Shame on Who? Experimentally Reducing Shame During Political Arguments on Twitter

AMANDA BAUGHAN, School of Computer Science & Engineering, University of Washington
KATHERINE CROSS, Information School, University of Washington
ELENA KHASANOVA, Department of Computational Linguistics, University of Washington
ALEXIS HINIKER, Information School, University of Washington

Online political arguments have a reputation for being futile exchanges, partially because people often respond more punitively to those who do not share their views, a phenomenon called ingroup bias. We explore how ingroup bias affects political disagreements online, and how respect can mitigate its effects. redTowards this goal, we conducted an experiment on Twitter systematically varying respectful versus neutral language across people who did and did not share views. We found that people who do not share views are most likely to reply to disagreements, and neutral disagreements generated more discussions than respectful disagreements. However, we also found that using respectful language reduces shaming responses overall, and reduces the effects of ingroup bias across conversations with people who do and do not share views. We conclude with recommendations to promote respectful language on social media and build shame resiliency online, such as through design that encourages thoughtful engagement, and a peer support network that allows users to share shame experiences online.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: social media, group dynamics, politics, shame, respect

ACM Reference Format:
Amanda Baughan, Katherine Cross, Elena Khasanova, and Alexis Hiniker. 2018. Shame on Who? Experimentally Reducing Shame During Political Arguments on Twitter. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

People express widespread frustration with the experience of discussing political differences online. They report that online political disagreements escalate into hostile and unproductive clashes [23, 49], rife with personal insults and shaming pejoratives [13]. These concerns are common; in one national survey, more than 80% of Americans said they felt incivility in online discourse was a serious or very serious problem [30]. Analysis of online content supports this perception, finding that incivility is widespread in online forums for news and politics [13]. As a result, many people say they avoid political discussions online altogether so as not to risk participating in a disagreement that might lead to personal insults and shaming statements [3].

At first glance, these past findings might seem to suggest that discussions of political differences are best reserved for offline interactions or set aside altogether. However, other work shows that

CSCW '21, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

supporting constructive political disagreement online is a worthy goal. For example, although many people currently avoid political conversations, they also say they *wish* they could debate political issues and other challenging topics online [3]. This interest in arguing constructively is well-founded, as debate and deliberation are essential to healthy democracy [26, 51].

What leads political disagreements online to shift from respectful debate to hostile shaming? And what, if anything, can be done to help people disagree and debate more productively? In offline life, disagreements are more likely to turn hostile because of *ingroup bias*, a phenomenon in which people respond more favorably to those with whom they share a group identity. As a result of ingroup bias, people respond more constructively to conflict with ingroup members compared to outgroup members [61]. People are also more likely to humanize and forgive flaws in ingroup members than outgroup members [37], creating more room for respectful disagreement within identity groups. People can form ingroups and display ingroup bias across both arbitrary groups [58] and those based on demographics, ideology, or politics [21]. Thus, ingroup bias might make it easier for people who share a group identity to disagree constructively online and harder for those with differing group identities to do so.

In this work, we first examine how one type of group membership—political identity—affects civility during political disagreements online. Political affiliation is an increasingly polarizing aspect of one's identity, and contempt for people who associate with an opposing political party has steadily increased in recent years [17, 18]. This strong and growing antipathy across groups suggests the potential for ingroup bias to affect online disagreements, such that people respond more favorably and constructively to disagreements with someone who shares their political affiliation. Specifically we asked:

> **RQ1:** Holding constant the substance of a disagreement, how likely are people to use respectful and shaming language when disagreeing with someone who shares their political identity as compared to someone who does not?

Second, we investigate one potential intervention to combat incivility and increase respectful disagreement. Offline, ingroup bias can be reduced through explicit statements of respect toward outgroups, which helps to signal equality between groups and reduces the need for ingroups to strongly defend and distance themselves from an outgroup [21]. Here, we examine whether this phenomenon translates to a digital context, asking:

> **RQ2:** Does adding a respectful preamble to a statement of disagreement online increase civility from the respondent (as measured by the prevalence of respectful and shaming language)?
>
> **RQ3:** Does adding this respectful preamble to a statement of disagreement online increase civility between people with different political identities?

By investigating these questions, we sought to understand whether differences in political identity make civil, productive disagreement more difficult, irrespective of the substance of the disagreement. And if so, we sought to understand whether a simple intervention could alleviate some of this difficulty, an important goal, given the value in debating political issues with others and the growing antipathy across the political divide.

To address these questions, we conducted an experimental study of political disagreements on Twitter. Using two artificial Twitter personas (one representing a mainstream American conservative and one representing a mainstream American liberal), we systematically disagreed publicly with others online. These disagreements were counterbalanced across Twitter users who both did and did not share the political identity of the responding persona. We used scripted responses for all of our statements of disagreement, but we varied whether or not these scripts began with a respectful preamble. We hypothesized that Twitter users' responses to our statements of disagreement 1)

would be less respectful and more likely to use shaming language when political identities differed, 2) would be more respectful and less shaming when a respectful preamble was used, and 3) the difference in responses across ingroup and outgroup disagreements would be reduced in the presence of a respectful preamble. We found that all of these hypotheses held.

Even a very brief preamble signaling respect significantly shifted the tone of conversations away from shame and toward respect. This has implications for both designing social platforms and for engaging in conversations on these platforms. For example, social media platforms may be able to influence discourse to become more respectful through micro-interventions that suggest respectful language use. While we were encouraged to see a simple intervention systematically promoted more respectful disagreements in online political discussions, we also note that this did not eliminate problematic discourse altogether, and future work will need to consider many avenues for fostering productive political disagreements online.

2 BACKGROUND

Here we review literature related to our goal of exploring how respectful language can shape the direction of political disagreements online. First, we review the large body of literature regarding online political conflicts. Next, we discuss how the perception of shared ingroup identity (or lack thereof) relates to people's use of respectful or shaming language in general. We formulate our hypotheses in the context of this prior work.

2.1 Online Conflicts

When people interact online, subtle but valuable nonverbal cues that characterize in-person communication are stripped away. Specific affordances of online contexts, such as invisibility, perceived anonymity, and lack of eye contact decrease empathy online and reduce people's likelihood of adhering to social norms, a phenomenon known as "online disinhibition" [27, 40, 57, 63]. As a result, people tend to find less common ground online [41], and online conflict is considered one of the biggest stressors of using social media [23].

Because of these affordances that undermine mutual understanding, when people are online, they tend to avoid debate and disagreement, particularly when discussing polarizing topics like politics [1, 14, 24, 64]. As a result, people most often see and interact with those they share views with, which is often referred to as "echo chambers" [1, 14]. When people do engage "across the aisle," political online arguments have garnered a reputation for being unproductive exchanges that devolve into "shouting matches" [49] and "comment wars" [23], and rude posts are particularly likely to garner engagement when they are about politics [9]. However, recent research has also found that people *want* to be able to discuss their views openly [3], and many people want to use online communities to gain exposure to opinions different than their own [32, 47, 52], suggesting it is hostility and lack of respect rather than differences of opinion that people are avoiding. And even if users do not usually change their mind, they find that the process of engaging in a debate online can be rewarding if conducted civilly [32]. Thus, regardless of the outcome, whether an argument is conducted respectfully is an important measure of its success.

