Incivility in Reddit's Top Political and News Subreddits: Prevalence, Moderation, and Engagement

Abstract: *We examine incivility across twenty high-traffic political and news subreddits to test how platform governance and social identity cues relate to on-platform discourse and participation. Building on theories of democratic communication and incivility, platform affordances and moderation, and uses-and-gratifications/network externalities, we specify how decentralized community rules and explicit in-group orientations could shape both the prevalence of uncivil language and patterns of engagement. We analyze a year-long, random sample of submissions and comments scored with established computational measures of incivility, and we link these scores to subreddit-level rule regimes and identity signaling. By distinguishing interpersonal impoliteness from democratic norm violations and by evaluating moderation complexity at the community level, this work clarifies when and how community governance relates to discourse quality and participation dynamics on Reddit. Findings inform ongoing debates about the efficacy of hybrid, human-centered moderation and the role of explicit identity norms in large online communities.*

Researchers, policy makers, and members of the public have long debated whether and how social media platforms should intervene to maintain civil discourse. Communication scholarship defines incivility as discourse that violates norms of respect and rational argument (e.g., Hopp, 2019), while media affordance theory suggests that moderation policies and community guidelines influence the prevalence of uncivil language (e.g., Oz et. al, 2017). Platform-use theories (e.g., uses-and-gratifications; network externalities) imply that design choices and community norms condition not just what people say but how much others respond. If uncivil language systematically attracts more replies, moderators face perverse incentives: removing uncivil content could reduce visible activity, while tolerating it may erode discourse quality. Reddit's volunteer-driven, rule-based governance is an informative contrast case because community rules—and their complexity—vary across subreddits. We therefore examine whether incivility is associated with higher engagement and whether that association is attenuated in communities with clearer rules or explicit in-group orientations.

In public opinion, moderation and censorship are often viewed as intertwined. As a general rule, social media platforms moderate accounts to reduce hate speech, trolling, personal attacks, and other types of incivility, a strategy that appears, in most cases, to be effective (Jhaver et al., 2021). That being said, over the past several years, large social media platforms such as Facebook, Instagram, and Twitter/X have, for a variety of commercial, political, and social reasons, subjected their moderation policies to extensive changes (e.g., Milmo, 2025), leading some to remark that newly these instituted policies of centralized moderation do little to support free speech, expose vulnerable users to potential harms, and are unduly tethered to political and financial concerns (e.g., Electronic Frontier Foundation, 2025). To that end, Reddit's decentralized moderation regime makes it a distinctive case for studying platform governance.

Unlike algorithmically curated networks such as Facebook or Twitter, where recommendation algorithms amplify content that generates high engagement and thereby privilege highly emotional and uncivil posts, Reddit delegates most moderation decisions to volunteer community moderators. Its human-centered moderation model allows subreddits to set bespoke rules and enforce them with human judgment, rather than rely solely on automated ranking and recommendation systems. This distinction aligns with scholarship on platform affordances and incivility, which argues that the technical architecture and governance structures of online platforms shape the communicative behaviors they afford (e.g., Papacharissi, 2004; Sydnor, 2018). By focusing on Reddit, our study thus examines incivility in a context where moderation practices are decentralized and community-driven, providing a contrast to algorithmically curated environments.

Reddit's content moderation is, as mentioned above, characteristically 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. However, many have noted that these individuals commonly seek to remove illegal content, not posts relating to specific community guidelines for a subreddit (Gibson, 2019). Therefore, subreddits rely heavily on volunteers, known as moderators, who make guidelines for their subreddits that 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 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, incivility on Reddit persists. Early work identified persistent uncivil behaviors across subreddits (Davidson et al., 2020; Hansen, 2022; Stevens et al., 2021). Building on this literature, recent studies have further unpacked the dynamics of Reddit incivility. For instance, Salgado et al. (2025) compare echo-chamber and cross-ideological subreddits and show that incivility is more prevalent in ideologically homogeneous communities. Gao et al. (2024) introduce a fine-grained taxonomy and model forty million Reddit comments, finding that discussions that begin uncivil tend to become more uncivil and that different incivility categories have distinct effects on engagement. Hanley and Durumeric (2023) analyse posts linking to unreliable news sources and show that comments on these posts are 32 percent more likely to be toxic and that toxic subreddits attract more engagement with low-quality information. Mamakos and Finkel (2023) examine millions of comments across partisan and non-partisan subreddits and find that highly engaged partisans are consistently toxic across contexts, suggesting that incivility may reflect individual traits rather than situational factors. Narimanzadeh et al. (2025) show that incivility and contentiousness triggered by COVID-19 discussions spill over into climate change engagement across platforms.

