van der Linden. Breaking harmony square: A game that "inoculates" against political misinformation. *The Harvard Kennedy School Misinformation Review*, 2020.

[187] Patricia Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425, 2022.

[188] Erin Rossiter. The consequences of interparty conversation on outparty affect and stereotypes. Technical report.

[189] Gina Roussos and John F Dovidio. Playing below the poverty line: Investigating an online game as a way to reduce prejudice toward the poor. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(2), 2016.

[190] Kai Ruggeri, Bojana Većkalov, Lana Bojanić, Thomas L Andersen, Sarah Ashcroft-Jones, Nélida Ayacaxli, Paula Barea-Arroyo, Mari Louise Berge, Ludvig D Bjørndal, Aslı Bursalıoğlu, et al. The general fault in our fault lines. *Nature Human Behaviour*, 5(10):1369–1380, 2021.

[191] Dana Ruggiero. The effect of a persuasive social impact game on affective learning and attitude. *Computers in Human Behavior*, 45:213–221, 2015.

[192] Dana Ruggiero. The effect of a persuasive social impact game on affective learning and attitude. *Computers in Human Behavior*, 45:213–221, Apr 2015.

[193] Tamar Saguy, Nicole Tausch, John F Dovidio, and Felicia Pratto. The irony of harmony: Intergroup contact can produce false expectations for equality. *Psychological Science*, 20(1):114–121, 2009.

[194] Kathy Sanford, Lisa J Starr, Liz Merkel, and Sarah Bonsor Kurki. Serious games: video games for good? *E-Learning and Digital Media*, 12(1):90–106, 2015.

[195] David Schleifer, Will Friedman, and Erin McNally. Putting partisan animosity in perspective: A hidden common ground report. 2021.

[196] Rüdiger Schmitt-Beck and Christiane Grill. From the living room to the meeting hall? citizens' political talk in the deliberative system. *Political Communication*, 37(6):832–851, 2020.

[197] Donald D Searing, Frederick Solt, Pamela Johnston Conover, and Ivor Crewe. Public discussion in the deliberative system: does it make better citizens? *British Journal of Political Science*, 2007.

[198] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong'Cherie' Chen, Likang Sun, and Geoff Kaufman. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 606. ACM, 2019.

[199] Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 111–125, 2017.

[200] Bryan Semaan, Heather Faucett, Scott P. Robertson, Misa Maruyama, and Sara Douglas. Designing political deliberation environments to support interactions in the public sphere. CHI '15, page 3167–3176. ACM, 2015.

[201] Bryan C Semaan, Scott P Robertson, Sara Douglas, and Misa Maruyama. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1409–1421, 2014.

[202] Bryan C. Semaan, Scott P. Robertson, Sara Douglas, and Misa Maruyama. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work  social computing*, page 1409–1421. ACM, 2014.

[203] Jaime E Settle. *Frenemies: How social media polarizes America*. Cambridge University Press, 2018.

[204] Jaime E Settle and Taylor N Carlson. Opting out of political discussions. *Political Communication*, 36(3):476–496, 2019.

[205] Richard M Shafranek. Political considerations in nonpolitical decisions: a conjoint analysis of roommate choice. *Political Behavior*, pages 1–30, 2019.

[206] Dhavan V Shah, Jaeho Cho, William P Eveland Jr, and Nojin Kwak. Information and expression in a digital age: Modeling internet effects on civic participation. *Communication research*, 32(5):531–565, 2005.

[207] Frank M Shipman and Catherine C Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, 1999.

[208] Simon Buckingham Shum et al. Cohere: Towards web 2.0 argumentation. *COMMA*, 8:97–108, 2008.

[209] Aaron Smuts. Are video games art? *Contemporary Aesthetics (Journal Archive)*, 3(1):6, 2005.

[210] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 2019.

[211] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[212] Natalie Jomini Stroud, Ashley Muddiman, and Joshua M Scacco. Like, recommend, or respect? altering political behavior in news comment sections. *New media & society*, 19(11):1727–1743, 2017.

[213] Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, 23(1):95–115, 2018.

[214] Jiyoun Suk, David Coppini, Carlos Muñiz, and Hernando Rojas. The more you know, the less you like: A comparative study of how news and political conversation shape political knowledge and affective polarization. *Communication and the Public*, page 20570473211063237, 2021.

[215] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3401–3410, 2011.

[216] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16, 1979.