Prior work has also examined these themes specifically in the context of Twitter, one of the most common places in which online political arguments occur [19], and the platform we used to conduct our study. The design of any platform impacts how its users communicate [3], and prior work has critiqued Twitter's design for its (in)ability to support constructive disagreements. For example, Liu and Weber [42] found that Twitter is "not an ideal public sphere for democratic conversations" because of the social hierarchies that exist on the platform and the frequency of low-quality comments. Prior work also shows that there is also a high degree of political homophily on Twitter, as people tend to mostly follow those that they agree with [14]. Similarly, Yardi and Boyd

[64] found that Twitter arguments most often reinforce pre-existing views, potentially because tweets lack the contextual depth necessary for engaging in constructive dialogue [51]. Users have also said that Twitter's design makes it very easy to see people's opinions and start arguments, because users' posts are more visible in comparison to other social networking sites [3]. On Twitter, users' awareness of their audience also decreases their willingness to be vulnerable [3] and freely share their opinions [16, 44]. Collectively, this work highlights some of the challenges to arguing well that are inherent in online environments. We build on this prior research by exploring a potential intervention for increasing respect during online arguments, which, if successful, has the potential to enable users to move past their current avoidance and to engage in the conversations they wish they could have.

2.2 The Role of Shame and Respect in Intergroup Conflicts

Prior works have found that ingroup bias appears both in online and offline political discussions [21, 64]. People negatively evaluate information that threatens their political ingroup [15, 21], thus arguments may be threatening when they imply the ingroup's convictions are false, inferior, or even morally objectionable [6], or when they call the ingroup's identity-defining opinions or values into question [55]. These threats may be interpreted as "face threats" in which a person's desired image ("face") is undermined [25], and people often experience shame in response to face threats [33]. Shame may be defined as an intensely painful feeling or experience of believing we are flawed and therefore unworthy of acceptance and belonging [7]. Shame is inextricably interpersonal and tied to people's need for social connection [7]. An "unwanted identity" can be a powerful elicitor of shame, in which people perceive others ascribing to them characteristics that undermine their desired identity or self-ideals [7]. It is through this lens that we introduce our first hypothesis:

H1: Holding the substance of the disagreement equal, subjects will use more respectful language when disagreeing with those who share their political identity, and conversely, use more shaming language with those who do not share their political identity.

However, when people's collective identity is buffered against threat, they are less likely to discredit contradictory opinions [48]. In a lab study, Eschert and Simon [21] found that people rated ingroup arguments as stronger on a Likert scale than those of an outgroup, demonstrating ingroup bias. However, by letting participants know all views would be respected and taken seriously before their ratings, participants demonstrated less biased evaluations of strength of outgroup arguments. This shows that respect can play a pivotal role in people's judgements of each other's views. Eschert and Simon [21] argued that public political disagreements are struggles for equality, not superiority, and respect signals equality between disagreeing parties. Huo and Molina [29] argued that respect reduces ingroup bias and improves intergroup relations because members of respected groups have less psychological need to show ingroup bias as a means to defend their collective identity. Thus, we introduce our second hypothesis:

H2: When disagreements include respectful preambles, subjects will respond with more respectful and less shaming language.

Online, when an outgroup states contrary opinions, ingroup members re-establish shared values of the ingroup and distance themselves from the outgroup [64]. An unstudied aspect of this phenomena is how shame is used during this response. Public shaming is used to sanction "unacceptable" views and behavior [4], and the use of shame online is often perceived as justified [4]. Users who shame others on Twitter tend to have a faster and more sharp increase in followers than those of non-shamers [2]. Such shaming of others online has lead to increased criticism of "cancel culture," and researchers have shown that being "cancelled" online can lead to overly attributing blame to

individuals and de-contextualizing discussions of nuanced topics [5]. On Twitter in particular, while those who publicly shame others may "*revel in and enjoy their shared moral position*," the tweets can also have the effect of misrepresenting and distracting from the actual issues [5]. Ingroup bias may lead ingroups to focus on displaying their moral capital, rather than deconstructing how someone in an outgroup can come to display the views and behaviors that are found to be intolerable [5]. Put simply, "*Shaming and dehumanizing people and holding them accountable* are *mutually exclusive. Shame is not a social justice tool*" [8]. The widely popular use of shame to sanction others online is especially troubling given that when individuals believe they are stigmatized offline for their views, they often adopt and endorse even more extreme views, meaning that shame may further polarize people [35]. Therefore, we introduce a third hypothesis:

H3: Explicit cues towards respectful consideration of all viewpoints will mitigate the effects of ingroup bias by increasing respect and reducing shame.

3 METHODS

To answer our research questions, we conducted a between-subjects experimental study on Twitter. We artificially constructed two Twitter personas, one intended to represent a user with U.S. liberal views and one intended to represent a user with U.S. conservative views. Our two artificial personas responded to the tweets of others by disagreeing. Our responses systematically engaged with those in their ingroup or outgroup, and in each case, responses were counterbalanced to add a respectful preamble half the time. This created a two-by-two experimental design in which the two variables we manipulated were: relationship (ingroup or outgroup) and tone (respectful or neutral). Below we describe our procedures for creating these profiles, responding to other users' tweets, ethical considerations, and analysis.

3.1 Materials

We created two Twitter accounts, one appearing to be a mainstream American liberal, the other a mainstream American conservative. These accounts had the same location of a city known to be split politically, and they followed either the Democratic (U.S. liberal) or Republican (U.S. conservative) party and national figures, as well as state-level politicians and a collection of interests such as science, weather, and humor accounts. Each Twitter profile had an English male name among the top ten most popular names for babies born in the United States in 1967. Each Twitter account had a dog for a profile photo to keep race and appearance ambiguous, as prior work has shown that the gender and race of online accounts can influence users' responses, even if the content is the same [62]. The bios of the Twitter accounts featured common hashtags or phrases from liberal or conservative Twitter accounts, such as "#BlackLivesMatter" and "Proud Patriot," respectively. These hashtags were selected after a contextual analysis of a randomly selected sample of Twitter accounts that followed one of the 2020 presidential nominees at the time: Joe Biden or Donald Trump. The choice of using the #BlackLivesMatter hashtag is additionally supported by Dunn [20]'s work on the most politically divisive issues in 2020, in which 76% of Democrats think how racial minorities are treated in the justice system is a very big problem, compared to 20% of Republicans.

To validate that users would perceive the personas as their intended political affiliation, we asked 53 survey respondents on Mechanical Turk to rate these profiles as right- or left-leaning. Respondents accurately identified and were confident in their identification of the political party of the personas at a similar level to a randomly chosen group of 18 other left or right leaning Twitter accounts. These randomly chosen accounts were labeled by the criteria in Table 1. In the analysis of

our results, we discuss "ingroups" as generally expressing similar views, supporting similar party candidates, and perceived to be part of the same political party as validated in this survey.