Although our empirical focus is Reddit, the theoretical mechanisms we invoke generalize across social platforms: governance arrangements and community norms condition both the prevalence of incivility and audience responses to it. Work on platform affordances shows that rule clarity and enforcement shape what communicative acts are seen as acceptable (Gibson, 2019; Papacharissi, 2004), and that when uncivil cues do appear, they can alter participation dynamics by encouraging replies or discouraging deliberation (Hansen, 2022; Shmargad et al.,

2022; Theocharis et al., 2020). Because Reddit's moderation is decentralized and largely volunteer-driven, local rules and identity signals should be especially consequential for whether incivility surfaces and whether it is rewarded with engagement (Gibson, 2019; Hopp, 2019). We therefore integrate two complementary perspectives to guide expectations in this setting: (a) a *platform affordances and governance* view, which links community rules to observable rates of incivility; and (b) an *audience response* view, which links uncivil cues to comment volume and conversational persistence (Hansen, 2022; Shmargad et al., 2022; Theocharis et al., 2020).

Studies of mobile social media highlight the role of platform affordances in shaping continued use (Pang et al. 2024). For example, research on WeChat finds that network externalities, such as the size of a user's network and the availability of complementary services, increase hedonic, social and utilitarian gratifications, which in turn influence attitudes toward the platform and sustained usage. Likewise, analyses of mobile short video apps show that social connectivity and system interactivity enhance perceived benefits and continuance intentions (Pang & Ruan, 2024). Service quality has been pointed to as via a survey of WeChat users that suggests that responsiveness and reliability strengthen identification, belonging and satisfaction, and that emotional attachment fosters continued service attachment.

Alongside the platform-governance perspective outlined above, a second line of scholarship centers on user psychology—especially cognitive overload and social norms—and its consequences for continued participation. In online political talk, exposure to incivility and norm violations can shift perceptions of what is acceptable, sometimes amplifying hostile replies and at other times discouraging entry into a thread (Hopp, 2019; Shmargad et al., 2022). This perspective suggests that even when rules are clear, audience-level processes—fatigue, perceived hostility, or normative pressure—shape whether incivility is produced and whether others engage

with it (Coe et al., 2014; Theocharis et al., 2020). We draw on this work to anticipate when uncivil cues will correlate positively with engagement and when they will not.

Cognitive overload—both informational and communicative—can deplete users and lead to social network exhaustion, privacy invasion and discontinuance intentions. Studies among university students further reveal that depressive mood and self-disclosure contribute to information and social overload; these strains encourage problematic mobile app use and are associated with declines in academic performance. Other work links mobile app addiction, privacy concerns and cognitive overload to perceived technostress, which reduces subjective well-being and academic expectancy and mediates the effects of addiction and privacy concerns. Together, these findings underscore that both platform affordances and user psychology shape engagement and discontinuance and are influenced by echo-chamber dynamics, information reliability, individual traits, and cross-domain spillovers. To this end, this paper explores the relationship between the incivility of a subreddit and 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 whether they identify as having a specific 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. We use the term “out-group” to refer to individuals or communities that do not align with the subreddit’s explicit in-group and are often viewed as outsiders by subreddit members.

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 hand, 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 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 the civility of a subreddit may hinge on whether it is objectively trying to advance a group or is agnostic to group memberships. (see Vallone et al., 1985 for hostile-media perceptions)

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.

