World Wide Web (2024) 27:31
https://doi.org/10.1007/s11280-024-01269-0

The medium is the message: toxicity declines in structured vs unstructured online deliberations

Mark Klein[1,2] · **Nouhayla Majdoubi**[2]

Received: 12 January 2024 / Revised: 27 March 2024 / Accepted: 9 April 2024 /
Published online: 8 May 2024

Abstract
Humanity needs to deliberate effectively *at scale* about highly complex and contentious problems. Current online deliberation tools—such as email, chatrooms, and forums—are however plagued by levels of discussion toxicity that deeply undercut the willingness and ability of the participants to engage in thoughtful, meaningful, deliberations. This has led many organizations to either shut down their forums or invest in expensive, frequently unreliable, and ethically fraught moderation of people's contributions in their forums. This paper includes a comprehensive review on online toxicity, and describes how a structured deliberation process can substantially reduce toxicity compared to current approaches. The key underlying insight is that unstructured conversations create, especially at scale, an "attention wars" dynamic wherein people are often incented to resort to extremified language in order to get visibility for their postings. A structured deliberation process wherein people collaboratively create a compact organized collection of answers and arguments *removes* this underlying incentive, and results, in our evaluation, in a 50% reduction of high-toxicity posts.

Keywords Collective intelligence · Crowd-scale deliberation · Toxicity

1 Introduction

Deliberation processes have changed little in centuries, perhaps even millennia. Typically, small groups of powerful stakeholders and designated experts craft solutions behind closed doors. Most people affected by the decisions have limited input, so important ideas and perspectives do not get incorporated, and there is often substantial resistance to *implementing* the ideas from those who were frozen from the process.

✉ Mark Klein
m_klein@mit.edu

Nouhayla Majdoubi
Nouhayla.MAJDOUBI@emines.um6p.ma

[1] Center for Collective Intelligence, Massachusetts Institute of Technology, Cambridge, MA, USA

[2] School of Collective Intelligence, University Mohammed VI Polytechnic, Rabat, Morocco

Humanity now however needs to deliberate effectively about highly complex, contentious, and existentially important problems – such as climate change, security, and poverty – where a small-circle process is no longer adequate. We need to find a way to effectively integrate the expertise and preferences of tens, hundreds or even thousands of individuals in our most consequential deliberations.

This paper addresses one important barrier to creating this capability: toxicity[1] in online deliberations. Online technology seems to represent our best hope for scaling up deliberations, but it has been plagued by debilitating levels of toxic comments. How can we fix that? As part of that discussion, we will cover:

- Goal: defining deliberation, and why scale is so important
- Challenge: the toxicity trap of existing deliberation technologies
- Solution: an introduction to deliberation mapping, a solution to online toxicity:
- Assessment: an evaluation of how well deliberation mapping reduces toxicity
- Conclusions: lessons learned and next steps

2 The goal: effective deliberation at scale

Let us define deliberation as the activity where groups of people (1) *identify* possible solutions for a problem, (2) *evaluate* these alternatives, and (3) *select* the solution(s) that best meet their needs (4).

Research from the field of collective intelligence has shown that engaging crowds in the way has the potential to unleash such powerful benefits as [1]:

- *many hands*: the advent of cheap digital communication and ubiquitous personal computing has revealed the existence of a massive cognitive surplus: very large numbers of people with deep and diverse skill sets are eager to participate in collective tasks, driven by such non-monetary incentives as contributing to problem or communities they care about [2, 3]. Wikipedia is an excellent example of this.
- *casting a wide net:* frequently, solutions for difficult problems can be found by consulting outside of the usual small circle of conventional experts in that field [4]. Innocentive is one example of a company that has been very successful exploiting this phenomenon.
- *idea synergy:* out-of-the-box solutions can often be achieved by bringing together many individuals and engaging them in *combining* and *refining* each other's ideas. The Matlab Coding Coopetition is a spectacular example of the power of this effect [5]
- *wisdom of crowds:* large numbers of suitably diverse, motivated and independent raters have been shown to produce assessment accuracy—e.g. for prediction and estimation tasks—that exceeds that of experts [6]. Prediction markets are a powerful example of the value of this phenomenon.
- *many eyes:* our ability to detect possible problems in solution ideas increases dramatically by simply engaging more people in the task. This has been one of the key reasons for the success of such volunteer-created open-source software tools as Linux

[1] We define toxicity as the presence of rude, disrespectful, or unreasonable comments that are likely to make people leave a discussion.

(the dominant operating system for supercomputers), Apache (the most widely-used web server), MySQL (the most widely-used relational DB) and the web toolkits used by Chrome, Firefox (the most popular web browsers in the world). These open source tools have decisively out-competed software developed by massive software companies with thousands of highly-paid engineers [7].

Engaging the relevant stakeholders in making decisions also has the great advantage of reducing the resistance and confusion that can occur when trying to actually *implement* the solutions developed by the deliberation engagement.

3 The challenge: online toxicity in existing deliberation technologies

We conducted a systematic literature review, using the PRISMA 2020 methodology, to better understand the phenomenon of online toxicity. We queried 3 major databases (SCOPUS, Sage Journals, JSTOR) for articles available in English, using toxicity and incivility as key terms, and adding terms referring to online mediums and tools (e.g. social media, platform), toxicity/incivility-related terms (e.g. negative, offensive, toxic*, incivil*, uncivil*), and collective deliberation (e.g. debate, collaboration, deliberat*). We excluded articles from unrelated domains like chemistry or biology that potentially conflate toxicity with other terms, as well as toxicity studies in online gaming communities as those represent a different case study than deliberation. Our review included 91 articles meeting these criteria, the majority of which are recent.

3.1 Defining our terms

Incivility, recognized as a challenging concept to define, involves disruptive behaviors that induce physical and psychological stress [8]. This aligns with a parallel definition of toxicity, often characterized by the use of rude, disrespectful, or unreasonable language aimed at provoking or driving users away [9, 10]. This definition can be context-dependent, for example Hwang and Kim frame political incivility as the expression of disagreement through the denial and disrespect of opposing views [11]. Some of the more comprehensive definitions come from the works of Bormann [12] and Frischlich et al. [13]. Frischlich et al. distinguish types of incivility by categorizing violations into interpersonal and intergroup communication norms, including offensive speech and hate speech. This work draws highlights that these definitions call for a context-dependent exploration of online norms, varying across platforms, which introduces subjectivity in perceptions. Further complicating this is the fact that incivility is not only text-based, but encompasses various channels such as images, audio, or video material [13]. In an attempt to bypass these challenges, Bormann, opts for an experimental approach, directly inquiring about participants' perceptions of incivility. This leads to him identifying five categories: informational, formal, processual, personal, and anti-democratic incivility. His multidimensional model aligns with perceptions of norm violations [14]. We concluded that it's imperative to include both incivility and toxicity in online discussions.

3.2 The lost promise of online deliberation

Online spaces, particularly commenting boards, were initially considered promising avenues for democratic debate, aligning with Habermas' famous vision of an inclusive public sphere where citizens engage in discussions on social and political matters [15]. This optimism was grounded in the belief that these platforms could facilitate digital political participation and contribute to a deliberative public sphere [16, 17]. Theoretically, comment sections were envisioned as platforms for rational and respectful exchange of diverse viewpoints, where the power of the better argument prevails. Scholars hoped that user comments would enhance inclusivity, participation and deliberation in the public discourse [18]. They were seen as these spaces providing additional information, introducing journalistic content to a diverse pool of citizen opinions and broadening readers' interpretations of issues and events [19]. Moreover, user comments were expected to play a crucial role in opinion formation, influencing perceptions of journalistic quality, and fostering engagement in dialogues [19, 20]. Overall, initially, the internet was seen as a promising infrastructure for a deliberative public sphere accessible to everyone, fostering rational and respectful discussions on social and political issues [21]. Deliberation was considered a valuable input for a healthy democracy, and scholars frequently referred to deliberative norms to assess the quality of online user comments [17, 22].

In online deliberation, the quality of discourse is shaped by several key dimensions, and civility emerges as one such crucial factor. Friess and Eilders conducted a comprehensive analysis of theoretical and empirical research on online deliberation, identifying pillars for it such as rationality, interactivity, equality, inclusiveness, civility, common good reference, and constructiveness [16]. Civility, also considered a facilitating factor for constructive deliberation in studies like Santana [23] and Friess et al. [24], is deemed a prerequisite for deliberation, emphasizing the importance of mutual respect [25]. During deliberative exchanges, scholars have been critical of practices deemed disrespectful, underscoring the significance of maintaining a civil tone in such contexts. Moreover, some research suggests that civility plays a role in sparking conversations, as participants exposed to civil comments are more likely to engage in discussions, with this relationship mediated by elaboration [26]. While disagreement is inherent and valuable in deliberation, exposure to uncivil disagreement, characterized by personal attacks and derision, can have erosive consequences, challenging the core principles of constructive discourse [11].

In reality, civility, a fundamental norm for deliberative discourse, is frequently violated in online discussions, transforming platforms into areas rife with frustration, hate, and incivility, often displayed through insults and flaming. The escalating concern about online incivility is evident among the American public, with 68% identifying it as a "major problem" and nearly 90% recognizing its severe consequences, encompassing cyberbullying, harassment, violence, hate crimes, intimidation, threats, intolerance, and a diminished sense of safety in public spaces [27, 28]. Several scholars underscore that incivility, rather than civility, prevails in online conversations, despite the potential for political discussion on social media [26, 29, 30]. User comments in particular, once viewed optimistically, often do not deliver on their promise of civil exchange due to uncivil and aggressive discourse, and efforts to eliminate uncivil messages face significant challenges [19–21, 31]. This increasingly pessimistic perspective on user comments has gained traction, leading news outlets to move discussions to social networking sites

and shut down comment sections on their websites due to their perceived low quality [21]. But even on social networking sites, low-quality user discussions pose challenges to individual-level democratic benefits such as increased knowledge, tolerance, and familiarity with diverse viewpoints [32].

Some studies suggest that incivility lowers the deliberative quality of online discussions [33]. For example, Collins highlights that incivility is not always linked with substantive arguments, therefore playing a peripheral role in discussions with key comments characterized by more sophisticated argument structures [20]. Other studies however show that rational reasoning and incivility are not mutually exclusive [34, 35]. To reconcile these two perspectives, we can look to a particular study which uncovered that uncivil comments exhibit lower persuasiveness overall, which presents a hurdle to achieving mutual recognition and social cooperation [36]. Incivility and toxicity, while not inherently devoid of logic, pull discussions out of the realm of pure reason, hindering the potential for meaningful deliberation. Even when these comments contain valuable arguments, their emotional underpinnings lead to negative consequences since exposure to online incivility is linked to the release of negative emotions, hostile cognitions, and perceptions of polarization [37–40]. Moreover, it increases the likelihood of further uncivil reactions, discourages user participation in networked discussions, and serves as a key marker of opinion polarization [36]. Detailed exploration of these consequences will follow in a subsequent section.

The initial excitement around online discussions has thus given way to a gradual but definite disillusionment. Various studies conducted over a significant period of time have allowed us to analyze the evolution of this toxicity phenomenon. Our examination reveals that there has been a noticeable increase in toxicity. The contradiction between the potential and the actual state of online deliberation reflects a significant gap between the ideal and the practice of digital democratic discourse.

