Incivility on Popular Politics and News Subreddits: An Analysis of In-groups, Community Guidelines and Relationships with Social Media Engagement

Chris J. Vargo[1] and Toby Hopp[2]
Advertising, Public Relations and Media Design
College of Communication, Media, and Information
University of Colorado Boulder
[1] christopher.vargo@colorado.edu [2] tobias.hopp@colorado.edu

Abstract

Political and news subreddits are individualistic as it pertains to the incivility we might expect them to exhibit; some have clear in-group members, and all have varying degrees of content moderation policies. We sample submissions (n = 127,870) and comments (n = 2,576,049) from 20 of the most popular news and politics subreddits from June 4th, 2021, to June 4th, 2022. All subreddits appear to be mostly civil, with incivility most commonly occurring in comments. When incivility occurs, it tends to take on less-severe forms including insults, profanity, and general toxicity. Subreddits with with clear political in-groups did exhibit more insults, toxicity, profanity, and identity-based attacks. The more complex a subreddit's moderation policies, the less incivility was observed. Finally, uncivil submissions do result in a mild increase in engagement, but given the overall low prevalence of incivility observed, it appears not to be integral to a subreddit's overall engagement.

Keywords: incivility, Reddit, content moderation, community guidelines, in-groups

1. Introduction

The content moderation of social media platforms has once again been thrust into the public sphere, with Elon Musk recently calling the content moderation policies of Twitter, his soon-to-possibly-be company, an inhibitor of free speech (Ovide, 2022). This concerned many journalists, as Musk has been well known to use the platform to exhibit uncivil behavior, including likening Justin Trudeau to Hitler, using misogynistic stereotypes to describe Elizabeth Warren, making ageist remarks towards Bernie Sanders, posting a photo of Bill Gates and joking that his likeness causes male impotence, and perhaps, most famously, baselessly calling a British cave rescuer a "pedo" (Levin, 2022).

It appears, at least in public opinion, the issue of moderating social media platforms for civility and issue censorship remain intertwined. For instance, in June 2020, 90% of Republicans said that it was likely that social media sites censor political viewpoints (Vogels et al., 2020). Indeed, most political pundits that have been "deplatformed" on social media are affiliated with the Republican party, most famously former president Donald Trump. However, the platforms themselves do not assert any political motivation. Instead, they deplatform accounts to reduce hate speech, trolling, personal attacks and other types of incivility, a strategy that appears to be effective (Jhaver et al., 2021).

Given the divisions on whether social media platforms should do any moderation of the messages their users generate, it is important to document the extent to which uncivil communication occurs on various social media platforms, and the extent to which platform moderation policies are effective. Here we examine a year's worth of Reddit data from popular political and news-based communities.[1] Reddit provides an interesting case study of a social media platform that has largely left its platform to human moderation a process that varies from community to community. From what we can gather via public research and journalism, Reddit is not moderated by AI, as Musk criticizes Twitter and Facebook of, but instead by human moderators. These unpaid individuals have the power to remove posts and comments that violate a subreddit's rules and guidelines.

Reddit's content moderation is primarily decentralized and hybrid, with illegal content and objectionable behaviors prohibited (Caplan, 2018). Reddit has a small team of paid administrators (~10% of the workforce) who enforce content policies, but many have noted that these individuals commonly seek to remove illegal content, not posts relating to specific

[1] Registered users of Reddit submit content in the form of text, links, photos, and videos. Its main page features top submissions from its "subreddits," which are topic-based forums. Submissions can be engaged with in various ways. They can be nominated for an award, up or downvoted, and/or replied to in the form of a comment.

URI: https://hdl.handle.net/10125/102930
978-0-9981331-6-4

community guidelines for a subreddit (Gibson, 2019). Therefore, subreddits rely heavily on volunteers, known as moderators, that make guidelines for their subreddit which list responsibilities and expected behavior (Reddit Inc, 2017). Most content removal decisions are placed on moderators who grapple with how to best manage discussions relating to uncivil behaviors (Almerekhi et al., 2020).

Thus, it remains an open question as to the degree to which a subreddit's community guidelines are enforced, and the extent to which these policies ultimately curb incivility. For instance, in the subreddit known as /r/alltheleft, one guideline dictates that content shared in the subreddit must be related to the topic of politics. This includes posts about democratic politicians, policy, and ideological theory. Users are asked to avoid "hate speech or bigotry… classism, racism, sexism, Islamophobia, homophobia, and transphobia." On r/republican, users are also asked to be civil and to avoid personal attacks, racism, and violent content.

Despite these guidelines, recent work shows that incivility on Reddit persists (Davidson et al., 2020; Hansen, 2022; Stevens et al. 2021). Hmielowski et al. (2014) show that uncivil behaviors that are deeply traumatic to other users, such as flaming, can be socialized on Reddit in individuals if content moderation is not enforced. To this end, this paper explores the relationship the incivility of a subreddit has with the strictness of its community guidelines.