Table 1. Categorization guide for liberal and conservative Twitter accounts. We excluded leftists, alt-right, and QAnon accounts from our experiment.

	Liberal	**Conservative**
Inclusion	• States being liberal • Hashtags in support of Biden or Black Lives Matter (e.g. #RESIST, #BlueWave2020, #BLM) • Retweets/follows/voices support for Democratic politicians • Uses phrase "Trump Virus"	• States being conservative • Fox news followers • "Proud" American, Patriot • Mentions God and family values • Retweets/follows/voices support for Republican politicians, Trump
Exclusion	Any of the leftist identifiers: • States being leftist, socialist, or communist • Support of progressive Democratic politicians such as Bernie Sanders, Alexandria Ocasio-Cortez exclusively • Critique of mainstream liberals	Any of the alt-right or QAnon identifiers: • Explicit mentions of Gab, Parler accounts • Mentions "white nationalism" or other racist terms • Profile photo of Pepe the Frog meme, classical statues, or Nazis • Explicit mentions of known conspiracy websites and/or QAnon. (e.g. #WWG1WGA, human trafficking, distrust of Democratic party and celebrities) • Reference to the "truth" or "unity" beyond parties • Crititque of mainstream media (#MSM)
Example bio	Mom, Wife, Arts Adminstrator, Reader, Sailor, Traveler. I would like to see my kids have a bright future. #BlueWave2020	Family Man, Supporting our President everyday. Keep America Great, No Excuses.
Example tweet	*No family in middle-class America is saying: "Thank goodness Mitch McConnell is eliminating the duty of employers to act reasonably and protect employees from COVID." They want a good outcome; not the dangerous outcome businesses want in exchange for Senate campaign donations.*	*ACA is useless. It just continues to line the pockets of insurance companies because the deductibles are so high that lower income people still can't afford medical care. They still need to pay out of pocket. Raise Medicaid limits if you really want to help low-income people.*

To create the script of tweets that each account would send, we selected a subset of issues that American voters said they considered "*very big*" or "*moderately big*" problems according to a June 2020 survey by Pew Research Center [20]. These topics included the U.S. national response to coronavirus, federal budget deficit, ethics in government, healthcare affordability, unemployment, and mail-in voting. The statements each persona made on each issue were developed to be in alignment with prominent politicians from the Democratic and Republican party during August 2020 (see Table 2). We also reviewed left-leaning, right-leaning, and centrist media such as The New York Times, Fox News, and The Hill to validate that the statements we crafted aligned with the positions of liberal or conservative journalism. Determinations about media political biases were made by allsides.com[1] and confirmed by mediabiasfactcheck.com[2]. For each issue, the statements we created for the liberal and conservative accounts were designed to be similar in character count, formality, tone, and syntax.

[1]https://www.allsides.com
[2]https://www.mediabiasfactcheck.com

To verify that these statements were perceived in the way we intended, we asked 77 participants on Mechanical Turk to rate the political leanings of each statement on a 5-point Likert scale from "*very liberal*" to "*very conservative*" and their confidence in their ratings. In all cases except for the intentionally duplicated coronavirus prompt (see Table 2, row 3), there was a statistically significant result of participants accurately rating the political leaning of each tweet. Therefore, after the analysis we added explicit identifiers to coronavirus arguments that stated, "*As a [liberal/conservative].*" Our sixth topic, mail-in voting, was added after this confirmatory analysis in response to a rise in the partisan discussion of this topic on Twitter and a decrease in discussions of other topics. Thus, our statements on this topic were not validated by Mechanical Turk workers, however, we added an explicit identifier to the tweets from the conservative persona to reinforce its identity. As discussions on Twitter do not always reflect the most important issues to voters [65], we accepted this as a necessary mitigation strategy in order for our study to remain feasible.

When tweeting any one of these statements, we added either a neutral or respectful preamble per our experimental conditions. The respectful prompt was, "*I respect your views, and I think…*" The neutral prompt stated, "*I disagree, because…*" or the tweet was sent without any declaration of disagreement. These preambles were kept short, as prior work has shown that even a brief indicator of respect can have a significant effect [21] and because Twitter has a 240-character limit per tweet. The tweet content contained slight deviations to maintain a conversational tone and not sound too scripted, as we tweeted in response to slightly different points raised on these topics.

3.2 Subjects

Subjects were selected initially based on their stated stance in reply to a popular US politician or news site that tweeted on a topic listed in Table 2. We then evaluated their bios and recent tweets to categorize them as either liberal or conservative according to the criteria listed in Table 1. In developing a categorization method, three of the researchers on the project coded randomly sampled Twitter accounts as leftist, liberal, conservative, or alt-right/QAnon. The researchers iteratively revised code definitions and re-coded new accounts until 80% interrater reliability was achieved. We aimed to tweet at users who were based in the US and not bots. Subjects who explicitly stated being nationals of a different country were excluded, as were accounts that had been created since June 2020 with a non-human profile picture and/or no followers, similar to how prior works have identified bot accounts [45, 46].

3.3 Procedure

Every day for two weeks, a researcher logged into the liberal and conservative accounts and found 20 tweets per profile to reply to on the topics in Table 2. Ten of these conversations were with their political ingroup, and ten were with their political outgroup. Of these, half were explicitly respectful, and half were neutral, and these conditions were randomly assigned. In all instances, the researcher's tweet expressed disagreement with the subject, and it reflected the stance of its persona (i.e., the liberal persona always responded with a statement from the "Liberal" column in Table 2). All tweets had to have occurred within the past 24 hours. We only responded to subjects one time and did not engage in follow-up. This resulted in 40 tweets in response to subjects per day for 14 days, with the exception of the first day and the day U.S. presidential candidate Joe Biden announced Kamala Harris as his vice-presidential running mate, which limited the amount of conversations on other political topics. One day we replied to tweets that occurred within the past 36 instead of 24 hours. This resulted in 531 total tweets sent during August 2020. We debriefed subjects and offered them $5 USD compensation if they had responded to our tweets. Not all subjects accepted direct messages on Twitter, so we were not able to contact all of our subjects, but every person who

accepted direct messages received debrief information and an offer for compensation, redthough none responded.

3.4 Ethical Considerations

Our study involved deceiving people on Twitter and collecting data without their informed consent or debriefing. We carefully considered the risks and benefits to subjects when creating the study design and looked to prior literature that has engaged in similar practices for guidance. Hudson and Bruckman [28] initially showed that obtaining consent is impractical in chatroom settings, as only 0.5% of participants chose to do so, and they conclude that obtaining a waiver of consent, which waives the requirement to obtain informed consent, is appropriate in such research. Munger [45] was the first researcher to our knowledge to create Twitter "sock puppet" accounts, followed by Siegel and Badaan [53] and Munger [46]. In these works, subjects were not debriefed on the deception used in the study. However, in our work we chose to debrief subjects and allow them several weeks to respond with questions or concerns. None of the subjects chose to receive compensation or remove their data.