3.3 Reasons for online incivility

In what follows, we investigate the origins of this prevalent and seemingly persistent online incivility. We will consider whether incivility is influenced by topics, platforms, users, or cultural factors. The evidence suggests, in fact, that the roots of online incivility are a combination of all of the above.

3.3.1 Topic-driven toxicity

Evidence for topic-driven toxicity/incivility in online discussions: Early observations by researchers, derived from a meticulous content analysis of comments across various topics, reveal that incivility is a frequent occurrence linked to key contextual factors, including the article's subject matter and the cited sources [15]. Notably, "hard news" topics tend to provoke greater incivility, while articles focusing on "lighter topic" such as lifestyle and technology exhibit considerably lower levels. An exception to this trend is sports, a lighter topic that paradoxically experiences the highest percentage of incivility. Subsequent research in 2020, involving the analysis of 7 million YouTube comments, supports these findings, with approximately 69% of the collected videos containing toxic comments [41]. Further analysis by the authors indicates that religion and crime-related news attract the highest rates of toxic comments, while economy-related news sees the lowest rate. Other studies point to various topics driving toxicity, including open-source programming, immigration, genetic

testing, climate change, wind energy, sexism, consumer dissatisfaction, history, disputed historical facts, investment stocks, racism, religion, abortion and sexual assault [42–50].

Political topics in particular have garnered considerable attention as being notoriously challenging to discuss civilly. Previous research underscores that people are significantly more likely to exhibit incivility online in the context of political topics and disagreements [51–53]. But even within politics, not all topics are equal; Analyzing 55,053 comments, one study emphasizes the relevance of the political news story's topic in sparking political disagreement [51].

Despite discussions on incivility prevalence, the overall level of incivility remains low online even around controversial topics [42, 54]. Moreover, in the political context, incivility occasionally extends to individual politicians, particularly their viewpoints on specific issues. For instance, one study examining comments in the political sphere shows most incivility is directed towards *political candidates* [41]. Another analysis of Facebook comments on parliamentary candidates during the Italian general election of 2018 found impolite comments, driven by feelings of partisanship, expressed mutual hostility, often overshadowing the issues at hand [54]. Another study analyzing 18 million tweets surrounding political candidates in America revealed associations between certain candidates and incivility discourse, with specific policy issues closely linked to uncivil discourse. Linking these results through k-means clustering, the study illustrates that gun control and immigration are closely related to mentions of controversial candidates [28]. Other research indicates there is an important distinction to make here; impoliteness is directed at individuals, while incivility is predominantly driven by the topic. This is further supported by the analysis of YouTube and newspaper comments under Al Jazeera, where toxicity appears to be topic-driven, rather than targeting any specific individual [46, 54].

Why could it be topic driven? As we just hinted at, certain topics can spill over into others. For example, discussions following news about immigration, especially when allowing anonymity, can lead to emotionally charged, uncivil comments directed at *Latinos* by supporters of strict immigration laws [47]. This illustrates how discussions on immigration can become about issues of race and minority targeting. Similarly, topics like genetic testing can extend into discussions of racism, with certain subreddits associating genetic testing discussions with hateful, racist, and sexist content, and suggesting potential links to racist and anti-Semitic agendas on platforms such as Twitter [48].

In politics, one theoretical explanation suggests that the topic itself is inherently uncivil/toxic, and this is further exacerbated by the heterogeneity and polarization of the American political context. In summarizing multiple theoretical perspectives, Hopp and Vargo [55] suggest that the broadening of political participation, especially among those with high prior levels of political interest, contributes to increased incivility. Heterogeneity, linked to incivility through sociopsychological theories, suggests that communication among diverse individuals lacks the essential social tools for civil conversation, such as trust and mutual feelings of obligation. This leads to an exacerbation of conflicts, fostering a more uncivil discourse.

Another interesting theoretical perspective on why certain topics could drive incivility comes from a study on discussions of sexual assault. The authors survey literature around the just-world bias to conclude that the belief that good things happen to good people and vice versa, is linked to incivility [44]. The authors refer to past research indicating that individuals react with empathetic anger, hostility, and aggression when faced with situations that challenge their belief in a just world. In discussions of sexual assault, just-world bias can manifest as individuals siding with in-group members and blaming out-group members, serving as a coping mechanism for managing distressing emotions [44].

We can summarize the key insights so far as follows:

Toxicity and incivility are linked to the nature of topics. Hard topics encompass such areas as politics, law and order, taxes, foreign affairs, sports, climate change, wind energy, sexism, consumer dissatisfaction, history, historical facts, racism, religion, abortion, and rape. Lighter topics include humanistic stories, health, lifestyle, journalism. Incivility related to topics tends to persist across years and platforms for hard topics, spanning newspaper websites, Facebook, Twitter, Reddit, and 4Chan.

Whereas impoliteness can be personal, incivility is topic-driven. The former is directed at specific individuals, notably political candidates. In contrast, when incivility is aimed at political candidates, it often is closely associated with their positions on hard topics.

Theoretical explanations for topic-driven toxicity include sociopsychological factors like heterogeneity and polarization, as well as psychological factors like just-world bias, which can intensify conflicts and incivility.

3.3.2 Platform-driven toxicity

The impact of online platforms on driving incivility has been a consistent concern, dating back to YouTube's portrayal as an unregulated hub of hostility in 2013 [56]. As stated by Murthy and Sharma, theoretical perspectives from scholars like Gilroy, Nakamura, and Chow-White emphasize the role of the web as a significant space for public conversations about race, contributing to the emergence of new forms of racism [57]. Recent studies, including a 2021 investigation of 18 million tweets, reinforce the notion that incivility is contingent on environmental conditions, such as the specific place and time of online discussions. This idea finds support in diverse contexts, as demonstrated by research on Online Learning Environments for dentistry students, which, despite being seemingly unrelated to contentious topics, highlights the role of platform affordances in driving incivility [58]. Yet another fascinating study revolves around a design intervention using images, demonstrating its effectiveness in mitigating online incivility. The findings indicated that positive backgrounds, both in color and grayscale, were successful in reducing incivility in online experiments which offers a promising direction for the design of more civil discussion platforms in online settings [59]. Additionally, a study linking social media use to uncivil participation, particularly on a Russian platform with lenient moderation rules and ultra-right-wing content, further underscores the association between platform features and incivility [13]. A prevailing view in the literature is that the spread of incivility is linked to inherent features of computer-mediated communication, as argued by one author [34]. Finally, in a meta-synthesis conducted by Ng et al. [60], exploring 36 research articles and 42 studies with 19,464 participants and 11,287,011 online comments, platforms emerged as significant contributors, with comments displaying higher levels of incivility on Twitter compared to Facebook for example.

To further demonstrate the impact of different online platforms, an examination of comments on Danish news sites and Facebook accounts uncovers variations in comment frequency, elaboration, and liveliness [61]. Notable distinctions in incivility levels emerge between platforms like Facebook and news websites, as evidenced in studies comparing the Washington Post's website and Facebook accounts [22, 34]. The deliberative quality is notably lower in Facebook comments than on news websites. Other inquiries reveal diverse expressions of incivility across platforms, with political blogs exhibiting more instances compared to mainstream outlets, particularly involving insulting language, vulgarity, and stereotyping of political parties [62]. Expanding on platform comparisons, Facebook

demonstrates lower levels of incivility than Twitter, while a comparison between Twitter and Reddit highlights differing manifestations of incivility on both platforms [44, 63]. In another study examining COVID-19 discussions on Twitter and Parler, higher toxicity levels are observed on Twitter than on Parler [64]. A compelling case study analyzed 100 toxic discussions from GitHub, revealing distinct forms of toxicity in open source compared to platforms like Reddit or Wikipedia. This includes entitled, demanding, and arrogant comments from project users, as well as insults stemming from technical disagreements [50].

The research strongly suggests a significant influence of platforms on incivility. But what platform affordances are shaping or driving this incivility and toxicity? Platform affordances refer to the unique capabilities and features that a digital platform provides to its users. These affordances shape the ways in which users interact with the platform and influence the types of activities and behaviors that are facilitated. In what follows, we will look at the platform affordances mentioned in the literature and the influence they may or may not have on provoking uncivil behavior.

The influence of moderation The moderation approach adopted by online platforms emerges as an important factor influencing the levels of incivility in user comments. Different content moderation strategies, now combining artificial intelligence and human moderators, contribute to significant variations in platforms' incivility levels. For example, the Der Standard website and Facebook comments, employing distinct moderation methods, are observed to have differing levels of incivility, suggesting a potential correlation [61]. Examining the disparity in incivility between Twitter and Reddit, another study emphasizes the role of moderation rules. The lenient stance on profanity on Twitter, with lower expectations of being banned, may foster more profane responses. In contrast, Reddit users may self-censor profanity, adhering to subreddit rules threatening bans for such behavior [44]. However, moderation rules alone don't offer a complete explanation, as evidenced by Twitter's stricter rules but higher toxicity compared to Parler [64].

The influence of anonymity The impact of anonymity is a widely discussed affordance in online discussions. Halpern and Gibbs argue that the level of identifiability vs. anonymity serves as a media affordance influencing online interactions [65]. Drawing on deindividuation theories and platform affordances there is an anticipation of heightened incivility in anonymous settings. Deindividuation theories suggest that anonymity leads to socially deregulated behavior due to reduced self-assessment [61]. Moreover, computer-mediated communication, lacking non-verbal cues present in face-to-face interactions, coupled with platform affordances such as anonymous participation, contributes to the inherent incivility in online discussions [34]. Another hypothesis [44] is that the anonymity prevalent in various online communication spaces (e.g., Reddit) enables users to engage without risking damage to their reputation, especially when making uncivil comments. The reduced cues model in social science and the online disinhibition effect [66] collectively suggest that the removal of identity lowers inhibitions, often resulting in incivility, though caution is advised in attributing causation. In response to this, many media organizations have shifted toward social platforms like Facebook to enhance comment quality by leveraging personal accounts and reducing anonymity as reported by researchers [61].

In settings where users can remain anonymous on social media, many studies indeed report a higher incidence of incivility compared to non-anonymous counterparts [22, 30, 47, 65, 67–69]. Delving deeper into specific studies, an analysis of 4,800 comments from online commenting forums of top news sites highlighted anonymity as a key factor, with

anonymous commenters being not only more likely to exhibit incivility but also less likely to meet the academic literature's criteria for quality dialogue [23]. The study concludes that anonymous commenting boards on news sites are more prone to incivility, while non-anonymous forums tend to host more civil sentiments. Another notable study found a significant increase in uncivil behavior in reader comments on platforms that allow user anonymity compared to those where commenters are identified and accountable for their content [22]. In the anonymous condition, uncivil behavior was more likely directed at discussion participants, while the non-anonymous condition saw instances of incivility aimed at individuals not involved in the discussion or used to articulate an argument without intending offense.

As a most studied affordance, the role of anonymity in online incivility sparked some scientific disagreement. Despite the common attribution of aggressive behavior to internet anonymity, we can all attest to the fact that contemporary less anonymous platforms, like social networking sites, witness users generating numerous aggressive comments, triggering negative responses [39, 70, 71]. Contrary to the assumption that Facebook's public nature would constrain antinormative behaviors, a study challenging this hypothesis emphasizes that uncility is still prevalent on the platform [34]. These observations extend to Twitter, where there seems to be no clear relationship between anonymity and incivility, as uncivil tweets are as likely to come from identified users as from anonymous or pseudonymous accounts [28].