Incivility

The definition of political incivility is subject to contest (e.g., Herbst, 2010; Hopp 2019). Papacharissi (2004) has argued that incivility should be specifically understood 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 et al., 2020; Muddiman, 2017; Rossini, 2022; Sydnor, 2018) has arrived at a related but slightly different conclusion, suggesting that incivility is fundamentally tied to a

motivation to exclude disconsonant voices from the public sphere, and can manifest 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 that incivility is a multi-dimensional construct (e.g., Hopp, 2019). Frequently identified uncivil communication acts include the use of name-calling or insulting language (e.g., Coe et al., 2014; Kenski et al., 2020; Sydnor, 2018), the use of vulgarity and profanity (Coe et al., 2014; Stryker et al., 2016), the use of threatening language (Massaro & Stryker, 2012; Santana, 2014), racism, xenophobia or other identity-based attacks (abbreviated in this paper as IBAs; e.g., Santana, 2014; Theocharis et al., 2020), the attempt to undermine faith in democratic systems (e.g., Bentivegna & Rega, 2022; Gervais, 2016; Papacharissi, 2004), and the rhetorical designation of those with opposing views as illegitimate (e.g., Bentivegna & Rega, 2022; Muddiman, 2017; Papacharissi, 2004). Scholars additionally distinguish between interpersonal impoliteness—rude or coarse language that reflects private disagreements—and democratic norm violations that reject principles of inclusion and equality. Interpersonal impoliteness may hurt feelings but does not necessarily undermine deliberation, whereas democratic norm violations delegitimize opponents or institutions and threaten democratic discourse. Our operationalization therefore focuses on expressive acts that transgress democratic norms while acknowledging that impolite exchanges can contribute to perceptions of incivility or, in some cases, serve as the mechanism by which incivility is pursued.

Incivility on Reddit

In their attempt to develop a taxonomy of incivility specifically for Reddit, Davidson et al. (2020) analyzed 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. Some forms, such as name-calling or profanity, may be perceived as less harmful than others like identity-based attacks or threats, underscoring the importance of distinguishing between interpersonal impoliteness and more serious norm violations.

RQ1: What proportion of Reddit posts contain at least one of the five Perspective-coded incivility attributes (identity attack, insult, threat, profanity, toxicity)?

In-Groups and Incivility

General use subreddits like r/news or r/politics aim to be independent of group memberships (Rathje et al., 2021). As such, they have cross-cutting 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 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. We use the term 'out-group' to refer to individuals or communities that do not align with the subreddit's explicit ideological identity and are often viewed as outsiders by subreddit members.

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. (2021) found that 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: Across the sampled subreddits, does incivility vary between subreddits with and without in-group designations?

RQ3: Across the sampled subreddits, do communities with clear in-groups receive more engagement than those without clear in-groups?

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). As it currently stands, 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 parameterize on-platform expression as they see fit (Carlson & Cousineau, 2020). Even on Facebook, the largest social media platform globally, 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 the 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 dissatisfaction with Reddit's content moderation. 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 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 greatly influence what types of content is allowed to flourish. These policies, the literature suggests, affect self-censorship and language use in online spaces (e.g., 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 do 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. It examines whether communities with more guidelines exhibit less incivility.

RQ4: Does the number of stated moderation rules correlate with observed incivility?

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 rewarded with engagement, such as upvotes and recommendations, commenters tend to take that as an incentive to post more uncivil content. It is then possible 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.

Beyond network externalities and uses-and-gratifications evidence in WeChat (Pang, Qiao, Li, & Wang, 2024), research on mobile short-video platforms demonstrates that social connectivity and system interactivity increase users' perceived benefits and, through them, continuance intention (Pang & Ruan, 2024). Complementary work in mobile social media services shows that service quality strengthens identification, belongingness, and satisfaction (Pang & Zhang, 2024a), while functional, psychosocial, and hedonic benefits raise cumulative satisfaction and eWOM engagement (Pang & Zhang, 2024b). On the risk side, cyberbullying perpetration and communication overload elevate app-switching intention (Pang, Zhao, & Yang, 2025); depressive mood and self-disclosure increase overload and problematic app use, which in turn predicts declines in educational attainment (Pang, 2024); and social comparison–driven problematic WeChat use undermines academic achievement via attention-related mechanisms (Pang & Hu, 2025). Together, these findings situate engagement and governance dynamics on Reddit within a broader social-media ecology where platform affordances, benefits, and stressors jointly shape participation and outcomes.