Moreover, one of the biggest differences that can be readily observed by reviewing the most active subreddits is that some have clear intentions to cater to, and include certain members based on based on whether they identify as having a certain identity or belonging to a group. In these subreddits, in-group members are invited to participate, while often out-group members are instructed not to engage or participate.

On one hand, a subreddit with a clear in-group may reduce the number of cross-cutting conversations that occur thus reducing the number of conflicts and confrontation that comes with them (Himelboim, et al., 2013). On the other, moderators may be inclined to allow incivility, such as insults or hate speech, to be directed towards an out-group to appease users of the subreddit. For instance, political out-group content is engaged with more so than other political content on Twitter, and as such, Reddit may have incentives for content expressing out-group animosity (Rathje et al., 2021). It stands to reason that civility of a subreddit may hinge on whether it is objectively trying to advance a group, or whether it agnostic to group memberships.

To answer these questions, we draw on a sample of 20 major subreddits that cover politics and news. We assess the degree to which incivility exists in these subcommunities and the degree to which community policies and in-group status mediate that incivility. We inspect the two main ways Reddit users generate content, by making a submission to a subreddit and commenting under submissions. Jigsaw's Perspective Application Programming Interface (API) — a suite of computational annotation tools that detect toxic online comments — is used to assess differing types of incivility. Using metadata associated with each submission and comment, we conclude with an assessment of the degree to which incivility is associated with engagement on the platform, assessing the degree to which each subreddit relies on uncivil content to garner social media engagement, a necessity of a successful online community.

2. Incivility

The definition of political incivility is subject to contest (e.g., Herbst, 2010). Papachraissi (2004) has argued that incivility should be understood specifically as the intentional rejection of democratic communicative norms around inclusion and equality, and should be differentiated from interpersonal impoliteness, which tends to both pertain primarily to interpersonal conflict and to take on a spontaneous character. More recent work (e.g., Bentivegna & Rega, 2022; Kenski, Coe, & Rains, 2020; Muddiman, 2017; Rossini, 2022; Sydnor, 2018) has arrived at a slightly different conclusion, suggesting that incivility manifests across a wide array of communicative acts, including impoliteness. Therein, incivility should be understood as the attempt to delegitimize individual communicators, political actors, and/or democratic institutions. On the operational level, these approaches suggest the existence of multidimensionality of uncivil expressive forms. Frequently identified uncivil communication acts include the use of name-calling or insulting language (e.g., Coe, Kenski, & Rains, 2014; Kenski et al., 2020; Sydnor, 2018), the use of vulgarity and profanity (Coe et al., 2014; Stryker, Conway, & Danielson, 2016), the use of threating language (Massaro & Stryker, 2012; Santana, 2014), racism, xenophobia or other identity-based attacks (here in this paper abbreviated as IBAs) (e.g., Santana, 2014; Theocharis, Barberá, Fazekas, & Popa, 2020), the attempt to undermine faith in democratic systems (e.g., Bentivegna & Rega, 2022; Gervais, 2015; Papacharissi, 2004), and the rhetorical designation of those with opposing views as illegitimate (e.g., Bentivegna & Rega, 2022; Muddiman, 2017; Papacharissi, 2004).

2.1 Incivility on Reddit

In their attempt to develop a taxonomy of incivility specifically for Reddit, Davidson et al. (2020) read approximately 4,000 Reddit comments selected at random across 9,355 subreddits from 2016 to 2019. They identified name-calling, aspersion (or attacks on integrity), disparaging remarks, and general vulgarity. They found that 9.21% of all non-political comments were uncivil, and 14.75% of political posts were uncivil, suggesting that incivility on Reddit was quite widespread.

Not all incivility on Reddit is equal in terms of the consequences it can have on recipients. For instance, while some individuals are sensitive to profanity, many are not. Name calling can temporarily hurt individuals, but other behaviors can have serious consequences for other users. Hmielowski et al. (2014) documents that flaming, or repeatedly insulting an individual or group with the aim of starting a conflict, is quite common. Stevens et al. (2021) found that news content posted to Reddit that discussed sexual assault was often met with "rape culture," or uncivil responses downplaying the severity or legitimacy of sexual assault claims. Taken together, the academic literature paints a picture of abundant incivility on Reddit with varying forms of severity. We ask, to what extent is this true in our sample, and to what extent are the behaviors observed problematic to individuals who receive it?

RQ1: To what extent are Reddit comments and submissions uncivil?

2.2. In-groups and incivility

Subreddits like /r/news, or r/politics, aim to be independent of group memberships (Rathje et al., 2021). As such, they have crosscutting conversations between individuals that vary across political groups. However, many subreddits do have in-group memberships, often drawn along political or ideological lines. The subreddit r/socialism, for instance, boldly states in its guidelines "no liberalism," designating a clear in-group and out-groups. That is, the specification of the subreddit specifies who should *and should not* participate in the conversation. Another, r/alltheleft describes itself as a safe space for all left-minded individuals. It is political in nature, with a clear in-group drawn along these lines. In the context of incivility, if a Reddit user perceives that an in-group exists, in this example the political left, they are more likely to view opposing groups (a.k.a., out-groups) as an obstacle (e.g., an enemy) and become angered towards them. It is common for group members to attack the obstacle (Dillard & Peck, 2001). Thus, there is a motivation for in-groups to collectively target incivility towards out-groups in these subreddits.