Taking a more provocative stance, one paper from the literature references research suggesting that anonymity might enhance public deliberation by fostering inclusivity and encouraging participation [30]. These divergent findings raise questions about the conventional understanding of anonymity's role. Acknowledging this, some researchers highlight a limitation in previous studies, noting that the purported effect of anonymity is primarily based on observational cross-sectional studies comparing user behavior across media channels. These cross-platform research designs struggle to disentangle the impact of anonymity on political discussions from other affordances or contextual factors [30].

Perhaps, however, there is no disagreement here. Many studies highlighting anonymity as an issue date back to the early years of the internet. In contrast, more recent studies indicate a pattern where identifiable users, including verified Twitter accounts and political candidates, participated in uncivil discourse. This suggests an evolution over time. Additionally, we consistently observe authors noting the influence of platform norms, networks, in-group dynamics, and their interaction with the element of anonymity. Let's delve deeper into those aspects.

Communities and Norms The studies collectively underscore that the norms in online communities significantly influence the prevalence and acceptance of incivility, especially in political discourse. Hmielowski et al. note that regular engagement in online political discussions socializes individuals to perceive flaming as acceptable, a view that intensifies among those with high verbal aggression [72]. This is echoed by Rossini [34], who suggests that frequent engagement in such discussions leads to a perception of incivility as a normative behavior. Supporting the role of norms, Sobieraj & Berry [73] find a higher degree of incivility in independent blogs than in mainstream-affiliated ones, attributing this to the echo chamber effect in virtual communities. For their part, Ziegele et al. [21], along with Ruiz et al. [17], emphasize how the dynamics within individual news communities contribute to different qualities of discussion, evolving into either debate-oriented or echo chamber-like environments. Rossini's further investigation into the targets of uncivil discourse on online platforms like Facebook and news websites reveals that uncivil rhetoric,

particularly towards politicians, is considered acceptable by those engaging online, perpetuating a norm of incivility [34].

In-group Behavior and Social Network ties The influence of in-group behavior and networks on incivility is well documented. It is intertwined with and explanatory of the emergence of community norms. Identification with an in-group audience can stimulate expressions of commitment to one's in-group. These partisanship dynamics can subsequently foster incivility with individuals exhibiting confirmation and disconfirmation biases [40, 74]. One paper [75] discusses how uncivil behavior garners group support in large networks, particularly in Polish internet interactions which is problematic given that uncivil behavior is contagious through contacts in social networks, with reciprocity playing a major role in the contagion of political incivility [52]. Another way these networks form is thar uncivil commenters who get their account suspended, tend to form more closely-knit communities [49]. Simply put, in-group members are more likely to perpetuate the norms of incivility and uphold them if they are already stablished within the community. Conversely, people are more likely to be civil when constrained by the social ties and networks where uncivil behavior is not accepted [34]. In addition, Jaidka et al. demonstrate that social identifiability, coupled with personal anonymity, leads to greater rationality and civility in online discussions, suggesting a conformity to group norms [30]. The literature suggests other ways in which identification with the in-group can stimulate incivility. For example, reactions to incivility differ based on whether it aligns with one's political views, and users more prone to incivility within opposing communities [39, 76]. To add a little bit more nuance to these findings, Trifiro et al. and Oh et al. note that while incivility often occurs in heterogeneous settings, intolerance is more likely in homogeneous, echo chamber-like environments, with a small number of 'super-participants' driving much of the uncivil and intolerant discourse [28, 43].

The literature consistently demonstrates that incivility often begets further incivility, a complementary notion of community norms. Chang et al. observe that harmful comments and negative emotions create self-reinforcing feedback loops in forums [42]. Rösner et al. and Stevens et al. explain this through social influence research and conformity theories, where exposure to uncivil comments can serve as social modeling, legitimizing verbal aggression [39, 44]. Some researchers show that when uncivility is initially affirmed, it is more likely to be repeated [77]. However, even the victims of incivility, familiar with its negative consequences, can retaliate with the same. Political candidates who experience incivility from opponents are more likely to use incivility themselves [78] and Frischlich et al. find that nearly half of those who witnessed incivility contributed to its spread, with victimized participants being more likely to engage in uncivil behavior [13]. Interestingly, Rösner et al. report that the expected pattern of increased incivility following exposure to uncivil comments did not always occur, suggesting that modeling effects might be less severe than previously thought [39].

The key insights on platform-driven toxicity / incivility thus include:

- Online platforms profoundly influence incivility, with studies showing that platform-specific features like UI affordances, moderation, and anonymity significantly shape user interactions and the prevalence of uncivil behavior.
- Platform design elements (even in non-contentious settings) contribute to the emergence and escalation of incivility, with some interventions, like positive image backgrounds, effectively mitigating it.

- Anonymity is a crucial factor in driving incivility, with theories suggesting that less identifiability leads to deregulated behavior and heightened toxicity. Notably, efforts to reduce anonymity on platforms like Facebook have aimed to enhance comment quality. The impact of anonymity however, has dampened over the years, with identifiable commenters also resorting to incivility.
- Engagement in online political discussions often leads to the normalization of uncivil behavior, especially when such interactions occur in echo chambers or among like-minded individuals. In-group identification and dynamics play a crucial role in shaping attitudes toward incivility.
- Social networks facilitate the contagion of incivility, with the dynamics of reciprocity and group norms playing significant roles. Both victimization and witnessing incivility can lead to further participation in uncivil behavior.
- While incivility often begets more incivility, creating self-reinforcing cycles, exposure to civility can also positively influence online discussions, highlighting the potential for positive modeling effects.
- Some theoretical frameworks argue that inherent features of computer-mediated communication, coupled with the unique affordances of each platform, continue to foster environments where incivility can thrive.

3.3.3 User-driven toxicity: Who are the uncivil users?

Research to date helps us understand the profiles of uncivil commenters. One recurring theme suggests a positive association between increased online political participation and higher levels of incivility [72]. On Twitter, the manifestation of uncivil discourse is concentrated among a select group of 'super participants' [79]. Still on Twitter, analyzing tweets from the 2012 electoral campaign, a 2017 study revealed that designated market areas with elevated participation exhibited heightened political incivility [55]. A similar pattern emerges on Reddit, where author propensity and discussion context toxicity act as strong positive antecedents of language toxicity, impacting both volume and user evaluation in specific sub-communities [80]. More evidence comes from a study of YouTube comments, in which a positive correlation was discovered between the overall toxicity of an online discussion and its length, measured in both the number of comments and time. These findings echoed the concept encapsulated in Godwin's Law, coined by Mike Godwin in the 90 s, which posits that "As an online discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches to one." [76].

Contrasting these findings however, an earlier study challenges this notion, asserting that contrary to popular perception, frequent commenters are often more civil than their infrequent counterparts, and uncivil commenters show no significant difference in their use of evidence to support claims [15]. Even recent research on Reddit reveals that the depth of a comment in the reply structure and its length are significant predictors of toxic incivility, even when controlling for the political alignment of the subreddit. Interestingly, comments at a deeper level tend to be, on average, slightly less toxic. [81]

The link between higher participation levels and increased toxicity is complicated. A study suggests that this association is conditioned by economic status, with the indirect effect between negative advertising exposure and citizen incivility being most pronounced in areas with lower economic status [55]. In a recent exploration of political candidate tweets, challengers and candidates in less competitive races were found to be more prone to uncivil rhetoric, while women, racial minorities, and candidates in open seat races

exhibited lower tendencies toward incivility [78]. This observation of women displaying lower incivility is corroborated by another study conducted during a political campaign in Zambia, revealing stronger effects of exposure to online political campaign messages on incivility/hate speech among male participants [82].

Additional personal factors were found through a 2020 meta-synthesis on the antecedents of online incivility [60]. Dispositional predictors indicated that negative traits, including attentional impulsivity and boredom susceptibility, were positively associated with online incivility, while positive traits like openness to experience and agreeableness showed a negative correlation. Developmental predictors, such as education level and age, underscored that less educated and younger Twitter users were more prone to engage in uncivil discourse. Finally, social predictors encompassed factors like ethnic heterogeneity, residential tenure, and unemployment rate, demonstrating their noteworthy influence on the prevalence of online incivility.

What role does political affiliation play? Debates surrounding the reliable association of toxic behavior with specific political affiliations have sparked controversy, with limited evidence pointing towards either left or right-leaning ideologies. More studies have explored the connection between right-wing extremism and increased incivility and intolerance. Notably, reported micro-level analyses identified intolerance as a distinct pro-life behavior in American and Irish Twitter spheres, aligning with other extant literature linking uncivil society to reactionary right-wing discourse [43]. Similarly, research on frequent users of the Russian-based network VKontakte hosting ultra-right-wing content, reported more uncivil participation, implying that such platforms may attract individuals engaging in such behavior [13]. However, conflicting perspectives emerge, as other studies on a large corpus of tweets deny a direct association between toxic profiles and politics, emphasizing the complexity of this issue [83, 84]. This is all confirmed in the recent meta-synthesis on the antecedents of online incivility where political identity was found to elicit inconsistent findings regarding liberals and conservatives [60].

Several notable studies have delved into the detailed profiling of toxic users on specific platforms, shedding light on their distinctive characteristics across platforms like YouTube, Twitter, GitHub, and Reddit. One study meticulously scrutinized the top 1% most toxic Twitter profiles, unveiling distinctive patterns that go beyond mere tweet volume [84]. These profiles, while comparable in total tweets, exhibit lower retweet activity, intermittent tweeting without fixed intervals (suggestive of automation), and a penchant for shorter, more comprehensible language. Linked domains span diverse categories like pornography, news, and information technology. Strikingly, despite their verified status, these profiles maintain fewer connections and followers, and they are less politically affiliated, with a majority originating in the US and being created during U.S. election years. In the realm of toxic behavior on Reddit, some researchers introduced a nuanced categorization of users into four types: Steady Users, Fickle-Minded Users, Pacified Users, and Radicalized Users. Fickle-Minded Users emerge as the predominant group, dynamically oscillating between toxic and non-toxic commenting over time. This challenges conventional notions that users can be rigidly classified as either purely toxic or non-toxic [85]. This observation of toxicity regularly occurring was also persistent in an analysis of Italian YouTube comments in 2021. This study reveals that hate speech is occasionally triggered in regular users, emphasizing a nuanced relationship between toxicity and polarization [76]. And finally in a scrutiny of toxic behavior on GitHub, the authors challenged the notion that toxicity is solely external. Internal toxicity surfaces, with project members actively contributing to toxic comments. While toxicity predominantly resides in popular, active repositories, it also manifests in smaller, inactive

projects, especially those focused on libraries or end-users. Gaming projects, while less toxic, exhibit more severe language when toxicity is present [50]. The toxic behavior of these users seems distinct insofar as the most prevalent forms of toxicity are entitled, demanding, and arrogant comments from project users as well as insults arising from technical disagreements, as opposed to direct insults or flaming.

These studies collectively highlight commonalities in intermittent toxicity prevalence, fluid user types, and the presence of toxicity within both large and small projects or communities across platforms.