In a 2009 analysis of 180 different subreddits, another study assessed the relationship between uncivil behaviors and user engagement (Mohan et al., 2017). As the authors expected,

there was a *negative correlation* between uncivil behaviors and engagement. The researchers found that when a community's toxicity 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 receive more engagement than civil content?

Method

This study assessed commentary on the 20 most active US-oriented political and news subreddits. The rationale for focusing specifically on politics and news subreddits was based on the notion that incivility, at least as considered in this project, is an inherently political phenomenon insofar as it speaks to who does and does not have access to the public sphere (e.g., Hopp, 2019; Muddiman, 2017; Papacharissi, 2004). While it may well be the case that other topical subreddits (e.g., r/NFL, r/Funny, r/gaming) feature toxic, abusive, or impolite discussion, the non-political nature of these spaces offers limited utility regarding the democratic implications of social media moderation. To avoid issues with seasonality, we settled on sampling a 1-year period from June 4th, 2021, to June 4th, 2022. In addition to addressing seasonality concerns, the use of a single-year frame allowed for collection of a large corpus of data while simultaneously capturing a contemporaneous set of moderation practices. Reddit affords two distinct loci of expression: top-level submissions, which broadcast to the full subreddit and are more visible to non-participants, and nested comments, which unfold within threads and often reflect conversational back-and-forth. Because visibility, audience scope, and moderation practices can differ across these loci, we analyze submissions and comments separately throughout (Gibson, 2019).

To generate a manageable, but representative, sample, we used the Pushshift API's random-seed functionality to randomly select 20% of all submissions and comments from each of the 20 subreddits. Because every submission or comment posted between June 4 2021 and June 4 2022 had an equal probability of selection and because the sampling spanned the entire year, the resulting sample is representative of the full year's population (Morstatter et al., 2013).

Next, we considered the major data throughput limitations to assess how many submissions and comments we could process in the computing time the researchers had to collect data, which was one week. As described below, we relied on Google's Perspective application programming interface (API) to measure incivilityUsing Pushshift's random-seed functionality, we drew a 20% random sample of all submissions and comments posted in each of the 20 subreddits between June 4, 2021 and June 4, 2022. This procedure yielded 127,870 submissions and 2,576,049 comments (≈2.70 million items in total), which correspond to roughly one-fifth of all posts and comments created in those communities during the study window. Minor deviations from exactly 20% reflect API-level rate limits and the one-week data-collection window; because selection was random and spanned the entire year, the resulting corpus remains representative of activity across the period (Baumgartner et al., 2020; Morstatter et al., 2013).

Because several analyses involved regression models with highly skewed engagement measures, we estimated robust regression models to mitigate the influence of outliers. Robust regression minimizes the impact of extreme cases by applying M-estimation with a Huber weighting function, which downweights observations with large residuals. These models were fit using the rlm function from R's MASS package, which implements iteratively reweighted least squares to obtain robust estimates.

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 keep 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 reviewed to determine whether it was created 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 ideologies. A list of all 20 subreddits and the relevant grouping classifications can be found in Table 1. Subreddits were coded as "in-group present" when their description or rules explicitly encouraged participation from a defined in-group (such as a political ideology) and discouraged participation by outsiders; otherwise, they were coded as "in-group absent."

Table 1

Moderation policies of each subreddit

Subreddit	Bigotry	Civility	Personal Attacks	Hate Speech	Violence	Vulgarity	In-Group
alltheleft	✓	–	✓	✓	–	–	✓