Subreddits with in-groups often stoke tensions between out-groups. Extreme right groups have been more notorious for their ability to build collective identity. In their analysis of r/The_Donald, Gaudette et al. (2019) found Reddit's unique voting algorithm facilitated toxic "othering" discourse towards two groups, specifically Muslims and the left. Others have shown that with a clear out-group, redditors have the incentive to use inflammatory language, or low-quality, unnecessary aggressive insults (Hmielowski et al., 2014).

RQ2: To what extent does incivility vary by subreddits with and without in-group designation?

RQ3: Do subreddits with clear in-groups get more engagement than subreddits that do not?

2.3. Reddit content moderation policies

Content moderation has become a partisan issue in the United States, with conservatives accusing popular social media platforms of censorship (Buckley & Schafer, 2021). Social media companies in the United States are not legally liable for the speech or actions of the users on their platforms. They are free to censor expression as they see fit (Carlson & Cousineau, 2020). Even at Facebook, the largest social media platform, content moderation practices have been documented as rushed, ad-hoc, and at times incoherent (Langvardt, 2017). Because the content moderation process is one that often happens out of view of a social media platform's users, there are issues regarding the transparency of how most social media platforms handle policy violations, particularly as it pertains to violence, hate speech, and sexual content.

Moderators on Reddit struggle with managing uncivil content because decisions are often subjective, with two or more sides arguing for the removal or stay of content (Almerekhi et al., 2020). Both sides of the political spectrum have documented some type of dissatisfaction with content moderation on Reddit. Right-leaning users on Reddit desire less moderation, while left-leaning users highlight inconsistencies in how content policies are applied (Shen & Rose, 2019).

One major criticism of Reddit's decisions was that it could not justify why some subreddits were banned while others were maintained. Some Reddit communities are banned for violating content policies. For example, r/The_Donald was was banned for violating reddit-wide platform policies, specifically that it had continuously promoted hate speech. Reddit justified the new site-wide policy on hateful content as

necessary for platform health. It defined hate speech as content that, "encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group based on race, ethnicity, national origin, caste, sexual orientation, transgender status, religion, age, disability, serious medical conditions, or veteran status" (see Worstnerd, 2020 for a review of the policy). Reddit bans are not limited to the political right. r/ChapoTrapHouse, a community for left-leaning users, was also banned for violating subreddit rules around hate speech.

Beyond the hate speech policy, which applies to all subreddits, subreddit moderators largely propose, adopt and enforce their own policies. As such, moderators have a huge influence on which types of content flourish. These policies affect self-censorship and language use in online spaces (Gibson, 2019). For example, if a user is banned for promoting hate speech, other users in that community may self-censor their language to avoid being banned. However, the enforcement of subreddit specific community guidelines remains an open question. How to these policies vary? Do all subreddits enforce their community guidelines with equal rigor? There is evidence that Reddit moderators do not enforce community guidelines with equal rigor for all communities (Gibson, 2019). For example, moderators of the r/The_Donald community were less likely to enforce subreddit rules than moderators of other communities. Yet, little research has been done to assess the degree to which moderators enforce community guidelines in different ways for different communities. The current study seeks to address this research gap by investigating how Reddit moderators enforce community guidelines in different ways for different communities. This examines whether communities with more guidelines exhibit less incivility.

RQ4: Does the complexity of a subreddit's community guidelines correlate to its observed incivility?

2.4. Does uncivil content get engaged with more?

Bystanders can intervene when they observe a violation of a subreddit's community guidelines, but they can also choose to encourage and reward the behavior. These users fundamentally drive the success of the community through their engagements (Kim, 2021). There has been a growing concern that social media platforms, however inadvertently, are promoting uncivil discussion because the content is engaged with (Davidson et al., 2020). Incivility in the comment sections of newspapers has been shown to be infectious in individuals (Shmargad et al., 2022). If a user's incivility is awarded with engagement, such as upvotes and recommendations, commenters tend to take that as an incentive to post more uncivil content. It is possible then that if incivility is popular on Reddit, as the previous section of this literature review suggests, it is also rewarded by users in the form of engagement.

Turning to social media, Wang and Silva (2018) found that when participants observed angry political debates on Facebook, they became more engaged. Another study of Facebook users found engagement was higher when posts were uncivil. More recently, Hansen (2022) collected a one-month sample of Reddit submissions and comments for 71 subreddits across the political spectrum in 2020. Using the same pre-trained machine learning algorithms leveraged in this study, the author assessed the relationship between incivility in a Reddit submission and the number of upvotes that submission got. The author found a positive relationship, suggesting that "toxic incivility" led to more engagement.