What is the role of language and culture? Scholars have increasingly focused on toxicity studies, even within online gaming communities, across different global regions like Europe, Africa, MENA, Asia, and the Americas [61]. Cross-national examinations are powerful insofar as they consider regulatory practices, cultural norms, journalistic cultures, media systems, and political contexts [61]. For instance, one analysis revealed higher toxic outrage in comments in majoritarian democracies compared to consensus-oriented ones [86].

The U.S is a well-studied example, and examinations comparing it to other countries, show important variations. German news organizations' Facebook sites exhibit lower hostile emotions than those in the U.S [87]. Abortion discussions in the U.S. contain more incivility than in Ireland, attributed to America's history of violent abortion activism [43]. In Hungarian commentary culture, a study finds cruelty and rudeness akin to the U.S. context [15]. However, one study indicated more reasoned and polite online deliberation compared to Russia, showcasing differences in communicative culture and argumentation and arguing that deliberation is more developed in America as opposed to Russia where it is regarded more as entertainment [88]. Exploring the difference in toxicity between English and Polish users, one study unveils distinct patterns. English exchanges utilize sarcasm, irony, and hedging, often pointing to external opponents causally, while Polish comments exhibit a higher incidence of denigrating remarks and frequent hostility [75].

Further afield, a study in South Africa detects instances of collective ranting, threats to democracy, antagonistic stereotyping, and sarcasm [89]. It's interesting to note that here the data was collected from a South African newspaper Mail & Guardian Online, one of the few national media establishments that permitted user comments at that time of election which might hint at a level of censored speech. In China, two studies specifically investigated the unique landscape of online interactions in this authoritarian context [90]. One study presented a nuanced perspective, asserting that online incivility in China may not only be state sponsored but also initiated by netizens. In contrast to discussions in the American context, which often revolve around media manipulation in the realm of conspiracy theories, this article presents allegedly well-documented incidents of state-sponsored incivility as a strategic tactic [91].

The key insights on user-driven toxicity thus include:

- Incivility is more linked to extensive discussions and in particular, people with higher political participation. However, numerous factors, including platform norms, group norms, and personal characteristics such as impulsivity, boredom and social status, intertwine, making it nearly impossible to untangle them and pinpoint a single factor contributing to toxic user behavior.
- Some studies attempted to establish a link between political affiliation and incivility, but depending on one's political leanings may not be an adequate basis for drawing such conclusions; It appears that individuals of any political affiliation can display incivility under specific circumstances.

- Across platforms like Twitter, Reddit, YouTube, and GitHub, top toxic users often exhibit patterns of intermittent, lower-engagement posting, dynamic shifts between toxic and non-toxic behavior, and a tendency towards more direct and comprehensible language, challenging the notion of fixed toxic identities. This reveals a complex interplay between platform dynamics, particularly the established norms within them, and individual behavior.
- Cultural and linguistic differences significantly influence online incivility, with variations in toxic discourse styles across countries and within different political social contexts.

3.4 The multifaceted nature of online incivility

The literature on incivility online presents a multi-dimensional view, illustrating its complexity and connection to various factors. Szabo et al. identify platform, communicative situation, and surrounding comments as key influences on incivility, while noting that the timing of events like elections has less impact than expected [92]. Stevens et al. argue that there is no simple explanation for incivility, pointing out that linguistic features of news and disagreement provoke different modes of incivility, which manifest uniquely across platforms [44]. Jakob et al. show that the interaction effects between topic, platform, and country significantly impact the likelihood of toxic outrage in user comments [86]. Chen & Wang add that a variety of factors, including digital platform affordances, content types, intergroup dynamics, and partisanship, are potential triggers for online political incivility [52]. These studies collectively underline that incivility is a multi-faceted issue, interlinked with various elements that are challenging to disentangle.

3.5 The consequences of online incivility

3.5.1 Online Incivility: the bad and ugly

A plethora of studies collectively unveil the profound and varying repercussions of online incivility, as echoed in Ng et al.'s meta-synthesis [60]. Their examination delves into the moderating and mediating effects of emotional and cognitive factors on psychological and behavioral outcomes. These mediators unveil the impact of uncivil comments on readers' hostile cognition, perceived news quality, and induced negative emotions, a thematic resonance with our own research detailed in what follows.

As early as 2014, Coe et al. emphasize the counterproductive nature of incivility, noting its hindrance to meaningful conversation and the rarity of its ability to encourage speaking out against it [15]. Exposure to uncivil discourse can polarize opinions, especially in the case of little-known topics [32]. In other studies, Gervais et al. delve into the psychological aftermath by demonstrating that exposure to disagreeable uncivil political talk can induce anger and aversion, diminishing satisfaction with message board discourse [33, 37]. This is further corroborated by Rösner et al., who reveal that even a slight extent of incivility (one uncivil comment among six) can elicit hostile cognitions [39]. Uncivil discussion can also heighten negative emotions and closed-mindedness [11]. It has also been shown that people actively avoid engaging with comment threads starting with uncivil content, indicating a defensive response to toxic online environments [93]. Even those merely observing toxicity, as in the case of Maintainers on Github, experience substantial emotional taxing effort [94].

The repercussions of online incivility are not only evident for users, but also for the platform owners, compelling various prominent organizations to either completely shut down their online discussion forums or invest heavily in moderation to combat toxicity. This trend is evident in the actions of major platforms such as CNN which shut down its comment section since 2014[2] and NPR in 2016,[3] both opting to move the interactions to social media platforms, believed at the time to harbor less incivility. Multiple other websites are routinely forced to temporarily disable comments at times because of the internet trolls. In order to keep the comment sections up, they have to invest in heavy moderation; Such is the example of The New York Times.[4]

And the consequences extend beyond the digital realm. Previous research reported a negative relationship between exposure to uncivil online political comments and offline political participation or even worse, a potential causal connection to offline violence [76, 95]. Another study sheds light on the detrimental impact of incivility on democracy, fostering polarization in the audience, undermining public trust, leading to negative perceptions of news outlets and jeopardizing deliberation [92, 96, 97]. Notably, Sobieraj highlights how incivility acts as an exclusionary force, particularly affecting women in public conversations [98]. The interconnected web of repercussions extends further, as far as resulting in mental health issues, as mentioned by Cover [99]. Despite these well-documented negative effects, Rossini emphasizes the challenge of comparing findings due to different operationalizations of incivility, underscoring the lack of consensus on its definition [34].

3.5.2 Online Incivility: the good

One thing that incivility has been proven to be effective at time and again in the literature is getting your voice heard [38, 62]. Expanding on this, Rains et al. [100] put forth the idea that incivility serves as a mechanism for identity consolidation, particularly in visually anonymous computer-mediated communication. This underscores its significance in shaping both individual and collective identities within online interactions. Incivility seems to be this force, steering engagement across political issues and beyond. Various studies illuminate its impact, especially for political candidates. The findings reveal a paradox where uncivil tweets, adorned with rudeness, spark significantly more likes and retweets, hinting at incivility's effectiveness in fueling political discourse [55, 78, 82, 101]. This engrossing engagement might stem from the emotional triggers that incivility activates, propelling the sharing of news articles laden with sentiments such as anger, happiness, or awe as mentioned in one study [10]. Politicians strategically employ incivility as a tool to capture voter attention, raise awareness, and elevate content engagement in the political arena [55, 78, 82, 101]. Toxicity is also a tool for accruing social capital for political factions, as evidenced by its rewarding impact on news sources associated with it [10]. This strategic use not only benefits candidates but potentially the public as well. A fascinating study suggests that uncivil remarks toward politicians may act as a societal alarm, alerting and exposing others in the network to pertinent political issues, embodying the essence of monitorial citizenship [51].

[2] *"Online comments are being phased out", Doug Gross, CNN, 2014.*

[3] *"NPR Website To Get Rid Of Comments", Elizabeth Jensen, 2016.*

[4] *"How The New York Times moderates 12,000 comments a day", Lucia Moses, 2017.*

Beyond the political spectrum, incivility is shaping social network landscapes. Interactions with comments, measured by likes and replies, play a pivotal role in determining visibility on platforms like Facebook. One study reveals correlation between rational, impolite, and intolerant comment characteristics and the interaction count, affirming the intricate dance of incivility in online spaces [19]. Twitter mirrors this pattern, where many toxic profiles are verified, enhancing the virality of their content [102].

Contrary to common belief, one study suggests that even when people express their opinions in a less-than-polite way, it can actually lead to more reasonable and civil conversations later on. This idea of "robust civility" means that, according to a liberal perspective, societies should embrace opinions even if they're delivered uncivilly, but there's a catch; It works best when others respond in a calm and democratic manner [21]. Intriguingly, reading more rude comments doesn't always make people more hostile. It might depend on how invested someone is or how uncivil the content really is [11]. So, instead of incivility just causing more of the same, it can stir up diverse views and lively discussions.

In addition, incivility does not always hamper the quality of argumentation. Research by Rösner et al. asserts that incivility doesn't necessarily lead to a drop in argumentative standards [39]. In fact, encountering uncivil expressions may enhance the recall of opposing viewpoints, potentially fostering deliberation [11]. Furthermore, despite the presence of incivility, individuals were actively engaging in impassioned discussions, presenting evidence, and posing legitimate questions [38]. While the discourse occasionally took on an impolite tone, incivility did not dominate the conversation. This idea is supported in other observations associating incivility with justified opinion expression and genuine engagement in policy disagreements [30]. Moreover, a recent study conducted by Jiyoung Lee et al. indicates that incivility might lead to offline participation [103].

Let's take this one step further. What if the lack of consensus is due to the fact that we failed at step 1: defining toxicity/incivility. Recent studies propose a reevaluation of our initial definitions. Gondwe's [82] exploration of "good incivility" stands out, distinct from character-focused incivility and emphasizing a positive correlation with increased online participation. Another study contributes to this reexamination by scrutinizing discursive and contextual conditions related to interpersonal incivility in comparison to incivility directed at political elites on social media platforms. Rossini contends that the vitriol often directed at politicians is underlined by justified opinions rather than toxic behavior, suggesting that incivility serves as a tool for political critique and opinion justification rather than an inherently problematic feature of online discussions [34].

Overall, while earlier studies predominantly viewed various forms of online incivility as inherently detrimental to democracy, contemporary research has evolved to offer more nuanced conceptualizations, acknowledging at certain points the contributions it could have to deliberation [34, 60, 94].

The key insights on the consequences of online toxicity / incivility thus include:

- Literature reports mixed effects of incivility. It can lead to emotional distress and polarized opinions. However, in certain contexts, incivility can lead to more robust discussions and enhanced understanding of diverse viewpoints.
- The Engagement Paradox; Despite negative connotations, incivility often increases political and social engagement online. Due to online attention wars, incivility is one way to get your message heard.
- Definitional Challenges: Disagreement exists on what constitutes incivility, affecting the interpretation and study of its impacts.

3.6 Addressing online incivility

In the quest to mitigate online incivility within discussions and deliberations, a myriad of strategies have been employed, each with its unique focus and inherent limitations.

One of the earliest methods is the implementation of codes of engagement and civility statements. Clark et al. [8] highlighted strategies including clearly defined expectations in civility statements, faculty role modeling, and immediate addressing of incivility, alongside rewards for civil behavior [104]. Similarly, Sterrett et al. discussed the development of a policy against incivility in an online nursing program.