There is of course some risk of discomfort inherent to having one's views challenged. To mitigate the extent to which participants might experience discomfort or the potential for this discomfort to

Table 2. The scripts that each account executed daily. An explicitly neutral phrase ("*I disagree, because…*") or respectful phrase ("*I respect your views, and I think…*") was added as a preamble to these statements.

	Liberal	**Conservative**
Federal budget	The federal deficit will likely be over $3.7T, thanks to Trump's administration's spending. Democrats are working to make sure every American can survive the pandemic financially.	We already have a federal deficit, and liberals wants to increase spending. The GOP is protecting Americans and small businesses by keeping taxes and federal costs low during this recession.
Unemployment	Biden's unemployment plan can help workers and businesses through the pandemic and beyond. His plan's flexibility offers protection to workers and would lead to less layoffs, helping the economy.	Our economy is already in recession, the government can't risk giving incentives so people prefer not to work more than return to work. Unemployment not be a pay raise, just enough to get by in an emergency.
National Response to COVID-19	Deaths from the pandemic aren't due to any single party or political figure. This situation is changing so quickly, and we have to focus on our shared future rather than assigning blame. Our lives depend on it.	*The same as the left*
Ethics in Government	Trump's administration has shown too many suspicious behaviors to think there hasn't been something to cover up. We need an AG & justice system that holds everyone accountable equally, including our leaders.	Democrats claim they have more than circumstantial evidence that there was collusion between Trump and Russia, but we haven't seen it. We can't allow this discrediting of the legally elected President.
Healthcare	Healthcare should be affordable for everyone, especially during this pandemic. Building on the ACA and incorporating Medicare will help every American get the care they need. Healthcare is a human right.	Obamacare doesn't work - for the country or its citizens. The ACA has led to very high costs for American people and businesses. The middle class is suffering due to government failures.
Mail-In Voting	It is unfortunate to see that voting by mail is now political. I live in WA state, where mail-in voting has occurred securely for years, with any fraud caught quickly. Even Fox News reports that it is safe to vote by mail.	There's a lot of fear that mail-in voting could lead to fraud. I'm a conservative myself, and I've lived in WA for years, which has mail-in voting. It's never been an issue. These fraud criticisms are unfounded.

be harmful, we stated only established facts and expressed mainstream views, presented via neutral and respectful tones. We took great care not to incorporate inflammatory or shaming language in our study materials, or contribute to extremist narratives or disinformation campaigns, so as to not "poison the stream to see how the fish respond." This limits our study's scope, however, this limitation was ethically necessary.

Given the public and conversational nature of Twitter, we believe our research activities presented no further risk than is already present in using such a platform, which is known to contain both genuine and non-genuine political interactions, and they comprised a minute proportion of the overall volume of political conversations on Twitter. Our study script and subject responses have been paraphrased to protect subject privacy. Our university's IRB approved our data collection methods.

3.5 Analysis

redHere, we describe the data we collected and how we qualitatively coded it in preparation for quantitative analysis. Of the 531 tweets we posted, 141 received at least one reply. When two or more people tweeted separately in response to our prompt, those were considered separate responses. This resulted in 178 responses from unique subjects. Anyone who replied who was not the original subject was categorized as conservative, liberal, or neither based on the criteria in Table 1; anyone who was "neither" was removed, for a final total of 163 tweets used in the following analysis.

3.5.1 Qualitative Coding. We were interested primarily in observing instances of respectful or shaming responses, as respect creates tolerance for different perspectives [21, 32], whereas shame is used to police "unacceptable" outgroup views, and create distance from an outgroup [61]. As such, we coded the 163 responses for presence of language that was respectful, shaming, or neither. To do so, three members of the research team conducted iterative, blind closed-coding on the replies. We reviewed the codes three times separately, and iterated until we arrived at the following final codebook:

- **Shaming** classifications required 1) the user to directly reference the research persona, and 2) in referencing the research persona, describe them or their stance with language that imposed a socially undesirable identity or degraded them personally [39]. These slights might have been subtle or understated, as long as they satisfied these dimensions. Statements that shamed people or groups (such as political parties) other that the original poster were not automatically considering shaming. Tweets that showed strong negative affect, anger, or even hostility toward the original poster were not automatically considered shaming. An example shaming tweet is, "*Did you ever hear of the f-king pandemic. The budget grew faster under all presidents put together and that's the truth. I hope you're not one of leftwing loonies moving to FL because of the looters and high taxes in your leftwing state. If so, don't come.*"
- **Respectful** classifications required the user to be dignity-affirming and imply or state equal consideration of viewpoints, even under disagreement. Language markers of respect were drawn from prior work in natural language processing [60] and social psychology [21], including gratitude, semantic softeners (also known as "hedging"), an appeal towards shared values such as democracy, or understanding or agreement towards some aspect of the argument. For example, "*Fair enough. The problem is, what are the solutions? One side wants to lock down the country. Neither side wants to provide people with aid without adding in a separate agenda. So, given that dynamic, the real solution is (oddly enough) term limits and getting money out of politics.*"
- Tweets categorized as "**neither**" met none of the above requirements, and they encompassed a wide range of tone and sentiment, from openly hostile and challenging to simple agreements.

For example "*Biden has been in politics for what, 50 years? His time has come and gone. We need change and new ideas, whereas his are beyond dated. Have a good night,*" and "*They're lying to you. Democratic politicians all belong in prison with the RINOs! Trump 2020.*"

In our last round of coding, two coders were in absolute agreement on the categorization of tweets (Cohens κ = 1.00). The first author then re-coded the entire dataset according to the codebook, and any tweets that were considered unclear (n = 23) were isolated and re-coded by the three coders until agreement had been reached.

3.5.2 Quantitative Analysis. We used the coded responses to conduct a series of quantitative analysis, consisting of multinomial logistic regressions and Chi-square tests. Our multinomial logistic regression analysis contained response type [shame | respect | neither] as the dependent variable and prompt type [respectful | neutral], relationship [ingroup | outgroup], the interaction between prompt type and relationship as independent variables. When we designed this experiment, we anticipated that the people we replied to would be the only ones who responded to our tweets. However, we found that the original recipients and new users replied in similar proportion ($\chi^2(1) = 0.38, n.s.$), so we added this as an independent variable [recipient | newcomer] to our regression. There was no distinguishable interaction effect between prompt type and relationship, so it was removed from the model.

Chi-square and *post hoc* tests were corrected for multiple comparisons using the Holm-Bonferroni method. Note that multinomial logistic regressions generate model coefficients with regards to a "reference category." For instance, to understand the impact of various predictive variables on the responses we coded, we will see a coefficient for shaming and respectful responses, but not "neither" responses, as "neither" comprises the reference category. All statistical tests were conducted in R and are submitted as part of the supplementary materials.

4 RESULTS

In our experiment, we evaluated how introducing respect changed the tone of disagreements between people in political ingroups and outgroups. Of the 163 responses we received, 26 were respectful, 33 were shaming, and 104 were neither shaming nor respectful. See Table 3 for counts of how responses were distributed. and Table 4 for the results of the multinomial logistic regression.

Table 3. Number of responses received across categories.