In a 2009 analysis of 180 different subreddits, another study assessed the relationship between uncivil behaviors and the amount of user engagement exhibited on subreddits (Mohan et al., 2009). As the authors expected, there was a *negative correlation* between uncivil behaviors and engagement. The researchers found that when the toxicity of a community was stable, the growth of that subreddit flourished. Taken together, results are mixed. We reopen the question of whether uncivil submissions and comments on Reddit will receive more engagement than civil ones.

RQ5: Does uncivil content on Reddit will receive more engagement than civil content?

3. Method

Given our focus on political incivility, we set out to identify the most populous subreddits where this conversation was likely to occur. Redditlist.com enabled us to identify the most active subreddits centered on politics and news. To avoid issues with seasonality, we settled on sampling a 1-year period from June 4th, 2021, to June 4th, 2022.

Next, we needed to set a sampling percentage that would be representative, but still allow us to survey a large collection of Reddit data. Drawing from big data sampling strategies of Twitter data, we adopted a conservative percentage of 20%, one that is extremely likely to correlate with the distribution in which it was drawn from (Morstatter, et al., 2013).

Next, we considered the major data throughput limitations to assess how many messages we could process in the computing time the researchers had to

collect data, which was one week. As section 3.3 outlines, we relied on Google's Perspective application programming interface (API) to measure incivility. Its rate limit, combined with the limits of API we used to fetch reddit data, the pushshift API (Baumgartner et al., 2020), allowed our final analytic to be over 2.7 million messages (127,870 submissions and 2,576,049 comments).

This final sample size allowed us to fetch 20% samples from the top 20, most active political and news subreddits for an entire year. Examination of the extant literature on the computational detection of incivility employed analytic sample sizes that varied from several thousand messages to ~15 million messages (e.g., Almerekhi et al. 2020; Davidson, Sun, & Wojcieszak, 2020; Daxenberger, Ziegele, Gurevych, & Quiring, 2018; Hansen, 2022; Hopp et al., 2019). As such the final analytic sample comports to the median sample size found used in the literature.

3.1. Identifying moderation policies and in-group presence

Two researchers independently reviewed each subreddit's submission guidelines, and the official descriptions for each subreddit to determine the moderation policies and whether in-groups were clearly defined. The two researchers compiled their results and resolved all disagreements, which were limited to varying terminology for types of uncivil behaviors.

For content moderation, the most common trend that emerged was that (1) it was common for a subreddit to explicitly protect a gender or class. We labeled these subreddits as having some sense of aversion to bigotry. There were also (2) general calls to keeping conversations civil, (3) warnings against personally attacking individual users, (4) bans on hate speech, (5) bans on overly violent content, and (6) bans on vulgarity.

In addition, each subreddit was also reviewed for whether the subreddit was created with the intent to cater to a clear group of individuals. Given the political and news nature of this study, all in-group designations were political affiliations, or political ideology. A list of all 20 subreddits and the relevant grouping classifications can be found in Table 1.

Table 1. Moderation policies of each subreddit.

Subreddit	Bigotry	Civility	Personal Attacks	Hate Speech	Violence	Vulgarity	In-Group
alltheleft	1	0	1	1	0	0	1
americanpolitics	1	1	0	1	0	0	0
conservative	1	1	1	0	0	0	1
conspiracy	1	0	1	1	1	0	1
democrats	1	1	1	1	1	0	1
geopolitics	1	1	0	0	0	1	0
liberal	1	0	1	0	1	0	1
libertarian	0	0	1	0	0	0	1
neoliberal	1	1	0	0	1	0	1
news	1	1	0	0	0	0	0
politicalcompassmemes	1	0	0	1	0	0	1
politicaldiscussion	0	1	0	0	0	0	0
politics	0	1	1	1	0	0	0
progressive	1	1	1	1	0	0	1
republican	1	1	1	0	1	0	1
socialism	1	0	0	0	0	0	1
stupidpol	1	0	0	0	0	0	1
ukpolitics	1	1	1	0	1	0	0
uspolitics	1	1	1	1	0	0	0
worldnews	1	1	1	1	0	0	0
Total	17	13	12	9	6	1	12

3.3. The detection of uncivil content

As aforementioned, the Perspective API was used to detect incivility in this study. It was built and refined using hundreds of thousands of human-provided annotations across a wide range of Internet-based user-generated comments. The API returns a continuous probability value (P; theoretical range: 0-1.00) that represents the extent to which a given a text is likely to possess a specified attribute. The algorithm has been regularly used to assess uncivil online commentary (e.g., Hansen, 2022; Hopp et al., 2019; Kim, Guess, Nyhan, & Reifler, 2021; Theocharis et al., 2020), including incivility on Reddit (e.g., Almerekhi et al. 2020; Hansen, 2022; Stevens et al., 2021).