A broader application is the European Commission's code of conduct to combat online hate speech, engaging major social media platforms in a collaborative effort to uphold digital civility [105]. The 2022 evaluation of this initiative highlights that platforms are actively addressing hate speech notifications. Furthermore, the Santa Clara Principles for Transparency and Accountability in Content Moderation, have garnered support from key industry players like Apple, Facebook (Meta), Google, and Twitter, showcasing a widespread industry commitment to these standards. A 2020 report by InternetLab critically examines the implementation of these principles, voicing concerns over AI moderation's context sensitivity, the necessity for distinct principles in areas such as advertising, and the ongoing need to evolve these principles in response to the changing online environment. Another noteworthy initiative is the Global Internet Forum to Counter Terrorism (GIFCT). It aims to prevent the exploitation of digital platforms by violent extremists. A 2021 assessment report on GIFCT urges for greater inclusivity of voices affected by terrorism and enhanced transparency, for example through the publication of detailed information on operations [106].

Despite these efforts, research indicates that hate speech persists and is on the rise on social media platforms [107]. An analysis involving over 353 million records [107] from the largest social media platforms in the EU, submitted to the DSA Transparency Database, reveals partial compliance with the database's framework, inconsistency in moderation strategies, and raises questions about the reliability of these self-reported actions. This analysis makes apparent the inherent limitations of the implementation of codes of conduct in effectively restricting online toxicity and violent speech, when companies are in charge of reporting how they implement said regulations.

To curb incivility indirectly, some researchers tested design interventions. Park & Singh [59] found that positive imagery, both in color and grayscale, effectively reduces online incivility, indicating the potential for designing more civil discussion platforms. Along the same lines, Elsayed & Hollingshead [108] explored humor as a means to reduce incivility by diminishing anger and increasing affinity toward the author.

A more direct approach to dealing with incivility is the use of moderation. C. Miller et al. described various mechanisms on GitHub to manage uncivil interactions, including issue closing, comment editing, and user blocking [50]. However, these techniques face challenges such as lag time in posting due to heavy moderation in online platforms, can impact the flow of discussions [109]. A promising approach, crowdsourced civility, shows that large-scale civil participation is possible through distributed moderated systems enabling community enforcement of norms [110].

Building on the previous approach, the use of AI for moderation and toxicity detection has been gaining traction in recent years. However, challenges persist, such as the context-dependent nature of incivility and difficulty interpreting nuanced language [44]. Almerekhi et al. demonstrated a model's accuracy in identifying toxicity triggers from

Reddit posts, yet the adoption of automatic detection bots remains rare [50, 85]. This is partly due to the fact that understanding natural language requires a lot of nuance. It is context dependent, varies by culture and by platform norms. And while AI provides a proactive stance, Mall et al. [111] and Xia et al. [80] emphasize the importance of understanding and addressing the triggers and patterns of toxicity.

Some researchers advocate for a shift in focus, questioning the emphasis on eliminating incivility [38, 112]. Since incivility is unlikely to be eliminated completely, the authors argue for an approach that allows communities to define their own civility standards. For instance, they illustrate a platform wherein individuals were barred from commenting until they had assessed the level of incivility in other comments. This crowdsourced method creates an environment conducive to the organic development of norms. Additionally, the focus is shifted from chasing incivility to treating its outcomes like harassment and polarization.

In summary, each approach presents its own limitations. AI and generalized moderation strategies often struggle with context and linguistic nuances. Codes of conduct, while longstanding, may have limited impact in contemporary online environments. Behavioral nudges and platform interventions show promise but require further validation in diverse and real-world settings. The complexity of defining incivility complicates the effectiveness of any single strategy. As we navigate these challenges, it's evident that no singular solution suffices. Instead, a multifaceted and adaptive approach, integrating human insight with technological advancements, is essential.

We must also acknowledge that online platforms are intended to be a space for individuals to freely voice their opinions [62]. In the midst of the attention wars on these platforms, incivility is but a method employed, particularly by marginalized communities, to amplify their voices [38]. Additionally, it can serve as a means to alert others to significant concerns, as mentioned by Rossini [34]. Nevertheless, if this incivility persists over the long term, it leads to severe consequences. An illustrative commentary on this issue discusses the utilization of algorithms by Big Tech firms to enhance user engagement for financial gain. Specifically, in an effort to boost clicks, they inadvertently promote misinformation and polarizing political subjects, thereby diminishing content quality and overall societal well-being.

3.7 Consolidated takeaways from the literature

We begin with an observation that, initially, comments and unstructured forums were heralded as promising platforms to encourage constructive debates online. We acknowledge that not all online discussions are meant to be productive debates, but we do posit that most collective online deliberations happen in these unstructured environments. As the digital landscape matured, the promise of civil discourse was overshadowed by rising levels of toxicity and incivility. This escalation is attributed to multiple factors including emotionally charged topics, platform design that inadvertently breeds toxicity, and the fluid nature of toxic behavior among users, challenging the notion that certain profiles are inherently prone to toxicity.

We argue, based on extant literature, that by nature, the current setup is not conducive to civil deliberation. These platforms often lead to "attention wars" where extreme or toxic language is used to gain visibility, thereby escalating the toxicity levels. The fundamental challenge lies in maintaining an equitable playing field for all participants to have their voices heard, preventing them from resorting to extreme strategies.

Fig. 1 An example of a deliberation map, where: = question, = answer, = pro argument, = con argument, = criterion

Jakob et al., in their 2023 publication, propose that collective deliberations thrive in environments that incentivize individuals to seek compromise, prioritize pertinent topics, and facilitate public discourse while maintaining some separation from purely social interactions [citation]. One real-world example comes from a study examining the unresolved requests for comments (RfCs) on Wikipedia, where they suggested the effectiveness of the deliberation process could be improved with tools that aid participants in organizing their contributions [113]. Similarly, another study developed a theoretical framework for Open Civic Design emphasizing the need for a structured process to cultivate effective solutions [114].

The question then is: Can the transition from general conversation platforms to tools specifically designed for structured deliberation enhance both the tone and quality of discussions? Can we change the deliberation process in a way that prevents toxicity from happening in the first place?

4 Deliberation mapping: a solution to toxicity in online deliberation

The hypothesis explored in this paper is that the toxicity that is so prevalent in conventional online deliberations occurs because the tools that host them (forums, twitter, email, and so on) incorporate no model of what *kind* of conversations will lead crowds to quickly and efficiently find good solutions for complex problems. To test this hypothesis, we did a side-by-side comparison of the toxicity of unstructured conversations vs structured ones.

The structured conversation approach we use is "deliberation mapping" [115], a methodology that engages participants in co-creating *logically-organized knowledge structures* rather than conversation transcripts. As we will see below, the introduction of this structure fundamentally changes the participant incentives and results in a substantial reduction in toxicity. It is important to note however that this approach is aimed at the particular problem of *deliberation*, i.e. where the participants are trying to develop good solutions to particular problem(s). While better large-deliberation is of course an extremely important challenge, many online conversations are *not* solution-oriented

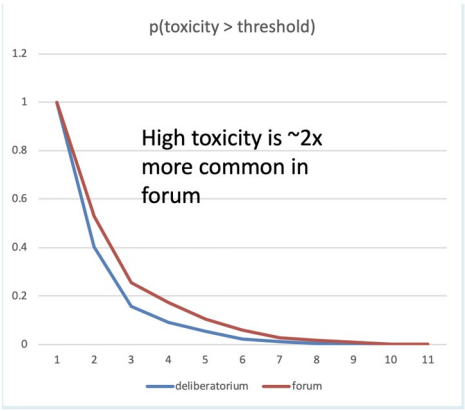


Fig. 2 Cumulative probability of posts above the given toxicity threshold, forum vs Deliberatorium

(e.g. they may simply be aimed at entertainment and socialization) and a structured deliberation approach would thus probably not be germane (Figures 1 and 2).

This work applied a deliberation mapping system called the "Deliberatorium" [116]. It represents the simplest form of deliberation map that, in our experience, enables effective crowd-scale deliberation. Our map schema is built of "QuAACRs", i.e. *questions* to be answered, possible *answers* for these questions, *criteria* that describe the attributes of good answers, *arguments* that support or rebut an answer or argument, and *ratings* that capture the importance of questions and criteria, the value of answers, and the strength of arguments:

Deliberation maps have many important advantages over conversation-centric approaches. All the points appear in the part of the map they logically belong to, e.g. all answers to a question are attached to that question in the map. It is therefore easy to find all the crowd's input on any given question, since it is collocated in the same branch. It's also easy to check if a point has already been contributed, and therefore to avoid *repeating* points, radically increasing the signal-to-noise ratio. Detecting and avoiding redundancy can in fact be mostly automated by the use of semantic similarity assessment tools based on text embedding technology [117]. *Gaps* in the deliberation—e.g. questions without any answers, or answers without any arguments—are easy to identify, so we can guide crowd members to fill in these gaps and foster more complete coverage. Making arguments into first-class map entities implicitly encourages participants to express the evidence and logic for or against competing answers [118], and means that arguments can be critiqued individually. Users, finally, can easily collaborate to refine proposed solutions. One user can, for example, propose an answer, a second raise a question about how that answer can achieve a given requirement, and a third propose possible answers for that sub-question.

Why should this approach reduce toxicity? As was pointed out by media theorist Marshall McLuhan in his 1964 book *Understanding Media* [119], the nature of the discussion medium we use can have a profound impact on *what* we communicate. In a sense, as he points out, the medium *is* the message. How, then, do online discussion media shape what we say? In such tools, one of the key questions for participants is: how do I win the **attention war** as new posts pile on? Our inputs can easily be overlooked unless we frame them in ways that are likely to gather more attention. One guaranteed way to do that is to be more *extreme/toxic* than the others in the discussion. But

if most people follow this individually rational strategy, the result is an upward toxicity spiral as contributors become more extreme in order to compete with other people using the same strategy.

Deliberation maps have different rules that in turn change the incentives (and thus typical behaviors) for contributors. Participants no longer need to engage in extremization in order to make themselves visible. Everybody's points on a given topic are co-located, right next to each other, and every unique idea appears just once, regardless of when or how often it was contributed. Deliberation maps make it immediately visible whether an individuals' postings have underlying (PRO or CON) arguments from the original poster and other participants. The "game" therefore changes *from* simply trying to get attention in a massive growing comment pile *to* creating points that people find compelling. In this context, less extremized and more carefully-argued points, we hypothesize, are instead more likely to receive positive evaluations. This suggests that toxicity in deliberation maps will be significantly less than that in conventional (conversation-centric) forums.

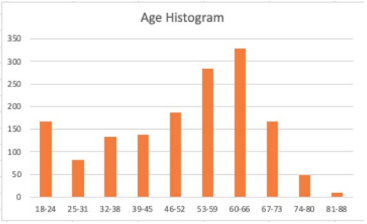
5 Experimental evaluation

We assessed the toxicity of the posts contributed in a random controlled trial consisting of two demographically matched experimental conditions of over 400 participants each:

- *Forum*: Participants used a forum (AKA threaded discussion) to submit posts as well as reply to other posts The posts and subsequent multiple levels of replies were viewed as an indented outline. Since users can contribute any kind of posts at any time, we considered this the "unstructured" condition.
- *Deliberatorium:* Participants used the Deliberatorium system, described above, to post questions answers and arguments in response to the newspaper articles. Since users are asked to contribute posts in a specific format (i.e. as questions answers and arguments in a logically-organized "map"), we considered this the "structured" condition.