	Relationship		**Prompt**		**Responder**	
	Outgroup	Ingroup	Neutral	Respectful	Recipient	New
Respectful	10	16	7	19	6	20
Shaming	28	5	23	10	17	16
Neither	55	49	73	31	56	48
Totals	**93**	**70**	**103**	**60**	**70**	**93**

4.1 H1: Shared Political Identity Correlates with More Respect and Less Shaming

The multinomial logistic regression revealed that when people shared the same political identity as our persona, they were more likely to reply to our prompts with respect than when their political identity differed ($\beta = 0.99, p = .048$, odds ratio = 2.69). When they shared the same political identity as the persona, they were also less likely to reply to our prompts with shaming responses ($\beta = -1.66, p = .0018$, odds ratio = 0.19). This is consistent with H1, in which we predicted shared political identity would be associated with more respectful language and differing political identity would be associated with more shaming language. When people's identity differed from that of the

Table 4. Results of the multinomial logistic regression.

Response Type	Independent Variables	β	se	Z	odds ratio	p
Respectful	**Shared Identity**	**0.99**	**0.50**	**1.98**	**2.69**	**0.048**
	Respectful Prompt	**1.91**	**0.52**	**3.67**	**6.75**	**0.0002**
	Original Recipient	0.80	0.54	1.47	2.22	0.14
	Intercept	-3.43	0.65	-5.26	0.03	0.00
Shaming	**Shared Identity**	**-1.66**	**0.53**	**-3.12**	**0.19**	**0.0018**
	Respectful Prompt	-0.20	0.46	-0.43	0.82	0.67
	Original Recipient	-0.20	0.42	-0.47	0.82	0.63
	Intercept	-0.50	0.35	-1.41	0.61	0.15

persona who disagreed with them, they were also marginally more likely to respond, as revealed by a Chi-square test ($\chi^2(1) = 3.25, p = 0.07$, see Table 3). This may mean people are more motivated to respond to defend themselves during disagreements with someone from an outgroup.

4.2 H2: Respect Leads to More Respect, but Does Not Impact Shaming Responses

Unsurprisingly, respectful prompts were positively correlated with respectful responses ($\beta = 1.91, p = .0002$, odds ratio = 6.75). However, there was not a significant correlation between respectful prompts and shaming responses ($\beta = -0.20, p = 0.67$, odds ratio = 0.82). This result is surprising, as we anticipated that respect would also result in reduced shaming language. A Chi-square analysis revealed that neutral prompts received more shaming replies ($n = 23$) than respectful prompts did ($n = 10, \chi^2(1) = 5.12, p = 0.037$), however, this result may not have been impactful enough to significantly factor into the multinomial model when considered amongst the other independent variables. We found that overall, people responded more to neutral prompts ($n = 103$) than respectful prompts ($n = 60$, $\chi^2(1) = 11.34, p < .001$). This is explained by people replying to neutral prompts with twice the number of responses that were *neither* shaming nor respectful ($n = 73$) relative to the response rate to respectful prompts ($n = 31, \chi^2(1) = 16.96, p < .001$).

4.3 H3: Using Respect with the Outgroup Yields Similar Quality Discourse to Neutrality with the Ingroup

Most notably, our results revealed no detectable difference between using *respectful* prompts with *outgroups* and using *neutral* prompts with *ingroups* ($\chi^2(6) = 6.85, p = 0.76$). See Table 5 for an overview of *post hoc* pairwise multinomial logistic regressions and Fig. 1 for how prompt type and relationship impact type of response received. This suggests that introducing respect into a conversation with an outgroup leads to similar quality discourse as neutral discussion in ingroups. While it appears that ingroup conversations may benefit from respectful cues more than outgroup conversations do ($\chi^2(6) = 14.87, p = .085$), it is promising to see that of respect has the ability to elevate the tenor of discourse in both ingroup and outgroup disagreements. It is also interesting that there is no significantly distinguishable difference in comparing outgroup conversations under neutral and respectful prompts ($\chi^2(6) = 3.94, p = 0.76$), but there is for ingroups ($\chi^2(6) = 19.56, p = .017$) indicating that ingroup conversations are most influenced to reciprocate prosocial cues online. In fact, when respect was used in ingroup disagreements, the amount of shaming responses was reduced to zero.

5 DISCUSSION

Our experimental results confirm the existence of ingroup bias in Twitter users' responses to political disagreements. Specifically, we saw that people were more likely to respond with respect

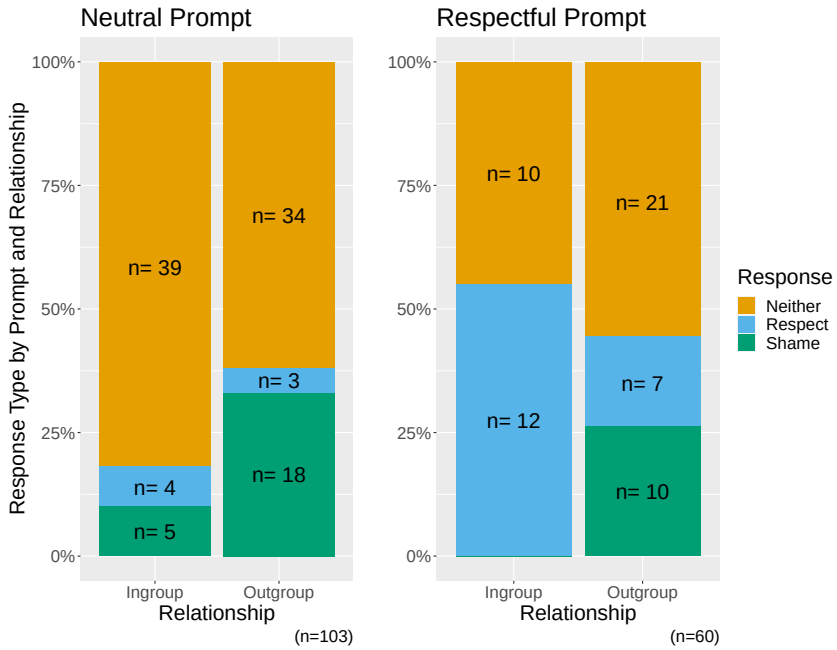

Fig. 1. Percentage of shaming and respectful responses across relationships and prompt types. We can see that outgroup responses with respectful prompts and ingroup responses with a neutral prompt have similar levels of shaming and respectful discourse. While not displayed in the graph "neither" responses are factored into the percentage calculations.

Table 5. Post hoc tests on how relationship and prompt type influence receiving respectful, shaming, and neither responses. Ingroups benefit the most from using respect, both compared to outgroup conversations with respect and neutrality within ingroup. Most notably, respectful outgroup disagreements yield similar quality discourse as neutral ingroup disagreements.