The general toxicity attribute is defined as a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion. In addition, Perspective can detect identity-based attacks (IBAs), such as racism, xenophobia. Moreover, it detects insulting language (i.e., name-calling), profanity, threatening language, profanity, and sexual explicitness. In a recent application that married self-response survey data with social media data from participants, Hopp et al. (2019) found that not only did toxicity attribute detect incivility in social media content as humans do, but scores also generally correlated to the perceptions individuals had *of their own incivility on social media*.

In the present study, we leveraged the typology and data key put forward by Stevens et al. (2021), that found the general "toxicity" algorithm was good at detecting comments that discourage replies. Its "insults" measure detected negative comments towards an opposing person and its "profanity" algorithm generally detected vulgarities and clever derivatives thereof. Its "threat" measure revealed desires to harm an individual or group, and its "identity attack" (here referred to as IBA) algorithm revealed negative identity-based comments.

These findings suggest that the tool is an imperfect, but acceptable means of identifying key interpersonal manifestations of political incivility (e.g., name-calling/insults, vulgarity and profanity, threatening language, racism and xenophobia, the delegitimization of oppositional others).

To externally validate the data to our present data set, two researchers manually and independently (from both one another and from the computer-derived annotations) reviewed a random sample of 2,000 positively flagged comments for these attributes (P > .50) and found that the precision for each of these five algorithms exceeded 70%.

4. Results

To address RQ1, we first examined the mean Perspective-generated probability values for both the comments and submissions data. All attributes across both datasets had mean scores of P < .20 (comments: *Midentity_attack* = 0.07, *Minsult* = 0.16, *Mthreat* = 0.05, *Mprofanity* = 0.11, *Mtoxicity* = 0.18; submissions: *Midentity_attack* = 0.07, *Minsult* = 0.09, *Mthreat* = 0.05, *Mprofanity* = 0.05, *Mtoxicity* = 0.13). Next, we converted the continuous Perspective-generated P values into binary outcome classifications (1 = uncivil message, 0 = civil message). As shown in Tables 2a and 2b, even under a low-confidence threshold (i.e., 1 = P > .50), incivility was somewhat infrequently observed in the data. Using the classification outcomes generated using a moderate (or, "compromise") confidence approach (i.e., 1 = P > .75), the observational range for message-based incivility ranged from 0.01% to 3.85%. Substantively speaking, this suggests that incivility is somewhat rare in Reddit messages. Therein, incivility appears to be more frequently observed in user comments rather than user submissions. Extreme forms of incivility (threat, identity attack, appeared substantially less frequently than impoliteness-style civility violations such as profanity and insulting language.

Table 2a. Incivility prevalence in sampled Reddit comments.

Comments (n = 2,576,049)			
Incivility Attribute	Low Confidence (P > .50)	Moderate Confidence (P > .75)	High Confidence (P > .90)
IBA	1.33%	0.01%	0.00%
Insult	10.00%	1.22%	0.00%
Threat	1.40%	0.01%	0.00%
Profanity	7.70%	1.79%	0.02%
Toxicity	11.19%	3.85%	0.74%

Note. P = Perspective API assigned probability of attribute presence.

Table 2b. Incivility prevalence in sampled Reddit submissions.

Submissions (n = 127,870)			
Incivility Attribute	Low Confidence (P > .50)	Moderate Confidence (P > .75)	High Confidence (P > .90)
IBA	0.84%	0.01%	0.00%
Insult	2.59%	1.22%	0.00%
Threat	0.67%	0.01%	0.00%
Profanity	1.29%	0.03%	0.03%
Toxicity	2.86%	0.74%	0.74%

The second research question was interested in the relationship between incivility and the presence of a dominant in-group within a given subreddit. To empirically assess this question, mean incivility values for in-group and non-in-group subreddits were compared. Given the large sample size, frequentist-based p-values were not considered. Instead, our interpretation focused on effect size estimates (here, Cohen's d). Generally speaking, *d* values between 0 and .10 represent negligible/no effect; values between .10 and .50 are indicative of a small effect; values between .50 and .80 represent a moderate effect, and values greater than .80 describe a strong effect. In the comments data, we observed a trend in which in-group dominance was associated with higher levels of incivility. However, the effect sizes for these differences were all negligible (range: 0.02 - 0.09). In the submissions data, we found that in-group dominance was positively but weakly associated with the use of insults ($d = 0.22$), general toxicity ($d = 0.22$), the use of profanity ($d = 0.16$), and the use of IBAs ($d = 0.12$). See Tables 3a and 3b for a complete report of these analyses.

Table 3a. Relationship between in-group status and subreddit Incivility in sampled comments.

Comments (n = 2,576,049)			
Incivility Attribute	In-Group High	In-Group Low	*d*
IBA	0.08	0.07	0.09
Insult	0.16	0.16	0.03
Threat	0.05	0.05	0.02
Profanity	0.11	0.10	0.08
Toxicity	0.19	0.18	0.07

Note. Cell entries under the high and low in-group headers contain mean values for each group; bolded entry indicates the sole scenario in which low in-group mean values were greater than high in-group mean values.