The participants were recruited using ads on a range of social media platforms including Facebook. The participants were demographically matched across the two conditions, as much as possible, with respect to the participants' age, gender, and country of origin:

Age



Average: 51.5 years
Standard Deviation: 16.3 years
Minimum: 18 years
Maximum: 88 years

Gender

Female: 347
Male: 435
Other: 18

Country of Origin

United Kingdom	284
United States	198
Australia	105
Canada	84
New Zealand	30
Ireland	26
Italy, Germany	8 each
Afghanistan	6
Netherlands	4
Mexico, Belgium, Switzerland, Brazil, Poland, Pakistan, France, Lebanon, Argentina, India Luxembourg, Thailand, Bangladesh, Austria, China, Georgia, Taiwan, Philippines, Spain, Nicaragua, Azerbaijan, Faroe Islands, Portugal, Uruguay, Finland, Somalia, Antarctica, Hungary, Kenya, Iceland, Honduras, Costa Rica, Uzbekistan, Trinidad & Tobago, Iraq, Turkey, Russia, Kazakhstan, Denmark, South Africa, Sri Lanka, Peru, Sweden, Kuwait, Rwanda, Egypt, Palestinian Territories, Malaysia	1 or 2 each

Participants in each condition were asked to discuss, using their assigned tools, the content of the following eight newspaper articles (used with permission from the New York Times):

Finding Compassion for 'Vaccine-Hesitant' Parents
By Wajahat Ali
We've All Just Made Fools of Ourselves — Again
By David Brooks
Why Are Young People Pretending to Love Work?
By Erin Griffith
New Zealand Massacre Highlights Global Reach of White Extremism
By Patrick Kingsley
The India-Pakistan Conflict Was a Parade of Lies.
By Farhad Manjoo

The West Doesn't Want ISIS Members to Return. Why Should the Syrians Put Up With Them?
By Abdalaziz Alhamza 3/14/2019 at 10:52:22 pm
Britain Is Drowning Itself in Nostalgia
By Sam Byers 3/24/2019 at 4:34:26 pm
If Stalin Had a Smartphone.
By David Brooks 3/13/2019 at 0:19:42 am

Participants in both conditions were asked to discuss all eight articles. The discussions remained open for two weeks, and none of the participants were compensated. Neither of these conditions were moderated: so participants were free to take any tone they chose in their postings.

In the Deliberatorium condition, participants were asked to add new posts to the map for the article they were currently discussing. Clicking on a question allowed them to add an answer as a response. Clicking on an answer or argument allowed them to add a PRO or CON argument as a response. When adding responses, users were presented with textual clues encouraging them to follow the system schema, which means entering text of the right type (answer, pro or con) as well as including only one thought in each response. While we did not include a mechanism that enforced this schema, we found that over 80% of participants followed it correctly.

We used the Google Perspective API (https://www.perspectiveapi.com/) to assess the toxicity of the posts from the two conditions on a scale from 0 (non-toxic) to 1 (highly toxic). While, as noted above, the Perspective API is imperfect, it is the acknowledged state-of-the-art tool for this purpose and is widely used. Our spot-checking of the Perspective API scores satisfied us that it does a good job of detecting overt toxicity, though it was less accurate at detecting more subtle forms of incivility such as irony or sarcasm. We deemed this acceptable since outright toxicity appears to have a much more pronounced negative effect on participation than irony and the like.

The average toxicity for the posts generated in the two conditions was as follows:

Platform	# posts	Average Toxicity	Standard deviation
forum	915	0.19	0.16
deliberatorium	812	0.14	0.12

While the overall toxicity levels were relatively low in our community, the average toxicity of the forum posts was 30% higher than the deliberation map posts: this difference was highly significant statistically ($p < 1.5 * 10^{-10}$) as assesses by a two-tailed T-test. We also found that high toxicity posts (i.e. with toxicity scores above 0.3) were twice as common in the forums than in the deliberation maps (also, of course, highly significant statistically).

6 Discussion

Toxicity has emerged as one of the major challenges for those who hope to enable useful crowd-scale online deliberations around complex and contentious topics. Our work has demonstrated that the level of toxicity in online discussions is deeply affected by the *way in which the discussions take place*. The structured nature of deliberation mapping, we

believe, changes the rules of the game in a way that makes toxic comments no longer part of a winning strategy. Our data provides initial support for this hypothesis, based on a carefully designed randomized control trial experiment involving over 800 participants.

This approach is not, however, a panacea. We have shown that structured discussions can enable a large drop in toxicity in the particular case of *deliberation*, i.e. where the participants are trying to develop good solutions to particular problem(s). Many online conversations are *not* solution-oriented and a structured deliberation approach would thus probably not be germane.

For future work, we would like to reproduce these experiments with communities and topics where the base toxicity level in the online forums is substantially higher, so we can assess the power of structuring conversations on reducing toxicity in more severely challenging contexts.

Acknowledgements The authors would like to acknowledge the contributions of Arnab Sircar and Paolo Spada to this work. The authors also gratefully acknowledge the financial support provided by the Templeton Foundations' "Intellectual Humility" grant program.

Author contributions MK ran the study, which included writing the deliberation software, designing and running the experiment, analyzing the data, and writing most of the paper sections. NM wrote the extensive literature review on online toxicity.

Funding 'Open Access funding provided by the MIT Libraries' This work was supported by the Templeton Foundations' "Intellectual Humility" grant program.

Data Availability The data collected from this study, with anonymized subject IDs, is available upon request from the corresponding author.

Declarations

Competing interests Dr. Klein is chief scientist and stockholder for a small startup (HiveWise LLC) that is commercializing an earlier version of the software described in this paper.

Ethical Approval This research was granted "exempt" status by the MIT IRB.

Participants were presented with the following statement when entering the system:
Use of this system is entirely voluntary, and you can stop at any time. Data from the system may be analyzed as part of ongoing research on how to better support large-scale online deliberation. Proper measures will be taken to safeguard data, and all data used in this analysis will be anonymized. Logging in to this system indicates that you consent to participate in this research.
Participation was uncompensated. The subjects were adults. We did not collect personal information on the subjects, and the data has been stored on secured servers.

References

1. Klein, M., Convertino, G.: An embarrassment of riches: A critical review of open innovation systems. ACM, Commun (2014)

2. Shirky, C.: Here Comes Everybody: The Power of Organizing Without Organizations. Penguin Press, (2008)
3. Tapscott, D., Williams, A. D.: Wikinomics: How Mass Collaboration Changes Everything. Portfolio, (2006)
4. Jeppesen, L.B., Lakhani, K.R.: Marginality and Problem-Solving Effectiveness in Broadcast Search. Organ. Sci. **21**(5), 1016–1033 (2010)
5. Gulley, N.: Patterns of innovation: a web-based MATLAB programming contest. In: CHI'01 extended abstracts on human factors in computing systems, pp. 337–338. ACM, Seattle (2001). https://doi.org/10.1145/634067.634266
6. Surowiecki, J.: The Wisdom of Crowds. Knopf Doubleday Publishing Group, (2005)
7. Raymond, E.: The cathedral and the bazaar. Knowl. Technol. Policy **12**(3), 23–49 (1999). https://doi.org/10.1007/s12130-999-1026-0
8. Clark, C.M., Ahten, S., Werth, L.: Cyber-Bullying and Incivility in an Online Learning Environment, Part 2: Promoting Student Success in the Virtual Classroom. Nurse Educ. **37**(5), 192 (2012). https://doi.org/10.1097/NNE.0b013e318262eb2b
9. Obadimu, A., Mead, E., Agarwal, N.: Identifying latent toxic features on YouTube using non-negative matrix factorization. In: The ninth international conference on social media technologies, communication, and informatics. IEEE (2019)
10. Chipidza, W.: The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach. Int. J. Inf. Manag. **61**, 102397 (2021). https://doi.org/10.1016/j.ijinfomgt.2021.102397
11. Hwang, H., Kim, Y.: Influence of Discussion Incivility on Deliberation: An Examination of the Mediating Role of Moral Indignation, Commun. Res., vol. 45, (2016), https://doi.org/10.1177/0093650215616861
12. Bormann, M.: Perceptions and evaluations of incivility in public online discussions—insights from focus groups with different online actors, Front. Polit. Sci., vol. 4. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpos.2022.812145. (2022). Accessed 18 Nov 2023
13. Frischlich, L., Schatto-Eckrodt, T., Boberg, S., Wintterlin, F.: Roots of Incivility: How Personality, Media Use, and Online Experiences Shape Uncivil Participation. Media Commun. **9**(1), 195–208 (2021). https://doi.org/10.17645/mac.v9i1.3360
14. Bormann, M., Tranow, U., Vowe, G., Ziegele, M.: Incivility as a Violation of Communication Norms—A Typology Based on Normative Expectations toward Political Communication. Commun. Theory **32**(3), 332–362 (2021). https://doi.org/10.1093/ct/qtab018
15. Coe, K., Kenski, K., Rains, S.A.: Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J. Commun. **64**(4), 658–679 (2014). https://doi.org/10.1111/jcom.12104
16. Friess, D., Eilders, C.: A Systematic Review of Online Deliberation Research, Policy Internet, vol. 7, (2015), https://doi.org/10.1002/poi3.95
17. Ruiz, C., Domingo, D., Micó, J.L., Díaz-Noci, J., Meso, K., Masip, P.: Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. Int. J. Press. **16**(4), 463–487 (2011). https://doi.org/10.1177/1940161211415849
18. Thiele, D., Turnšek, T.: How Right-Wing Populist Comments Affect Online Deliberation on News Media Facebook Pages. Media Commun. **10**(4), 141–154 (2022)
19. Jost, P., Ziegele, M.: How to get on top – the effect of rationality and incivility of user comments on their visibility in political online discussions on Facebook. Commun. Res. Rep. **39**(4), 224–235 (2022). https://doi.org/10.1080/08824096.2022.2120861
20. Collins, L., Nerlich, B.: Examining User Comments for Deliberative Democracy: A Corpus-driven Analysis of the Climate Change Debate Online. Environ. Commun. **9**(2), 189–207 (2015). https://doi.org/10.1080/17524032.2014.981560
21. Ziegele, M., Quiring, O., Esau, K., Friess, D.: Linking news value theory with online deliberation: How news factors and illustration factors in news articles affect the deliberative quality of user discussions in SNS' comment sections, Commun. Res., vol. 47, no. 6, Art. no. 6, (2020) https://doi.org/10.1177/0093650218797884
22. Rowe, I.: Civility 2.0: A comparative analysis of incivility in online political discussion. Inf. Commun. Soc. **18**(2), 121–138 (2015). https://doi.org/10.1080/1369118X.2014.940365
23. Santana, A.D.: Toward quality discourse: Measuring the effect of user identity in commenting forums. Newsp. Res. J. **40**(4), 467–486 (2019). https://doi.org/10.1177/0739532919873089
24. Friess, D., Ziegele, M., Heinbach, D.: Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. Polit. Commun. (2020). https://doi.org/10.1080/10584609.2020.1830322