[217] Martin Tanis and Tom Postmes. A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European Journal of Social Psychology*, 35(3):413–424, 2005.

[218] Katie Salen Tekinbas and Eric Zimmerman. *Rules of play: Game design fundamentals*. MIT press, 2003.

[219] Emily Thorson. Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3):460–480, 2016.

[220] Chau Tong, Hyungjin Gill, Jianing Li, Sebastián Valenzuela, and Hernando Rojas. "fake news is anything they say!"—conceptualization and weaponization of fake news among the american public. *Mass Communication and Society*, 23(5):755–778, 2020.

[221] W Ben Towne and James D Herbsleb. Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, 9(1):97–115, 2012.

[222] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.

[223] Sahana Udupa. Nationalism in the digital age: Fun as a metapractice of extreme speech. *International Journal of Communication*, pages 3143–3163, 2019.

[224] Nicholas A Valentino, Ted Brader, Eric W Groenendyk, Krysha Gregorowicz, and Vincent L Hutchings. Election night's alright for fighting: The role of emotions in political participation. *The Journal of Politics*, 73(1):156–170, 2011.

[225] Aart van Stekelenburg, Gabi Schaap, Harm Veling, and Moniek Buijzen. Correcting misperceptions: The causal role of motivation in corrective science communication about vaccine and food safety. *Science Communication*, 42(1):31–60, 2020.

[226] Sai Wang. The influence of anonymity and incivility on perceptions of user comments on news websites. *Mass Communication and Society*, 23(6):912–936, 2020.

[227] Steven W Webster and Alan I Abramowitz. The ideological foundations of affective polarization in the us electorate. *American Politics Research*, 45(4):621–647, 2017.

[228] Anita Whiting and David Williams. Why people use social media: a uses and gratifications approach. *Qualitative market research: an international journal*, 2013.

[229] Jason Wilson. Hiding in plain sight: How the 'alt-right'is weaponizing irony to spread fascism'. *The Guardian*, 23, 2017.

[230] Magdalena Wojcieszak and R Kelly Garrett. Social identity, selective exposure, and affective polarization: How priming national identity shapes attitudes toward immigrants via news selection. *Human communication research*, 44(3):247–273, 2018.

[231] Magdalena Wojcieszak and Diana Mutz. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 2009.

[232] Magdalena Wojcieszak and Benjamin R Warner. Can interparty contact reduce affective polarization? a systematic test of different forms of intergroup contact. *Political Communication*, 37(6):789–811, 2020.

[233] Magdalena Wojcieszak, Stephan Winter, and Xudong Yu. Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. *Mass Communication and Society*, 23(4):455–483, 2020.

[234] Scott Wright. From "third place" to "third space": Everyday political talk in non-political online spaces. *Javnost-the public*, 19(3):5–20, 2012.

[235] Scott Wright, Todd Graham, and Dan Jackson. Third space, social media, and everyday political talk. In *The Routledge companion to social media and politics*, pages 74–88. Routledge, 2015.

[236] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, 2017.

[237] Robert O Wyatt, Elihu Katz, and Joohan Kim. Bridging the spheres: Political and personal conversation in public and private spaces. *Journal of communication*, 50(1):71–92, 2000.

[238] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 2020.

[239] Wenjie Yan, Gayathri Sivakumar, and Michael A Xenos. It's not cricket: examining political discussion in nonpolitical online space. *Information, Communication & Society*, 21(11), 2018.

[240] Michael Yeomans, Julia Minson, Hanne Collins, Frances Chen, and Francesca Gino. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, 2020.

[241] Iris Marion Young. *Inclusion and democracy*. Oxford University press on demand, 2002.

[242] Iris Marion Young. Difference as a resource for democratic communication. In *Not for sale*, pages 109–129. Routledge, 2020.

[243] Xudong Yu, Magdalena Wojcieszak, Seungsu Lee, Andreu Casas, Rachid Azrout, and Tomasz Gackowski. The (null) effects of happiness on affective polarization, conspiracy endorsement, and deep fake recognition: Evidence from five survey experiments in three countries. *Political behavior*, 43(3):1265–1287, 2021.

[244] Jason Shuo Zhang, Brian C Keegan, Qin Lv, and Chenhao Tan. Understanding the diverging user trajectories in highly-related online communities during the covid-19 pandemic. *International AAAI Conference on Web and Social Media*, 2021.

[245] Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540, 2009.