Table 3b. Relationship between in-group status and subreddit Incivility in sampled submissions.

Submissions (n = 127,870)			
Incivility Attribute	In-Group High	In-Group Low	*d*
IBA	0.07	0.06	0.12
Insult	0.11	0.08	0.22
Threat	**0.04**	**0.05**	**0.06**
Profanity	0.06	0.04	0.16
Toxicity	0.14	0.11	0.22

RQ3 was concerned with the relationship between subreddit in-group status and engagement. Notably, the comments API does not return a valid (at least in our estimation) measure of engagement, so we focused our efforts specifically on the submission data, which contained a measure of the number of comments associated with each user submission (i.e., more comments = higher engagement). Simple comparison of group means indicated that engagement was similar across groups (in-group $M = 26.20$, out-group $M = 26.00$; $d = 0.00$). Notably, however, the in-group category was associated with a number of outlying cases (in-group max = 18,831 comments, non-in-group max = 11,785 comments). The extent to which outliers influenced the results was addressed using a simple robust regression model (e.g., Fox, 1997). The results of this model ($RSE = 4.45$) indicated that submissions posted in in-group dominant settings received, on average, 1.25 more comments than those posted in subreddits not linked to a dominant in-group.

RQ4 was addressed next. To generate a basic measure of moderation complexity and depth, we summed the number of identifiable moderation policies for each subreddit (see Table 1; observed range: 1-5). Spearman rank order coefficients (ρ) were used to assess the relational magnitude between moderation complexity and the presence of the incivility attributes of central interest to this study. Again, given the large sample size, frequentist-based *p* values were not considered. In the comments data, a clear trend was observed wherein moderation complexity was associated with subtle decreases in incivility (ρidentity_attack = -0.04, ρinsult = -0.05, ρthreat = -0.03, ρprofanity = -0.02, ρtoxicity = -0.05). This trend was not, however, apparent in the submissions data. Specifically, in several cases the relationship between moderation complexity and incivility was positive (ρinsult = 0.04, ρthreat = 0.06, ρtoxicityt = 0.06) while in the remaining cases the relationship was essentially zero (ρidentity_insult = -0.01, ρprofanity = 0.00).

Finally, the extent to which incivility was associated with engagement was evaluated (RQ5). Again, given the limitations associated with the comments API, we focused specifically on the submissions data. Basic regression diagnostics indicated moderate to severe amounts of multicollinearity among the incivility attributes (*VIF* range: 1.39 - 9.51; mean *VIF* = 4.32); as such, the relationships between the incivility attributes and comment frequency were examined individually. A series of Spearman rank-order correlations indicated weak but positive associations between several of the incivility measures and the number of submission-associated comments (ρidentity_attack = 0.04, ρinsult = 0.07, ρthreat = 0.02, ρprofanity = 0.00, ρtoxicity = 0.05). Given the presence of outlying cases in the data, these relationships were re-examined using a series of 5 discrete robust regression models. Unlike RQ3, addressing the impact of extreme cases via robust

regression modeling had a generally negligible impact on the relational magnitude between incivility and comment generation frequency (identity attack: *RSE* = 3.45, *b* = 0.26, β = 0.00; insult: *RSE* = 3.34, *b* = 0.77, β = 0.00; threat: *RSE* = 3.47, *b* = 0.22, β = 0.00; profanity: *RSE* = 3.48, *b* = 0.05, β = 0.00; toxicity: *RSE* = 3.36, *b* = 0.56, β = 0.00).

5. Discussion

The goal of this study was to better understand patterns and correlates of incivility on the most active subreddits centered on politics and news. Broadly speaking, and somewhat against what the literature has documented, our results suggest that incivility is not frequently observed on these subreddits.

It is important to note that not all forms of incivility have the same types of harm on individuals who are the recipients of it, and on Reddit, when incivility does occur, it tends to take on forms that have less harm in the individuals that consume it. Insults, profanity, and general toxicity were more common than more targeted forms such as threats or identity-based attacks. These findings conform with prior work (e.g., Hopp et al., 2019) which have similarly shown that incivility is not extensively apparent on Facebook and Twitter. At the same time, we urge caution when interpreting this observation. Any retrospective analysis of moderated trace data will naturally be unable to retrieve instances in which uncivil or otherwise noxious commentary was removed from the platform. As such, we are studying the moderated incivility that remains on Reddit. Moreover, the viral affordances of social and digital media and the human negativity biases mean that a small amount of uncivil commentary can play an outsized role in user's attentional patterns, and, as such, play a disproportionately strong role in establishing user interpretations of platform culture.

Our data also shows under certain conditions, incivility can be associated with or drive user engagement. Specifically, we observed subtle, albeit systematically positive, associations between the presence of incivility and user engagement with Reddit submissions. Obviously, the lack of strong bivariate associations observed in this study limit our ability to make strong claims about the strength of the incivility-engagement relationship. One potentially important observation pertains to the notion that different types of incivility are disproportionately associated with engagement outcomes. Insults were much more robustly associated with comment generation than comparatively more severe forms of incivility such as interpersonal threats or mundane forms of uncivil behavior such as profanity.