Conversation Type	**Conversation Type**	df	χ^2	p
Ingroup & Respectful (n = 22)	Outgroup & Respectful (n = 38)	6	14.865	0.085
Ingroup & Respectful (n = 22)	Ingroup & Neutral (n = 48)	6	19.56	0.017
Ingroup & Respectful (n = 22)	Outgroup & Neutral (n = 55)	6	29.96	<.001
Ingroup & Neutral (n = 48)	Outgroup & Respectful (n = 38)	6	6.85	0.76
Outgroup & Respectful (n = 38)	Outgroup & Neutral (n = 55)	6	3.94	0.76
Ingroup & Neutral (n = 48)	Outgroup & Neutral (n = 55)	6	7.81	0.76

when we used an ingroup profile and more likely to respond with shaming language when we used an outgroup profile. One surprising result was that outgroup members responded to our prompts far more often than ingroup members. We were also surprised to find that people were more likely to reply to neutral prompts than respectful ones. We theorize that outgroup status and neutral prompts create more face-threatening situations for recipients, and as such, people may feel more compelled to defend themselves from these face-threats and associated feelings of shame. This is supported by research which shows that ingroup members are motivated to negatively evaluate information that threatens their collective identities [15].

However, by adding respect to a statement of disagreement, the tone of the conversation shifted such that there was no statistically detectable difference between outgroup members' responses to respectful prompts and ingroup members' responses to neutral prompts. Prior work has shown that manipulating an individual's perception that they are respected and considered as equally as outgroup members causally influences them to have a less biased view of outgroup arguments [21]. Our results confirm this work and take it a step further: introducing respect also impacts people's *behavior*, such that it lessens the effects of ingroup bias. We further show that an online, text-based environment provides sufficient affordances for respect cues to trigger this response. This suggests a mechanism by which users might improve the tenor of disagreements online—conversations which are a crucial element of a pluralistic society and, today, fraught with hostility.

Surprisingly, respectful and shaming language made up a relatively small proportion of the responses we received (16% and 20% respectively). While analysis along other dimensions of language online were ultimately out of scope for this study, it does raise the question of what other elements of users' responses are worth evaluating during political disagreements. In particular, it raises the question of how people experience shaming online – are there messages people receive that are not shaming under our definition, but still produce feelings of shame for recipients? Does hostility towards another member of an ingroup evoke feelings of shame? Future works could more deeply explore what leads to feelings of shame online, and how people absorb or distance themselves from shame directed towards others online. Additionally, the fact that using respect did *not* have an influence on receiving shaming responses indicates that a different intervention may be more effective for reducing shaming online. Future works could evaluate people's motivations for using shame during disagreements and how to build resilience in the face of shame.

Our experiment shows that a relatively small adjustment to language ("*I respect your views*") can yield significant positive changes to the tone of discourse. And, it has been shown that online social cues tend to yield reciprocal social cues [10], which suggests that once respectful dialogue begins, it will continue. To invoke such respectful dialogue, a variety of implicit and explicit nudges can be employed through interaction design, as prior works have found that the design of online platforms influences perceptions of arguments [3]. For instance, organization of content on a page and how interactions with other users are scaffolded can have drastic impacts on how users behave on a site [38, 54, 56]. Interfaces that are intentionally designed to promote thoughtfulness can inspire users to be more thoughtful [56], and interaction design that explicitly encourages reflective "listening" online causes users to consider others' perspectives more deeply [38]. Even micro-interventions such as CAPTCHAs that prime users to experience low-arousal positive emotions can significantly increase the positivity and social connectedness in subjects' posts online [50]. As such, it may be possible for designers to craft interventions on social media to encourage the respectful cues that we found to improve discourse in our study.

Social media platforms might also promote respectful disagreements through algorithmic adjustments. Prior works have found that, today, more negative and emotionally divergent (high positive and negative sentiment) tweets are more likely to go viral [31], and users who post shaming content tend to get more followers [2]. Twitter's timeline algorithm decides what is shown to users based on features of tweets (such as the number retweets or likes), features of the tweet's author, and tweets each user has found engaging in the past [36]. These algorithmic decisions may currently reinforce a cycle of ingroup and outgroup ideals, in which users respond by shaming views or people they disagree with, which distances them from the outgroup [64]. In turn, algorithms promote content users have interacted with, which may continue to strengthen the ingroup identity in a way that highlights negative content [31], further driving a divide between ingroups and outgroups. This may lead people to further dehumanize and become frustrated towards outgroup members, leading

to more interpretation of neutral stances as shaming or negative, sparking defensiveness, and continuing a cycle of shaming and polarization.

redIt is also important to note that our analysis revealed that outgroup disagreements did not measurably improve through the introduction of respectful language. While shaming responses did decrease (from n =18 to n =10) and respectful responses increased (from n = 3 to n = 7), these changes were not statistically significant. In contrast, respectful disagreement did measurably improve the tone of ingroup responses. This may explain some of the feelings of futility of online political disagreements - efforts to improve discourse do not necessarily result in improved outcomes. One possible approach to address this is to emphasize a shared, broader group identity during disagreements. For instance, in the context of our study, emphasizing shared identity as Americans, or along other demographics and interests, could have been more impactful. Emphasizing common ground with others could be more important than equality-based respect, and future works could explore how such an intervention elevates discourse.

Future works may also focus on building *shame resilience* online. Brown [7]'s work identifies "critical awareness" of the socio-cultural influences on shame as a mechanism for normalizing and contextualizing shame experiences, as opposed to low critical awareness, which leads to reinforcing shame. That is, by helping people understand the universality of shaming experiences and giving them language to deconstruct it, they become more skilled in processing and moving past feelings of shame. An intervention as simple as informing users about how people's behaviors become more uninhibited online, and how ingroup dynamics impact how people communicate, may lead to more critical awareness of the prevalence of shaming online and help users maintain more emotional distance from shaming messages they receive. Another aspect of shame resilience is the construct of "reaching out" [7], in which one person seeks support about shame experience from another. With the right relational support, people can name and identify common experiences, create change, and even share knowing laughter about their shame experiences. This opposes remaining alone and feeding shame with secrecy and silence, which limits opportunities for change [7]. A potential intervention for people experiencing shame during online discussions may be to connect to a peer support network, dedicated to supporting people feeling shame as a result of interactions in the public sphere online. Given the rise of concern over public shaming online, this could be a very timely sociotechnical intervention.

Some may disagree that social media platforms should try to nudge users towards particular behavior or styles of discourse. However, we argue that this is likely already occurring, given the prevalence of negative, shaming, and argumentative content online [2, 10, 31]. Social media platforms are normative and influence behavior (whether they intend to or not), platforms and their users could benefit from intentionality regarding such norms. Stanfill [54] states that by utilizing Foucault and Ewald [22]'s concept of power as productive, designers can encourage particular actions (such as respect) in addition to forbidding undesired ones (such as abusive language). Additionally, researchers have found evidence that popular social media sites are not as value-neutral as they might aspire to be [12, 34, 43]. As long as products exist which scaffolding users' interactions online, there is a responsibility to do so in the least harmful and biased way possible. Incorporating changes towards more intentional and respectful interactions may change the tenor of online arguments for the better.