25. Sarmento, M.: Disrespect in Online Deliberation: Inducing Factors and Democratic Potentials, (2016). https://doi.org/10.4067/S0718-090X2016000300005
26. Galarza Molina, R., Jennings, F.: The Role of Civility and Metacommunication in Facebook Discussions, Commun. Stud., pp. 1–25, (2017), https://doi.org/10.1080/10510974.2017.1397038
27. Civility in America: Solutions for tomorrow, Weber Shandwick. [Online]. Available: https://cms.webershandwick.com/news/civility-in-america-2019-solutions-for-tomorrow/(2019). Accessed 5 Dec 2023
28. Trifiro, B.M., Paik, S., Fang, Z., Zhang, L.: Politics and Politeness: Analysis of Incivility on Twitter During the 2020 Democratic Presidential Primary. Soc. Media Soc. **7**(3), 20563051211036940 (2021). https://doi.org/10.1177/20563051211036939
29. Lück, J., Nardi, C.: Incivility in user comments on online news articles: Investigating the role of opinion dissonance for the effects of incivility on attitudes, emotions and the willingness to participate. Stud. Commun. Media **8**(3), 311–337 (2019). https://doi.org/10.5771/2192-4007-2019-3-311
30. Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., Lecheler, S.: Beyond Anonymity: Network Affordances, Under Deindividuation, Improve Social Media Discussion Quality. J. Comput.-Mediat. Commun. **27**(1), zmab019 (2022). https://doi.org/10.1093/jcmc/zmab019
31. Chen, G. M.: Online Incivility and Public Debate: Nasty Talk. Springer, (2017). https://doi.org/10.1007/978-3-319-56273-5
32. Anderson, A.A., Brossard, D., Scheufele, D.A., Xenos, M.A., Ladwig, P.: The 'Nasty Effect:' Online Incivility and Risk Perceptions of Emerging Technologies. J. Comput.-Mediat. Commun. **19**(3), 373–387 (2014). https://doi.org/10.1111/jcc4.12009
33. Gervais, B.: Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment. J. Inf. Technol. Polit. (2015). https://doi.org/10.1080/19331681.2014.997416
34. Rossini, P.: More Than Just Shouting? Distinguishing Interpersonal-Directed and Elite-Directed Incivility in Online Political Talk. Soc. Media Soc. **7**(2), 20563051211008828 (2021). https://doi.org/10.1177/20563051211008827
35. Jost, P., Köhler, C.: Warum so garstig? Zum Einfluss von realweltlichen, medien- und diskussionsimmanenten sowie situativen Faktoren auf die (In)Zivilität von Onlinediskussionen, in Politische Partizipation im Medienwandel, vol. 6, I. Engelmann, M. Legrand, and H. Marzinkowski, Eds., in Digital Communication Research, vol. 6. , Berlin, (2019), pp. 321–344. https://doi.org/10.17174/dcr.v6.13
36. Daniil Volkovskii, "(PDF) Low Civility and High Incivility in Russian Online Deliberation: A Case of Political Talk in Vkontakte Social Network." [Online]. Available: https://www.researchgate.net/publication/372273776_Low_Civility_and_High_Incivility_in_Russian_Online_Deliberation_A_Case_of_Political_Talk_in_Vkontakte_Social_Network (2023). Accessed 18 Nov 2023
37. Gervais, B.T.: Rousing the Partisan Combatant: Elite Incivility, Anger, and Antideliberative Attitudes. Polit. Psychol. **40**(3), 637–655 (2019). https://doi.org/10.1111/pops.12532
38. Masullo Chen, G., Riedl, M.J., Shermak, J.L., Brown, J., Tenenboim, O.: Breakdown of Democratic Norms? Understanding the 2016 US Presidential Election Through Online Comments. Soc. Media Soc. **5**(2), 2056305119843637 (2019). https://doi.org/10.1177/2056305119843637
39. Rösner, L., Krämer, N.C.: Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. Soc. Media Soc. **2**(3), 2056305116664220 (2016). https://doi.org/10.1177/2056305116664220
40. Won Kim, J., Park, S.: "How perceptions of incivility and social endorsement in online comments (Dis) encourage engagements. Behav. Inf. Technol. **38**(3), 217–229 (2019). https://doi.org/10.1080/0144929X.2018.1523464
41. Alshamrani, S., Abuhamad, M., Abusnaina, A., Mohaisen, D.: Investigating online toxicity in users interactions with the mainstream media channels on YouTube. In: Proceedings from CIKM 2020 workshops. Galway, Ireland (2020)
42. Anderson, A., Huntington, H.: Social Media, Science, and Attack Discourse: How Twitter Discussions of Climate Change Use Sarcasm and Incivility. Sci. Commun. **39**, 598–620 (2017). https://doi.org/10.1177/1075547017735113
43. Oh, D., Elayan, S., Sykora, M.: Deliberative Qualities of Online Abortion Discourse: Incivility and Intolerance in the American and Irish Abortion Discussions on Twitter. J. Deliberative Democr., vol. 19, no. 1 (2023), https://doi.org/10.16997/jdd.1413
44. Stevens, H., Acic, I., Taylor, L.D.: Uncivil Reactions to Sexual Assault Online: Linguistic Features of News Reports Predict Discourse Incivility. Cyberpsychology Behav. Soc. Netw. **24**(12), 815–821 (2021). https://doi.org/10.1089/cyber.2021.0075
45. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean Birds: Detecting Aggression and Bullying on Twitter," in Proceedings of the 2017 ACM on Web

Science Conference, in WebSci '17. New York, NY, USA: Association for Computing Machinery, (2017), pp. 13–22. https://doi.org/10.1145/3091478.3091487
46. Salminen, J., Sengün, S., Corporan, J., Jung, S., Jansen, B.J.: Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. PLoS ONE **15**(2), e0228723 (2020). https://doi.org/10.1371/journal.pone.0228723
47. Santana, A.: Incivility Dominates Online Comments on Immigration. Newsp. Res. J. **36**, 92–107 (2015). https://doi.org/10.1177/073953291503600107
48. Mittos, A., Zannettou, S., Blackburn, J., Cristofaro, E.D.: Analyzing Genetic Testing Discourse on the Web Through the Lens of Twitter, Reddit, and 4chan. ACM Trans. Web **14**(4), 17:1-17:38 (2020). https://doi.org/10.1145/3404994
49. Tsuchiya, T., Cuevas, A., Magelinski, T., Christin, N.: Misbehavior and Account Suspension in an Online Financial Communication Platform, in Proceedings of the ACM Web Conference 2023, in WWW '23. New York, NY, USA: Association for Computing Machinery, (2023), pp. 2686–2697. https://doi.org/10.1145/3543507.3583385
50. Miller, C., Cohen, S., Klug, D., Vasilescu, B., Kästner, C.: Did You Miss My Comment or What?' Understanding Toxicity in Open Source Discussions, in 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), (2022), pp. 710–722. https://doi.org/10.1145/3510003.3510111
51. Rossini , P. Maia, R.: Characterizing Disagreement in Online Political Talk: Examining Incivility and Opinion Expression on News Websites and Facebook in Brazil, Regul. Issue, vol. 17, no. 1, Art. no. 1, (2021). https://doi.org/10.16997/10.16997/jdd.967
52. Chen, Y., Wang, L.: Misleading political advertising fuels incivility online: A social network analysis of 2020 U.S. presidential election campaign video comments on YouTube. Comput. Hum. Behav. **131**, 107202 (2022). https://doi.org/10.1016/j.chb.2022.107202
53. Stromer-Galley, J., Bryant, L., Bimber, B.: Context and Medium Matter: Expressing Disagreements Online and Face-to-Face in Political Deliberations, J. Deliberative Democr., vol. 11, no. 1, Art. no. 1, (2015), https://doi.org/10.16997/jdd.218
54. Rega, R., Marchetti, R.: The strategic use of incivility in contemporary politics. The case of the 2018 Italian general election on Facebook. Commun. Rev. **24**(2), 107–132 (2021). https://doi.org/10.1080/10714421.2021.1938464
55. Hopp, T., Vargo, C.J.: Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. Comput. Hum. Behav. **68**, 368–377 (2017). https://doi.org/10.1016/j.chb.2016.11.034
56. Wolf, A., Foxman, A.H.: Viral Hate: Containing Its Spread on the Internet. By Abraham H. Foxman and. J. Contemp. Antisemitism **2**(1), 87–92 (2013). https://doi.org/10.26613/jca/2.1.26
57. Murthy, D., Sharma, S.: Visualizing YouTube's comment space: online hostility as a networked phenomena. New Media Soc. **21**(1), 191–213 (2019). https://doi.org/10.1177/1461444818792393
58. Farid, H., Hasan, S.J., Naveed, A., Hyder, P.R., Shaikh, G.M., Pasha, L.: Incivility in online learning environment: Perception of dental students and faculty. J. Dent. Educ. **86**(12), 1591–1601 (2022). https://doi.org/10.1002/jdd.13031
59. Park, J., Singh, V.K.: How Background Images Impact Online Incivility. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2), 444:1-444:23 (2022). https://doi.org/10.1145/3555545
60. Ng, Y. L., Song, Y., Kwon, K. H., Huang, Y.: Toward an integrative model for online incivility research: A review and synthesis of empirical studies on the antecedents and consequences of uncivil discussions online. Telemat. Inform., vol. 47, no. 101323, (2020), https://doi.org/10.1016/j.tele.2019.101323
61. Schroll, C., Huber, B.: Assessing levels and forms of incivility and deliberative quality in online discussions on COVID-19: A Cross-Platform Analysis, Front. Polit. Sci., vol. 4, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpos.2022.814002 (2022). Accessed 18 Nov 2023
62. Seely, N.: Virtual Vitriol: A Comparative Analysis of Incivility Within Political News Discussion Forums. Electron. News **12**, 193124311773906 (2017). https://doi.org/10.1177/1931243117739060
63. Oz, Pei Zheng, M., Masullo Chen, G.: Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes - Mustafa Oz, Pei Zheng, Gina Masullo Chen, Accessed: Dec. 04, 2023. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/1461444817749516 (2018)
64. DiCicco, K., et al.: Toxicity and Networks of COVID-19 discourse communities: a tale of two social media platforms. In: Proceedings of ceur-ws.org ISSN 1613 (2020). https://doi.org/10.48550/arXiv.2302.14270
65. Halpern, D., Gibbs, J.: Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. Comput. Hum. Behav. **29**(3), 1159–1168 (2013). https://doi.org/10.1016/j.chb.2012.10.008