We urge scholars to develop more nuanced hypotheses around studying uncivil content and its relationship to platform behaviors. Not all behaviors appear to be equal. Many have noted that insults are common on Reddit. For instance, this was a cited reason for /r/chapotraphouse being banned (Shen, 2021). To better understand these relationships, we encourage scholars to focus on the behaviors with positive associations identified in this paper.

Incivility is also not uniform across user-generated content submissions. In our analysis, we see that incivility appears to be more frequently observed in user comments rather than in user submissions. Submissions are prominently displayed on Reddit, for the entire subreddit to see, whereas comments live nested therein. It stands to reason that moderators prioritize the review of submissions for these reasons.

Indeed, the more complex the moderation policies were in a subreddit's community guidelines, the less incivility observed. This provides a direct incentive for social media platforms to develop and publish exhaustive moderation policies. For instance, at the time of this paper's writing, Twitter has several policies that touch on civility. Coordinated efforts to harm individuals are not allowed.[2] Violence is prohibited and is not to be celebrated. Violent groups are not allowed to exist or be promoted on Twitter,[3] and users may not, "engage in the targeted harassment of someone, or incite other people to do so."[4] Subreddits are, in a way, like mini social media platforms, at least as it pertains to the standards they put in place for their community. Reddit is an excellent case study of how the rigor of policies can change behavior. Given our findings here, we suggest that should Twitter repeal their policies in the name of "free speech," incivility in the replies to tweets would likely increase. If Twitter wants to maintain its levels of incivility, it should consider more policies, not less.

Moreover, when a subreddit is created with the clear intention to cater to a specific group — here a here a political alignment such as /r/socialism — more insults, toxicity, profanity and IBAs result. In-group presence results in a *mild increase* in incivility across the board, from less serious behaviors like vulgarity to more serious identity-based hate. As previous literature suggests, this is likely because these groups

[2] Policy available here: https://help.twitter.com/en/rules-and-policies/violent-threats-glorification

[3] Policy available here: https://help.twitter.com/en/rules-and-policies/violent-groups

[4] Policy available here: https://help.twitter.com/en/rules-and-policies/abusive-behavior

have a clear and common political enemy, use the subreddit to exhibit anger and disgust toward their political opponent, and attack it through all means necessary. However, we must temper our claims because these relationships were true of submissions in the subreddit but not of comments. But again, given that submissions get far more impressions than nested comments, the exposure to this behavior remains quite large and problematic.

As other scholars have suggested, there does appear to be a perverse reason to allow incivility to exist on social media platforms. Incivility in subreddit submissions do have slightly higher engagement. However, while these numbers are higher, they are not so much so that it would ruin the health of a subreddit to remove them. As Mohan et al. (2009) suggests, the health of a subreddit is likely due to toxicity being relatively rare. Much of Reddit's success can be attributed to the fact that it has made efforts to moderate incivility, and ban subreddits that violate its site-wide, and subreddit specific policies.

This study is, of course, associated with some limitations. The Perspective API, while somewhat precise, certainly suffers from its ability to recall all bad behaviors. Moreover, it does not allow us to assess certain types of incivility, such as rhetorical attempts to undermine faith in democratic institutions. Moreover, our data was drawn from 20 of the most popular politics and news oriented subreddits. It is not, representative of all commentary on Reddit as a whole. Reddit is a decentralized platform that employs a hybrid community moderation strategy. These unique attributes mean that the findings observed here may necessarily translate to other social media contexts.

In conclusion, news and politics subreddits were less uncivil than expected. The configuration of a subreddit — its alignment to a political side and its content moderation policies — do result in varied amounts of incivility, but more importantly, the platform's approach to incivility moderation appears to be effective, perhaps more so than the academic literature has written about, to date.

6. References

Almerekhi, H., Jansen, B.J., & Kwak, H. (2020). Investigating toxicity across multiple Reddit communities, users, and moderators. *WWW '20: Companion Proceedings of the Web Conference 2020*, 294–298. https://doi.org/10.1145/3366424.3382091

Bentivegna, S., & Rega, R. (2022). Searching for the dimensions of today's political incivility. *Social Media + Society, 8*(3), 1-12.

Buckley, N., & Schafer, J. S. (2022). "Censorship-free" platforms: Evaluating content moderation policies and practices of alternative social media. *For(e)dialogue, 4*(1). https://doi.org/10.21428/e3990ae6.483f18da

Caplan, R. (2018). *Content or context moderation? Artisanal, community-reliant, and industrial approaches*. https://apo.org.au/node/203666

Carlson, C. R., & Cousineau, L. S. (2020). Are you sure you want to view this community? Exploring the ethics of Reddit's quarantine practice. *Journal of Media Ethics, 35*(4), 202–213. https://doi.org/10.1080/23736992.2020.1819285

Coe, K., Kenski, K., Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679.