americanpolitics	✓	✓	–	✓	–	–	–
conservative	✓	✓	✓	–	–	–	✓
conspiracy	✓	–	✓	✓	✓	–	✓
democrats	✓	✓	✓	✓	✓	–	✓
geopolitics	✓	✓	–	–	–	✓	–
liberal	✓	–	✓	–	✓	–	✓
libertarian	–	–	✓	–	–	–	✓
neoliberal	✓	✓	–	–	✓	–	✓
news	✓	✓	–	–	–	–	–
politicalcompassmemes	✓	–	–	✓	–	–	✓
politicaldiscussion	–	✓	–	–	–	–	–
politics	–	✓	✓	✓	–	–	–
progressive	✓	✓	✓	✓	–	–	✓
republican	✓	✓	✓	–	✓	–	✓
socialism	✓	–	–	–	–	–	✓
stupidpol	✓	–	–	–	–	–	✓
ukpolitics	✓	✓	✓	–	✓	–	–

uspolitics	✓	✓	✓	✓	–	–	–
worldnews	✓	✓	✓	✓	–	–	–
Total	17	13	12	9	6	1	12

The Detection Of Uncivil Content

As briefly mentioned above, the Perspective API (https://perspectiveapi.com/) was used to approximate on-platform incivility. 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 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 et al., 2021; Theocharis et al., 2020), including incivility on Reddit (e.g., Almerekhi et al. 2020; Hansen, 2022; Stevens et al., 2021).

The Perspective API version employed in this study allowed for the identification of a variety of discursive attributes. This study focused on five particular attributes that have been shown as especially indicative of political incivility (e.g., Hopp et al., 2019): *identity-based attack*s (IBA; i.e., racism, xenophobia, homophobia), *insulting language* (i.e., directed name-calling), *threatening language* (i.e., directed use of menacing or intimidating language), *profanity* (i.e., undirected vulgar language), and general *toxicity* (i.e., rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion). 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. Stevens et al. (2021) found that the general "toxicity" algorithm effectively detected comments that discourage replies. Its "insults" measure detected negative comments toward an opposing person, and its "profanity" algorithm generally detected vulgarities and clever derivatives. 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. Taken as a whole, prior work suggests that the Perspective 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, and 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 (i.e., $P > .50$) submissions and comments and found that the precision for each of these five incivility detecting algorithms (i.e., IBA, insult, threat, profanity, general toxicity) exceeded 70%.

Results

To address RQ1, we first examined the averaged 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 (i.e., uncivil comment or submission, civil comment or submission) for both submissions and comments. As shown in Tables 2a and 2b, even under a fairly low-confidence threshold (i.e., incivility determined at $P >$

.50), incivility was observed somewhat infrequently in the data. Using the classification outcomes generated using a moderate (or "compromise") confidence approach (i.e., incivility determined at $P > .75$), the range for comments-based incivility was between 0.01% and 3.85% and for the submissions data the range was between 0.01% and 1.22%. Substantively speaking, this suggests that incivility is somewhat rare in both Reddit comments and submissions . Therein, incivility appears to be more frequently observed in user comments rather than in user submissions. Extreme forms of incivility (e.g., threats and identity attacks) appeared substantially less frequently than impoliteness-style civility violations such as profanity and insults.

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%

Notes: *P* denotes the Perspective API–assigned probability of attribute presence; low confidence threshold = *P* > .50; moderate confidence threshold = *P* > .75; high confidence threshold = *P* > .90; IBA = identity-based attack

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%

Notes: *P* denotes the Perspective API–assigned probability of attribute presence; low confidence threshold = *P* > .50; moderate confidence threshold = *P* > .75; high confidence threshold = *P* > .90; IBA = identity-based attack

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 emanating from subreddits characterized by a dominant in-group (n = 12) were compared to the mean incivility values emanating from subreddits not characterized by a dominant in-group (n = 8). Given the substantive sample size, frequentist-based critical values were not informative and therefore 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: d = 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 present	In-group absent	*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 In-group present and In-group absent headers contain mean values for each group

Table 3b

Relationship Between In-Group Status and Subreddit Incivility in Sampled Submissions

Submissions (n = 127,870)			
Incivility Attribute	In-group present	In-group absent	*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

Note. Cell entries under the In-group present and In-group absent headers contain mean values for each group