Limitations

There are a few notable limitations to consider in this work. The first is that we used profiles that implied a male identity to the user. The lack of diversity among personas prevents the study from making clear claims about race. We mitigated the effects of appearance in our interactions by using an animal profile photo, however, because of what is known about prejudice and its influence on

online harassment [11, 45, 62], having an English, masculine name likely protected these accounts from more frequent hostility and shaming. Future works could explore how intersectional identities across race and gender influence ingroup dynamics during political disagreements. Additionally, as we designed our study for ecological validity, there may be unmeasured demographic differences driving the use of respect or shame. For instance, it would be interesting to couple an experimental design such as ours with an evaluation of whether the users responding are more often challenge-seeking or challenge-averse [47] and how shame resilience [7] shapes argument tenor. Additionally, while we grouped subjects within categories of political ideology, ascription of values and identity based on tweets and Twitter profiles does not capture the full spectrum of political beliefs. Finally, we captured a relatively small sample of tweet responses ($n = 163$), which limits the statistical power of our sample. However, we did meet the sample size guidance for multinomial logistic regressions, which recommends at least 10 observations per independent variable.

Additionally, we approached political discourse with the assumption that shame is generally toxic and unwarranted, and respectful disagreement with tolerance for different opinions is desired. However, respect is not always clearly superior. At its worst, signifiers of respectful discussion can be manipulated by malevolent actors in order to win converts to an extreme cause [59]. The common alt-right tactic of "love bombing"–inundating a potentially receptive person with supportive commentary with an aim towards making an extremist position seem reasonable–is a crucial example in which respectful language can be harmful. Therefore, it remains important to note that in an age when irony, insincerity, and platform-manipulation are commonplace features of social media, we cannot always regard seemingly respectful or positive conversation at face value. Future works on the sincerity of the use of respect in online political discussions may yield important insights to combat these insidious interactions. Similarly, there is a limit to the power and suitability of politeness and respect to improve online discourse. For instance, when people hold political views which are intended to disempower others, such as through racism, sexism, ableism, transphobia, and/or homophobia, respect may actually make such ideas palatable to a mainstream audience and actually be dangerous. It is important to limit the applicability of our findings to mainstream political views on public policy that do not intend to disempower others.

6 CONCLUSION

We set out to discover whether respectful language impacts political arguments online across ingroup and outgroup discussions. We found that ingroup bias does impact the use of shaming and respectful language during political arguments online, by increasing the amount of shaming language and reducing respect. However, we also found that introducing respect can change the tone of a disagreement by increasing respectful language. We note with optimism the possibility that the tone of online disagreements may be influenced towards cycles of more respectful engagement and reduced shaming through targeted sociotechnical, algorithmic, and design interventions. We make the argument that social media platforms should be intentional about how their platforms influence norms and behavior, and that it is possible to influence people's language during online disagreements for the better.

ACKNOWLEDGMENTS

We would like to thank Casey Fiesler and the ACM SIGCHI Research Ethics Committee for their feedback on our experiment. We also would like to thank our reviewers who helped strengthen our paper. This work was partially funded by Facebook.

REFERENCES

[1] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[2] R. Basak, S. Sural, N. Ganguly, and S. K. Ghosh. 2019. Online Public Shaming on Twitter: Detection, Analysis, and Mitigation. *IEEE Transactions on Computational Social Systems* 6, 2 (2019), 208–220.

[3] Amanda Baughan, Justin Petelka, Catherine Jaekyung Yoo, Jack Lo, Shiyue Wang, Amulya Paramasivam, Ashley Zhou, and Alexis Hiniker. 2021. Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.

[4] L. Blackwell, Tianying Chen, S. Schoenebeck, and C. Lampe. 2018. When Online Harassment Is Perceived as Justified. In *ICWSM*.

[5] Gwen Bouvier. 2020. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media* 38 (2020), 100431. https://doi.org/10.1016/j.dcm.2020.100431

[6] Nyla R Branscombe, Naomi Ellemers, Russell Spears, Bertjan Doosje, et al. 1999. The context and content of social identity threat. *Social identity: Context, commitment, content* (1999), 35–58.

[7] B. Brown. 2006. Shame Resilience Theory: A Grounded Theory Study on Women and Shame. *Families in Society: The Journal of Contemporary Social Services* 87 (2006), 43 – 52.

[8] Brené Brown. 2021. Brené on Words, Actions, Dehumanization, and Accountability. https://brenebrown.com/podcast/brene-on-words-actions-dehumanization-and-accountability/

[9] Moira Burke and Robert Kraut. 2008. Mind Your Ps and Qs: The Impact of Politeness and Rudeness in Online Communities. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 281–284. https://doi.org/10.1145/1460563.1460609

[10] Gaowei Chen and Ming Ming Chiu. 2008. Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers & Education* 50, 3 (2008), 678 – 692. https://doi.org/10.1016/j.compedu.2006.07.007

[11] Danielle Keats Citron. 2009. Law's expressive value in combating cyber gender harassment. *Michigan Law Review* 108, 3 (2009), 373–415.

[12] Nick Clegg. 2019. Facebook, Elections and Political Speech. *https://about.fb.com/news/2019/09/elections-and-political-speech/* (2019).

[13] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4 (06 2014), 658–679. https://doi.org/10.1111/jcom.12104 arXiv:https://academic.oup.com/joc/article-pdf/64/4/658/22322347/jjnlcom0658.pdf

[14] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* 64, 2 (2014), 317–332.

[15] Natascha De Hoog. 2013. Processing of social identity threats. *Social Psychology* (2013).

[16] Michael A. DeVito, Jeremy Birnholtz, and Jeffery T. Hancock. 2017. Platforms, People, and Perception: Using Affordances to Understand Self-Presentation on Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 740–754. https://doi.org/10.1145/2998181.2998192

[17] Carroll Doherty, Jocelyn Kiley, and Bridget Jameson. 2016. Partisanship and Political Animosity in 2016. Pew Research Center. https://www.pewresearch.org/politics/2016/06/22/partisanship-and-political-animosity-in-2016/.

[18] Carroll Doherty, Jocelyn Kiley, and Bridget Jameson. 2019. Partisan Antipathy: More Intense, More Personal. Pew Research Center. https://www.pewresearch.org/politics/2019/10/10/partisan-antipathy-more-intense-more-personal/.

[19] Meave Duggan and Aaron Smith. 2016. The Political Environment on Social Media. https://www.pewresearch.org/internet/2016/10/25/the-political-environment-on-social-media/

[20] Amina Dunn. 2020. As the U.S. copes with multiple crises, partisans disagree sharply on severity of problems facing the nation. https://www.pewresearch.org/fact-tank/2020/07/14/as-the-u-s-copes-with-multiple-crises-partisans-disagree-sharply-on-severity-of-problems-facing-the-nation/

[21] Silke Eschert and Bernd Simon. 2019. Respect and political disagreement: Can intergroup respect reduce the biased evaluation of outgroup arguments? *PloS one* 14, 3 (2019), e0211556.

[22] Michel Foucault and François Ewald. 2003. *" Society Must Be Defended": Lectures at the Collège de France, 1975-1976*. Vol. 1. Macmillan.

[23] Jesse Fox and Jennifer J. Moreland. 2015. The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. *Computers in Human Behavior* 45 (2015), 168–176.