66. Suler, J.: The online disinhibition effect, Cyberpsychology Behav. Impact Internet Multimed. Virtual Real. Behav. Soc. **7**(3), 321–326 (2004). https://doi.org/10.1089/1094931041291295
67. Rowe, I.: Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. J. Broadcast. Electron. Media **59**(4), 539–555 (2015). https://doi.org/10.1080/08838151.2015.1093482
68. Su, L.Y.-F., Xenos, M.A., Rose, K.M., Wirz, C., Scheufele, D.A., Brossard, D.: Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. New Media Soc. **20**(10), 3678–3699 (2018). https://doi.org/10.1177/1461444818757205
69. Rabab'ah, G., Alali, N.: Impoliteness in reader comments on the Al-Jazeera channel news website. J. Politeness Res. **16**(1), 1–43 (2020). https://doi.org/10.1515/pr-2017-0028
70. Kim, Y.: Potentials and Limitations of Computer-Mediated Communication Theories for Online Incivility Research: A Focus on Bystander Dynamics, (2022). https://doi.org/10.24251/HICSS.2022.761
71. Almerekhi, H., Kwak, H., Jansen, B. J., Salminen, J.: Detecting Toxicity Triggers in Online Discussions, in Proceedings of the 30th ACM Conference on Hypertext and Social Media, in HT '19. New York, NY, USA: Association for Computing Machinery, (2019), pp. 291–292. https://doi.org/10.1145/3342220.3344933
72. Hmielowski, J.D., Hutchens, M.J., Cicchirillo, V.J.: Living in an age of online incivility: examining the conditional indirect effects of online discussion on political flaming. Inf. Commun. Soc. **17**(10), 1196–1211 (2014). https://doi.org/10.1080/1369118X.2014.899609
73. Sobieraj, S., Berry, J.M.: From Incivility to Outrage: Political Discourse in Blogs, Talk Radio, and Cable News. Polit. Commun. **28**(1), 19–41 (2011). https://doi.org/10.1080/10584609.2010.542360
74. Klein, O., Spears, R., Reicher, S.: Social Identity Performance: Extending the Strategic Side of SIDE. Personal. Soc. Psychol. Rev. **11**(1), 28–45 (2007). https://doi.org/10.1177/1088868306294588
75. Lewandowska-Tomaszczyk, B.: Incivility and confrontation in online conflict discourses. Lodz Pap. Pragmat., vol. 13, no. 2, (2017), https://doi.org/10.1515/lpp-2017-0017
76. Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P.K., Zollo, F.: Dynamics of online hate and misinformation. Sci. Rep. **11**(1), 22083 (2021). https://doi.org/10.1038/s41598-021-01487-w
77. Shmargad, Y., Coe, K., Kenski, K., Rains, S.A.: Social Norms and the Dynamics of Online Incivility. Soc. Sci. Comput. Rev. **40**(3), 717–735 (2022). https://doi.org/10.1177/0894439320985527
78. Heseltine, M., Dorsey, S.: Online Incivility in the 2020 Congressional Elections. Polit. Res. Q. **75**(2), 512–526 (2022). https://doi.org/10.1177/10659129221078863
79. Graham, T., Wright, S.: Discursive Equality and Everyday Talk Online: The Impact of 'Superparticipants'*. J. Comput.-Mediat. Commun. **19**(3), 625–642 (2014). https://doi.org/10.1111/jcc4.12016
80. Xia, Y., Zhu, H., Lu, T., Zhang, P., Gu, N.: Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. Proc. ACM Hum.-Comput. Interact. **4**(CSCW2), 108:1-108:23 (2020). https://doi.org/10.1145/3415179
81. Hansen, R.W.: You've never been welcome here: exploring the relationship between exclusivity and incivility in online forums. J. Inf. Technol. Polit. **20**(2), 139–153 (2023). https://doi.org/10.1080/19331681.2022.2069180
82. Gondwe, G.: Online incivility, hate speech, and political violence in Zambia: Examining the role of online political campaign messages. J. Afr. Media Stud. **13**, 35–51 (2021). https://doi.org/10.1386/jams_00032_1
83. Keller, T.R., Klinger, U.: Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications. Polit. Commun. **36**(1), 171–189 (2019). https://doi.org/10.1080/10584609.2018.1526238
84. Qayyum, H., Zi Hao Zhao, B., Wood, I., Ikram, M., Kourtellis, N., Ali Kaafar, M.: A longitudinal study of the top 1% toxic Twitter profiles," in Proceedings of the 15th ACM Web Science Conference 2023, in WebSci '23. New York, NY, USA: Association for Computing Machinery, (2023), pp. 292–303. https://doi.org/10.1145/3578503.3583619
85. Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S.-G., Jansen, B. J.: Four Types of Toxic People: Characterizing Online Users' Toxicity over Time, in Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, in NordiCHI '20. New York, NY, USA: Association for Computing Machinery, (2020), pp. 1–11. https://doi.org/10.1145/3419249.3420142
86. Jakob, J., Dobbrick, T., Freudenthaler, R., Haffner, P., Wessler, H.: Is Constructive Engagement Online a Lost Cause? Toxic Outrage in Online User Comments Across Democratic Political Systems and Discussion Arenas. Commun. Res. **50**(4), 508–531 (2023). https://doi.org/10.1177/00936502211062773

87. Humprecht, E., Hellmueller, L., Lischka, J.A.: Hostile Emotions in News Comments: A Cross-National Analysis of Facebook Discussions. Soc. Media Soc. **6**(1), 2056305120912481 (2020). https://doi.org/10.1177/2056305120912481
88. Volkovskii, D., Filatova, O., Bolgov, R.: Social media deliberation: civil or uncivil, reasoned or unreasoned?," in Proceedings of the Central and Eastern European eDem and eGov Days, Budapest Hungary: ACM, (2022), pp. 6–11. https://doi.org/10.1145/3551504.3551512
89. Brokensha, S.I., Conradie, M.S.: (In)civility and online deliberation: readers' reactions to race-related news stories. Safundi **18**(4), 327–348 (2017). https://doi.org/10.1080/17533171.2017.1335000
90. Min, C., Shen, F.: Online incivility, argument quality and public expression in China: Exploring the moderating role of education level and opinion congruency. Telemat. Inform. **82**, 102010 (2023). https://doi.org/10.1016/j.tele.2023.102010
91. Jiang, M., Esarey, A.: Uncivil society in digital China: Incivility, fragmentation, and political stability. Int. J. Commun. 12 (2018)
92. Szabo, G., Kmetty, Z., Molnár, E.K.: Politics and Incivility in the Online Comments: What is Beyond the Norm-Violation Approach? Int. J. Commun. **15**, 1659–1684 (2021)
93. Lu, S., Liang, H., Masullo, G. M.: Selective Avoidance: Understanding How Position and Proportion of Online Incivility Influence News Engagement. OSF Preprints, (2022). https://doi.org/10.31219/osf.io/syan5
94. Miller, M.L., Vaccari, C.: Digital Threats to Democracy: Comparative Lessons and Possible Remedies. Int. J. Press. **25**(3), 333–356 (2020). https://doi.org/10.1177/1940161220922323
95. Yamamoto, M., Dalisay, F., Kushin, M. J.: An examination of uncivil and reasoned comments and perceived civility in politics. Int. J. Commun. **14**, 279–298 (2020)
96. Druckman, J.N., Gubitz, S.R., Lloyd, A.M., Levendusky, M.S.: How Incivility on Partisan Media (De)Polarizes the Electorate. J. Polit. (2019). https://doi.org/10.1086/699912
97. Masullo, G.M., Tenenboim, O., Lu, S.: 'Toxic atmosphere effect': Uncivil online comments cue negative audience perceptions of news outlet credibility. Journalism **24**(1), 101–119 (2023). https://doi.org/10.1177/14648849211064001
98. Sobieraj, S.: Credible Threat: Attacks Against Women Online and the Future of Democracy. Oxford University Press (2020). https://doi.org/10.1093/oso/9780190089283.001.0001
99. Cover, R.: Digital hostility, subjectivity and ethics: Theorising the disruption of identity in instances of mass online abuse and hate speech. Convergence **29**(2), 308–321 (2023). https://doi.org/10.1177/13548565221122908
100. Rains, S.A., Kenski, K., Coe, K., Harwood, J.: Incivility and Political Identity on the Internet: Intergroup Factors as Predictors of Incivility in Discussions of News Online. J. Comput.-Mediat. Commun. **22**(4), 163–178 (2017). https://doi.org/10.1111/jcc4.12191
101. Frimer, J.A., et al.: Incivility Is Rising Among American Politicians on Twitter. Soc. Psychol. Personal. Sci. **14**(2), 259–269 (2023). https://doi.org/10.1177/19485506221083811
102. Mathew, B. et al.: Thou shalt not hate: Countering Online Hate Speech. arXiv, Apr. (2019). https://doi.org/10.48550/arXiv.1808.04409
103. Lee, J., Choi, J., Kim, J.: Effects of online incivility and emotions toward in-groups on cross-cutting attention and political participation: Behav. Inf. Technol.: Vol 41, No 14. Accessed: Nov. 20, 2023. [Online]. Available: https://doi.org/10.1080/0144929X.2021.1969429(2023)
104. Sterrett, S., Spadaro, K., Walter, L., Wasco, J., Hopkins, E., Fisher, M.: Incivility in the Online Classroom: A Guide for Policy Development, Nurs. Forum (Auckl.), vol. 52, (2017), https://doi.org/10.1111/nuf.12205
105. Code of Conduct on Countering Illegal Hate Speech Online. Downloaded from https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985 in March 2024
106. 2022 GIFCT Transparency Report. Downloaded from https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf in March 2024
107. Nazmine, Manan, K., Tareen, H.K, Noreen, S., Tariq, M.: Hate speech and social media: a systematic review. Turk. Online J. Qual. Inq. **12**, 5285–5294 (2021)
108. Elsayed, Y., Hollingshead, A.B.: Humor Reduces Online Incivility. J. Comput.-Mediat. Commun. **27**(3), zmac005 (2022). https://doi.org/10.1093/jcmc/zmac005
109. Popescu, D., Loveland, M.: Judging Deliberation An Assessment of the Crowdsourced Icelandic Constitutional Project. J. Deliberative Democr. **18**(1), 1 (2022). https://doi.org/10.16997/jdd.974
110. Lampe, C., Zube, P., Lee, J., Park, C.H., Johnston, E.: Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. Gov. Inf. Q. **31**(2), 317–326 (2014). https://doi.org/10.1016/j.giq.2013.11.005
111. Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S.-G., Jansen, B. J.: Four Types of Toxic People: Characterizing Online Users' Toxicity over Time, in Proceedings of the 11th Nordic

Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, Tallinn Estonia: ACM, (2020), pp. 1–11. https://doi.org/10.1145/3419249.3420142
112. Masullo Chen G., Muddiman, A., Wilner, T., Pariser, E., Stroud N.J.: We should not get rid of incivility online. Soc. Media Soc. **5**(3), 2056305119862641 (2019). https://doi.org/10.1177/2056305119862641
113. Im, J., Zhang, A. X., Schilling, C. J., Karger, D.: Deliberation and resolution on wikipedia: A case study of requests for comments. Proceedings of the ACM on Human-Computer Interaction, 2(74) (2018). https://doi.org/10.1145/3274343
114. Reynante, B., Dow, S.P., Mahyar, N.: A Framework for Open Civic Design: Integrating Public Participation, Crowdsourcing, and Design Thinking. In Digital Government: Research and Practice. **4**(31), 1–22 (2021). https://doi.org/10.1145/3487607
115. Shum, S.J.B., Selvin, A.M., Sierhuis, M., Conklin, J., Haley, C.B.: Hypermedia support for argumentation-based rationale. In: Dutoit, A.H., McCall, R., Mistrík, I., Paech, B. (eds.) Rationale management in software engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30998-7_5
116. Klein, M.: How to Harvest Collective Wisdom for Complex Problems: An Introduction to the MIT Deliberatorium. (2007). https://doi.org/10.13140/RG.2.2.32743.24489
117. Ravichandiran, S.: Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing, (2021)
118. Carr, C.: Using Computer Supported Argument Visualization to Teach Legal Argumentation, (2003), https://doi.org/10.1007/978-1-4471-0037-9_4
119. McLuhan, M.: Understanding Media: The Extensions of Man. Gingko Press, (2003)