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 submissions data, which contained a measure of the number of comments associated with each user submission (i.e., more comments = higher engagement). A simple comparison of group means indicated that engagement was similar across groups (in-group M = 26.20 comments; out-group M = 26.00 comments; d = 0.00). Notably, however, the in-group category was associated with several outlying cases (in-group max = 18,831 comments, non-in-group max = 11,785 comments). The extent to which outliers influenced the results was assessed using a simple robust regression model (e.g., Fox, 1997) in which in/out-group status was dichotomized and set as the predictor variable. 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.[1]

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. In the comments data, a clear trend was observed wherein moderation complexity was associated with subtle decreases in incivility ($\rho_{identity\_attack}$ = -0.04, ρ_{insult} = -0.05, ρ_{threat} = -0.03, $\rho_{profanity}$ = -0.02, $\rho_{toxicity}$ = -0.05). This trend was not, however, apparent in the submissions data. Specifically, in several cases, the relationship between moderation

[1] Robust regression minimizes the potential influence of outlying values by re-weighting extreme observations. In the current case, a robust regression model employing Huber weighting was employed via R's MASS package. Coefficients associated with robust regression models are analogous to those produced via OLS regression.

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_attack = -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; insult: *RSE* = 3.34, *b* = 0.77; threat: *RSE* = 3.47, *b* = 0.22; profanity: *RSE* = 3.48, *b* = 0.05; toxicity: *RSE* = 3.36, *b* = 0.56).

Discussion

The present analysis of 2.7 million Reddit submissions and comments advances scholarship on online incivility by showing that, within the twenty largest news and politics subreddits, uncivil content is the exception rather than the rule, and that volunteer moderation paired with clear, community-level policies appears to keep discussions largely civil.

First, the prevalence estimates challenge the view that political talk on social media is uniformly toxic. Even with a deliberately lenient classification threshold, fewer than one in ten comments and only about three in one hundred submissions contained insults, profanity, or generalized toxicity; severe forms such as threats and identity-based attacks were vanishingly rare. These rates are markedly lower than those reported by earlier work that either sampled more contentious electoral periods (Davidson et al., 2020) or focused on single partisan communities (Gaudette et al., 2021). The finding aligns with Hopp et al.'s (2019) claim that perceptions of pervasiveness can outstrip actual incidence once moderation removes the most egregious material. It also underscores the need to distinguish between raw exposure to incivility and the curated traces that remain after human review.

Second, the modest yet consistent increase in insults, profanity, and identity attacks within subreddits that signal an explicit in-group supports social-identity accounts of antagonism. When community membership is framed around ideological boundaries, users may experience normative permission to target perceived outsiders, thereby elevating low-grade hostility while stopping short of overtly sanctionable hate. That the effect is stronger in top-level submissions than in nested comments suggests strategic signaling to the broader group rather than spontaneous tit-for-tat escalation. At the same time, the weak correlations indicate that in-group status is neither a necessary nor a sufficient condition for incivility, moderating claims that echo-chamber design alone produces hostile climates.

Third, rule complexity shows an inverse relationship with comment-level incivility but little association with submission-level behavior, implying that granular guidelines may help moderators police the high-volume back-and-forth of comment threads where most incivility occurs. This pattern comports with platform-governance literature emphasizing that transparent procedural rules help volunteers act consistently (Gibson, 2019). It further suggests that, absent algorithmic ranking, the human labor of moderation can attenuate the engagement incentives that typically reward provocative content.

Relatedly, the engagement analyses reveal only minor rewards for uncivil submissions, and none for profanity alone. The strongest positive association emerges for insults, echoing findings that mild personal derogation is sufficiently arousing to elicit replies (Hansen, 2022) yet not so extreme as to trigger removal. These data complicate alarmist narratives that platforms necessarily profit from radical hostility; on Reddit, the commercial and reputational costs of unchecked toxicity may outweigh the incremental boost in comment counts.

Methodologically, the study demonstrates that combining a year-long random sample with the Perspective API and robust regression yields stable, interpretable estimates even in the presence of skew and outliers. Precision-validated automated coding supports scalable monitoring across communities, though the reliance on moderated traces and probability thresholds highlights continuing trade-offs between recall and false alarms. Future work could integrate deleted or quarantined content and apply domain-adapted language models to capture nuanced democratic norm violations beyond interpersonal slurs.