[24] Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. 2009. Blogs are echo chambers: Blogs are echo chambers. In *2009 42nd Hawaii International Conference on System Sciences*. IEEE, 1–10.

[25] Erving Goffman. 1967. On face-work. *Interaction ritual* (1967), 5–45.

[26] Jürgen Habermas. 1984. *The theory of communicative action: Reason and the rationalization of society*. Heneimann.

[27] Sameer Hinduja and Justin W. Patchin. 2008. Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behavior* 29, 2, 129–156. https://doi.org/10.1080/01639620701457816

[28] James M. Hudson and Amy Bruckman. 2004. "Go Away": Participant Objections to Being Studied and the Ethics of Chatroom Research. *The Information Society* 20, 2 (2004), 127–139.

[29] Yuen J Huo and Ludwin E Molina. 2006. Is pluralism a viable model of diversity? The benefits and limits of subgroup respect. *Group Processes & Intergroup Relations* 9, 3 (2006), 359–376.

[30] Public Religion Research Institute. 2010. PRRI/RNS religion news survey. www.publicreligion.org.

[31] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. 2013. Analyzing and Predicting Viral Tweets. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. Association for Computing Machinery, New York, NY, USA, 657–664. https://doi.org/10.1145/2487788.2488017

[32] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. Designing for Civil Conversations: Lessons Learned from ChangeMyView.

[33] Margaret E. Kemeny, Tara L. Gruenewald, and Sally S. Dickerson. 2004. Shame as the Emotional Response to Threat to the Social Self: Implications for Behavior, Physiology, and Health. *Psychological Inquiry* 15, 2 (2004), 153–160. http://www.jstor.org/stable/20447221

[34] Vanessa Kitzie and Debanjan Ghosh. 2015. #Criming and #Alive: Network and content analysis of two sides of a story on twitter. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–10. https://doi.org/10.1002/pra2.2015.145052010041 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.145052010041

[35] Willem De Koster and Dick Houtman. 2008. 'STORMFRONT IS LIKE A SECOND HOME TO ME'. *Information, Communication & Society* 11, 8 (2008), 1155–1176. https://doi.org/10.1080/13691180802266665 arXiv:https://doi.org/10.1080/13691180802266665

[36] N. Koumchatzky and A. Andryeyev. 2017. Using Deep Learning at Scale in Twitter's Timelines. https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

[37] Peter Koval, Simon M. Laham, Nick Haslam, Brock Bastian, and Jennifer A. Whelan. 2012. Our Flaws Are More Human Than Yours: Ingroup Bias in Humanizing Negative Characteristics. *Personality and Social Psychology Bulletin* 38, 3 (2012), 283–295. https://doi.org/10.1177/0146167211423777 arXiv:https://doi.org/10.1177/0146167211423777 PMID: 21940854.

[38] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012. Is This What You Meant? Promoting Listening on the Web with Reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1559–1568. https://doi.org/10.1145/2207676.2208621

[39] Emily B Laidlaw. 2017. Online shaming and the right to privacy. *Laws* 6, 1 (2017), 3.

[40] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28 (2012), 434–443.

[41] Maciek Lipinski-Harten and Romin W. Tafarodi. 2013. Attitude moderation: A comparison of online chat and face-to-face conversation. *Computers in Human Behavior* 29 (2013), 2490–2493.

[42] Zhe Liu and Ingmar Weber. 2014. *Is Twitter a Public Sphere for Online Conflicts? A Cross-Ideological and Cross-Hierarchical Look*. Springer International Publishing, Cham, 336–347. https://doi.org/10.1007/978-3-319-13734-6_25

[43] Sofia Lundmark and Maria Normark. 2014. Designing Gender in Social Media: Unpacking Interaction Design as a Carrier of Social Norms. *International Journal of Gender, Science and Technology* 6, 2 (2014), 223–241. http://genderandset.open.ac.uk/index.php/genderandset/article/view/345

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

[45] K. Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39 (2017), 629–649.

[46] Kevin Munger. 2020. Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter. *Journal of Experimental Political Science* (2020), 1–15. https://doi.org/10.1017/XPS.2020.14

[47] Sean A. Munson and Paul Resnick. 2010. Presenting Diverse Political Opinions: How and How Much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/1753326.1753543

[48] Peter Nauroth, Mario Gollwitzer, Jens Bender, and Tobias Rothmund. 2015. Social identity threat motivates science-discrediting online comments. *PLoS one* 10, 2 (2015), e0117476.

[49] Elia Powers, Michael Koliska, and Pallavi Guha. 2019. "Shouting Matches and Echo Chambers": Perceived Identity Threats and Political Self-Censorship on Social Media. *International Journal of Communication* 13 (2019), 20.

[50] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[51] Bryan Semaan, Heather Faucett, Scott P. Robertson, Misa Maruyama, and Sara Douglas. 2015. Designing Political Deliberation Environments to Support Interactions in the Public Sphere. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3167–3176. https://doi.org/10.1145/2702123.2702403

[52] Bryan C. Semaan, Scott P. Robertson, Sara Douglas, and Misa Maruyama. 2014. 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* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1409–1421. https://doi.org/10.1145/2531602.2531605

[53] Alexandra A Siegel and Vivienne Badaan. 2020. # No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review* 114, 3 (2020), 837–855.

[54] Mel Stanfill. 2015. The interface as discourse: The production of norms through web design. *New Media & Society* 17, 7 (2015), 1059–1074. https://doi.org/10.1177/1461444814520873

[55] Walter G Stephan, C Renfro, and Mark D Davis. 2008. The role of threat in intergroup relations. (2008).

[56] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative Influences on Thoughtful Online Participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3401–3410. https://doi.org/10.1145/1978942.1979450

[57] John Suler. 2004. The Online Disinhibition Effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* 7 3 (2004), 321–6.

[58] Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* 1, 2 (1971), 149–178.

[59] Dennis Tourish and Tim Wohlforth. 2000. *On the edge: Political cults right and left.* ME Sharpe.

[60] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114, 25 (2017), 6521–6526.

[61] Michael Wenzel, Tyler Okimoto, Norman Feather, and Michael Platow. 2007. Retributive and Restorative Justice. *Law and human behavior* 32 (11 2007), 375–89. https://doi.org/10.1007/s10979-007-9116-6

[62] Senuri Wijenayake, Niels van Berkel, and Jorge Goncalves. 2020. Bots for Research: Minimising the Experimenter Effect. In *International Workshop on Detection and Design for Cognitive Biases in People and Computing Systems (CHI'20 Workshop). ACM.*

[63] Sheng Wu, Tung-Ching Lin, and Jou-Fan Shih. 2017. Examining the antecedents of online disinhibition. *IT & People* 30 (2017), 189–209.

[64] Sarita Yardi and Danah Boyd. 2010. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society* 30, 5 (2010), 316–327. https://doi.org/10.1177/0270467610380011 arXiv:https://doi.org/10.1177/0270467610380011

[65] Pei Zheng and Saif Shahin. 2020. Live tweeting live debates: How Twitter reflects and refracts the US political climate in a campaign season. *Information, Communication & Society* 23, 3 (2020), 337–357.