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 95–101. https://doi.org/10.18653/v1/2020.alw-1.12

Daxenberger, J., Zigele, M., Gurevych, I., & Quiring, O. (2018). Automatically discussing incivility in online discussion of news media. *Proceedings of the 2018 IEEE 14th International Conference on e-Science*.

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society, 23*(12), 3491–3508. https://doi.org/10.1177/1461444820958123

Gervais, B. T. (2016). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research, 29*, 384–405.

Gibson, A. (2019). Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media + Society, 5*(1). https://doi.org/10.1177/2056305119832588

Google (2018). Example data. https://github.com/conversationai/perspectiveapi/blob/master/example_data/perspective_wikipedia_2k_score_sample_20180829.csv

Hansen, R. W. (2022). You've never been welcome here: Exploring the relationship between exclusivity and incivility in online forums. *Journal of Information Technology & Politics*. Advance online publication. https://doi.org/10.1080/19331681.2022.2069180

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of computer-mediated communication*, *18*(2), 154-174.

Hmielowski, J. D., Hutchens, M. J., & Cicchirillo, V. J. (2014). Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming. *Information, Communication, & Society, 17*(10), 1196–1211. https://doi.org/10.1080/1369118X.2014.899609

Hopp, T. (2019). A network analysis of incivility dimensions. *Communication and the Public, 4(*3), 204-223.

Hopp, T., Vargo, C. J., Dixon, L., & Thain, N. (2019). Correlating self-report and trace data measures of incivility: A proof of concept. *Social Science Computer Review, 38*(5), 584–599. https://doi.org/10.1177/0894439318814241

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-30.

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research, 47*(6), 795-814.

Kim, Y. (2021). Understanding the bystander audience in online incivility encounters: Conceptual issues and future research questions. *Proceedings of the 54th Hawaii International Conference on System Sciences, 2021*, 2934–2943. https://doi.org/10.24251/HICSS.2021.357

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

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal, 106*(5), 1353–1388. https://doi.org/10.2139/ssrn.3024739

Levin, B. (2022, April 26). *A reminder of just some of the terrible things Elon Musk has said and done*. Vanity Fair. https://www.vanityfair.com/news/2022/04/elon-musk-twitter-terrible-things-hes-said-and-done

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, 54, 375–441. Retrieved from http://nicd.arizona.edu/research-article/freedom-speech-liberaldemocracy

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The impact of toxic language on the health of reddit communities. In M. Mouhoub & P. Langlais (Eds.), *Lecture notes in computer science: Vol. 10233. Advances in artificial intelligence* (pp. 51-56). Springer. https://doi.org/10.1007/978-3-319-57351-9_6

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 400-408).

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, 11, 3182–3202.

Ovide (2022, April 26). *Elon Musk and the gray of 'free speech.'* New York Times. https://www.nytimes.com/2022/04/26/technology/elon-musk-free-speech.html

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283.

Rathje, S., Van Bavel, J.J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, USA*, *118*(26), e2024292188. https://doi.org/10.1073/pnas.2024292118

Reddit Inc. (2017, April 17). *Moderator guidelines for healthy communities*. Reddit Inc. https://www.redditinc.com/policies/moderator-guidelines-for-healthy-communities

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research, 49*(3), 399-425.

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice, 8*(1), 18–33.

Shen, Q. (2021). *Evaluating and Recontextualizing the Social Impacts of Moderating Online Discussions* (Doctoral dissertation, University of Michigan).

Shen, Q., & Rose, C. (2019, August). The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Third Workshop on Abusive Language Online* (pp. 58-69). https://doi.org/10.18653/v1/W19-3507

Shmargad, Y., Coe, K., Kenski, K., & Rains, S.A. (2022). Social norms and the dynamics of online incivility. *Social Science Computer Review, 40*(3), 717–735. https://doi.org/10.1177/0894439320985527

Sydnor, E. (2018). Platforms for incivility: Examining perceptions across media. *Political Communication, 35*(1): 97-116.

Stevens, H., Acic, I., & Taylor, L. D. (2021). Uncivil reactions to sexual assault online: Linguistic features of news reports predict discourse incivility. *Cyberpsychology, Behavior, and Social Networking*, *24*(12), 815–821. https://doi.org/10.1089/cyber.2021.0075

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). *Communication Monographs, 83*(4), 535-556.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open, 10*(2).

Vogels, E. A., Perrin, A., & Anderson, M. (2020, August 19). *Most Americans think social media sites censor political viewpoints*. Pew Research Center. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/

Wang, M. Y., & Silva, D. E. (2018). A slap or a jab: An experiment on viewing uncivil political discussions on Facebook. *Computers in Human Behavior, 81*, 73–83. https://doi.org/10.1016/j.chb.2017.11.041

Worstnerd. (2020, August 20). *Understanding hate on Reddit, and the impact of our new policy* [Online forum post]. Reddit. https://www.reddit.com/r/redditsecurity/comments/idclo1/understanding_hate_on_reddit_and_the_impact_of/