Practically, the results endorse a hybrid moderation regime in which volunteer moderators articulate concise but layered rules, focus efforts on comment streams, and treat mild incivility differently from threats or hate. Platforms considering governance reforms might prioritize tooling that surfaces potential rule breaches for human review and that scaffolds consistent enforcement rather than defaulting to fully automated bans or laissez-faire approaches.

Several limitations temper the conclusions. The observational design precludes causal inference; the Perspective API cannot capture rhetorical delegitimization or sarcasm; and the sample, though large, is restricted to English-language, high-traffic subreddits during a non-electoral year. Replication across election cycles, languages, and algorithmically ranked platforms such as Facebook would test generalizability.

In sum, the evidence paints a more optimistic portrait of political discourse on Reddit than popular discourse suggests. Incivility is neither rampant nor ungovernable; it clusters in predictable contexts and can be mitigated by transparent, community-specific rules enforced by motivated volunteers. These findings invite scholars to refine theories that conflate online political talk with toxicity and to investigate the socio-technical arrangements that allow large platforms to remain, for the most part, civil arenas for democratic exchange.

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

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). *The Pushshift Reddit dataset*. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Bentivegna, S., & Rega, R. (2022). Searching for the dimensions of today's political incivility. *Social Media + Society, 8*(3), 1-12. https://doi.org/10.1177/20563051221114430

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. https://doi.org/10.1111/jcom.12104

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*.

Dillard, J. P., & Peck, E. (2001). Persuasion and the structure of affect: Dual systems and discrete emotions as complementary models. *Human Communication Research, 27*(1), 38–68. https://doi.org/10.1111/j.1468-2958.2001.tb00775.x

Electronic Frontier Foundation (2025, January). Meta's new content policy will harm vulnerable users. Retrieved from: https://www.eff.org/deeplinks/2025/01/metas-new-content-policy-will-harm-vulnerable-users-if-it-really-valued-free

Forristal, L. (2025, July 15). Reddit rolls out age verification in the UK to comply with new rules. *TechCrunch*. https://techcrunch.com/2025/07/15/reddit-rolls-out-age-verification-in-the-uk-to-comply-with-new-rules/

Fox, J. (1997). *Applied regression analysis, linear models, and related methods* (2nd ed.). Sage.

Gao, Y., Qin, W., Murali, A., Eckart, C., Zhou, X., Beel, J. D., Wang, Y.-C., & Yang, D. (2024). A crisis of civility? Modeling incivility and its effects in political discourse online. Proceedings of the International AAAI Conference on Web and Social Media, 18(1), 408–421. https://doi.org/10.1609/icwsm.v18i1.31323

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.

https://doi.org/10.1093/ijpor/edw010

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

Hanley, H. W. A., & Durumeric, Z. (2023). Sub-standards and mal-practices: Misinformation's role in insular, polarized, and toxic interactions on Reddit. arXiv preprint arXiv:2301.11486. https://doi.org/10.48550/arXiv.2301.11486

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 20*(2), 139–153. 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, 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.

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

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

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy and emerging evidence on civility and effective democratic engagement. *Arizona Law Review, 54*(2), 375–441.

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. *Proceedings of the International AAAI Conference on Web and Social Media 7*(1), 400–408.

Mamakos, M., & Finkel, E. J. (2023). The social media discourse of engaged partisans is toxic even when politics are irrelevant. PNAS Nexus, 2(10). https://doi.org/10.1093/pnasnexus/pgad325

Milmo, D. (2025, April). Meta 'hastily' changed moderation policy with little regard to impact, says oversight board. *The Guardian*. Retrieved from: https://www.theguardian.com/technology/2025/apr/23/meta-hastily-changed-moderation-policy-with-little-regard-to-impact-says-oversight-board

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

Narimanzadeh, H., Badie-Modiri, A., Smirnova, I., & Chen, T. H. Y. (2025). Incivility and contentiousness spillover between COVID-19 and climate science engagement. arXiv preprint arXiv:2502.05255. https://doi.org/10.48550/arXiv.2502.05255

Newman, L. H., & Burgess, M. (2025, July 29). Age verification laws send VPN use soaring—and threaten the open internet. *Wired*. https://www.wired.com/story/vpn-use-spike-age-verification-laws-uk/

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society, 20*(9), 3400–3419. doi:10.1177/1461444817749516

Pang, H. (2024). Determining the influence of depressive mood and self disclosure on problematic mobile app use and declined educational attainment: Insight from stressor–strain–outcome perspective. *Education and Information Technologies, 29*(4), 4635–4656. doi:10.1007/s10639-023-12018-7

Pang, H., & Hu, Z. (2025). Detrimental influences of social comparison and problematic WeChat use on academic achievement: Significant role of awareness of inattention. *Online Information Review, 49*(3), 552–569. doi:10.1108/OIR-05-2024-0316

Pang, H., Qiao, Y., Li, Y., & Wang, L. (2024). Do network externalities and perceived gratifications boost WeChat continued use? Combining perspectives of network externalities and uses and gratifications. *Acta Psychologica, 251*, 104610. doi:10.1016/j.actpsy.2024.104610

Pang, H., & Ruan, Y. (2024). Disentangling composite influences of social connectivity and system interactivity on continuance intention in mobile short video applications: The pivotal moderation of user perceived benefits. *Journal of Retailing and Consumer Services, 80*, 103923. doi:10.1016/j.jretconser.2024.103923

Pang, H., & Wang, Y. (2025). Deciphering dynamic effects of mobile app addiction, privacy concern and cognitive overload on subjective well being and academic expectancy: The pivotal function of perceived technostress. *Technology in Society, 81*, 102861. doi:10.1016/j.techsoc.2025.102861

Pang, H., & Zhang, K. (2024a). Determining influence of service quality on user identification, belongingness, and satisfaction on mobile social media: Insight from emotional attachment perspective. *Journal of Retailing and Consumer Services, 77*, 103688. doi:10.1016/j.jretconser.2023.103688

Pang, H., & Zhang, K. (2024b). How multidimensional benefits determine cumulative satisfaction and eWOM engagement on mobile social media: Reconciling motivation and expectation disconfirmation perspectives. *Telematics and Informatics, 93*, 102174. doi:10.1016/j.tele.2024.102174

Pang, H., Zhao, Y., & Yang, Y. (2025). Struggling or shifting? Deciphering potential influences of cyberbullying perpetration and communication overload on mobile app switching intention through social cognitive approach. *Information Processing & Management, 62*(4), 104167. doi:10.1016/j.ipm.2025.104167

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

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. https://doi.org/10.1080/17512786.2013.813194

Siddique, H. (2025, August 2). Palestine Action ban coupled with Online Safety Act "a threat to public debate". *The Guardian*. https://www.theguardian.com/world/2025/aug/02/palestine-action-ban-coupled-with-online-safety-act-are-threat-to-public-debate

Shen, Q. (2021). *Evaluating and recontextualizing the social impacts of moderating online discussions* [Unpublished doctoral dissertation]. Carnegie Mellon University.

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

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (2020). Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse. AlgorithmWatch. https://algorithmwatch.org/en/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch.pdf

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

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

Salgado, S., Fernández-García, B., & Biscaia, A. (2025). Shades of incivility in Reddit: A comparison between echo chambers and plural spaces. *International Social Science Journal*, 75(255), 61–80. https://doi.org/10.1111/issj.12536

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs, 83*(4), 535–556. https://doi/10.1080/03637751.2016.1201207

Sydnor, E. (2018). Platforms for incivility: Examining perceptions across media. *Political Communication, 35*(1), 97–116. https://doi.org/10.1080/10584609.2017.1355857

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open, 10*(2). https://doi.org/10.1177/2158244020919447

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of*

Personality and Social Psychology, 49(3), 577–585. https://doi.org/10.1037/0022-3514.49.3.577

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

Weatherbed, J. (2025, July 15). Reddit is rolling out age verification in the UK. *The Verge*. https://www.theverge.com/news/707125/reddit-age-verification-uk-online-safety

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/