OXFORD

Interacting with Computers, 2024, 1–19
https://doi.org/10.1093/iwc/iwae036
Article

Examining the Impact of Digital Jury Moderation on the Polarization of U.S. Political Communities on Social Media

Christopher Micek* and Erin T. Solovey

Department of Computer Science, Worcester Polytechnic Institute (WPI), 100 Institute Road, Worcester, MA, USA, 01609-2280
*Corresponding author: cjmicek@wpi.edu

The increased prevalence of misinformation and inflammatory rhetoric online has amplified polarization on social media platforms in the United States, propelling a feedback loop resulting in the erosion of democratic norms. We conducted a study assessing how a social media platform employing appointed moderators would impact the polarization of its users compared to a peer-based *digital jury* moderation system, which may be better able to harness community knowledge and cultural nuances while fostering a sense of inclusion and trust in the moderation process. Although our study did not observe a significant impact on the polarization of moderators or users, moderators on average viewed the system as just, legitimate and effective at reducing harmful content. Furthermore, there were no significant differences between user perceptions of the content they were shown from either system, indicating that implementing such a peer-based system has the benefit of providing users agency in platform governance without adversely impacting user experience.

RESEARCH HIGHLIGHTS

- The choice of moderation system had no significant impact on the ideological or affective polarization of users.
- There was a significant interaction between partisan affiliation and moderation condition impacting users' social polarization ($p = 0.023$), with both liberal and conservative users who viewed top-down moderated content becoming slightly more polarized, and users viewing jury-moderated content becoming less polarized (though the effect size was small; $\omega_p^2 = 0.18$).
- Both liberal and conservative moderators thought jury moderation was fair, valued jurors' individual voices, and achieved satisfactory outcomes. Conservative moderators viewed jury moderation as less legitimate exercise of a social media platform's power than liberals.
- Users who viewed jury-moderated content perceived it similarly to those who viewed top-down moderated content, indicating either would be acceptable.

Keywords: *social media; content moderation; user agency; juries; democracy; polarization.*

1 Introduction

Despite their success in connecting users across the globe, social media platforms have also played a role in amplifying and disseminating sensational and divisive content Silverman (2016), Wilson *et al.* (2020). Recent examples include the spread of conspiracy theories along partisan lines regarding the existence of widespread voter fraud and foreign interference in the 2016 and 2020 U.S. presidential elections Frankovic (2016), the competing #BlackLivesMatter and #AllLivesMatter social media campaigns and the associated protests and counter-protests Gallagher *et al.* (2018), and a partisan divide in compliance with public health measures to mitigate the effects of the COVID-19 pandemic Milosh *et al.* (2020). Such content can contribute to political polarization, in turn exacerbating the phenomenon of identity politics among the electorate, fraying social cohesion Schirch (2023), and harming democratic norms and institutions (see Tucker *et al.* (2018) for a synthesis of the relevant literature).

The scale and ease with which harmful content can spread poses challenges for several platforms. To curb the spread of such content and ensure they remain safe and enjoyable environments for their users, platforms engage in moderation, whereby they detect, review and respond to content that may violate community standards Gillespie (2018). A common theme of existing social media moderation structures on most mainstream platforms is that they are *autocratic*: users interact in ecosystems where the rules and their enforcement are chiefly the responsibilities of the platforms themselves despite the fact that they purport to be neutral hosts of content generated by users Gillespie (2018), with employees or contractors performing moderator duties. However, content may require context-specific knowledge and information about local sociocultural norms to be moderated effectively Jiang *et al.* (2021), Koebler & Cox (2018), and enforcement of moderation policies can be uneven: an independent civil rights audit Murphy & Cacace (2020) found Facebook sometimes fails to enforce its own community standards, and that harmful content could be left on the platform for too long, especially if it targeted members of minority communities. Even on platforms where most moderation is performed by select user volunteers, such as Reddit,

Received: September 22, 2023. **Revised:** April 12, 2024. **Accepted:** July 27, 2024

content users perceive as toxic is still prevalent Cook *et al.* (2021), and moderation decisions may still lack transparency Juneja *et al.* (2020). Moreover, lack of user agency and trust in the mechanisms of platform governance has been shown to negatively impact social cohesion, contributing to polarization Schirch (2023).

Given the shortfalls of existing governance structures on social media platforms, could other more democratic methods prove more effective? Much like the development of digital tools that support democratic participation in politics generally Nelimarkka (2019), Vlachokyriakos *et al.* (2014), the development of systems enabling user participation in the governance of online ecosystems can benefit from a multidisciplinary sociotechnical approach leveraging both knowledge of political theory as well as design and research methods from the field of human-computer interaction. A moderation system upholding democratic values and processes would enable effective enforcement of community guidelines while also ensuring transparency, fairness and accountability throughout the decision-making process De Gregorio (2020).

Building from existing frameworks of digital constitutionalism and jury decision-making, Fan & Zhang (2020) developed and evaluated a model for digital jury moderation, a promising alternative to typical top-down moderation which places the responsibility for social media content moderation in the hands of end users. Instead of moderators existing apart from day-to-day users, a digital "jury of one's peers" is selected among users to moderate content and reach a consensus on any consequences to employ. Digital juries are perceived as more transparent and procedurally just than existing practices, ensuring that members of online communities play an active role in moderating the content they interact with. The agency and opportunity for civic participation that digital juries and other community-driven approaches provide are valuable for maintaining a sense of collective responsibility for the growth and culture of digital social spaces Seering (2020). Therefore, if the moderation outcomes from their implementation are similar to or better than existing approaches, they may be worthwhile to adopt by virtue of the democratic values they espouse. While Fan and Zhang explored the design considerations for such a system and assessed how moderators viewed it, no existing work compares the attitudes of end users interacting with political social media communities moderated by an implemented digital jury versus traditional, top-down moderation.

This paper takes initial steps in examining the impact of implementing a digital jury moderation system on the political polarization of social media end users as well as user perceptions of such a system. Specifically, we address the following research questions:

- **RQ1:** To what extent does the design of a social media moderation system (digital jury system vs. standard moderation) impact the polarization of social media users?
- **RQ2:** How do users perceive a digital jury moderation system? How do moderators?
- **RQ3:** Are there areas for improvement in the structure of such a system and how it can be integrated into existing social media platforms?

Our aim was to gain a greater understanding of how the design of social media moderation systems can lead to different societal impacts, and the extent to which designs that support democratic processes lead to more positive outcomes than existing systems. To do so, we conducted a pre-post study comparing measures of liberal and conservative participants' polarization before and after two weeks of interaction with social media feeds containing political posts that had been moderated with either Reddit's existing top-down moderation by appointed moderators, or our implementation of a digital jury moderation system, where other participants acted as jurors. We find that while neither system had a significant impact on users' polarization, our jurors regardless of partisanship were satisfied with the jury's verdicts, and believed the system was fair. Additionally, users had similar views of the content they observed for both systems, indicating digital jury moderation could be a plausible, more democratic alternative to existing systems. Our study also opened several areas that need further attention for effective use of such systems.

2 Background and Related Work

In this section, we first define the types of polarization that have been identified in existing literature that we also investigate in our work. We then draw from prior work to illustrate how content on social media platforms can impact polarization and describe some steps platforms have taken to address this. Finally, we outline existing approaches to moderation that platforms have employed to regulate the content users encounter, as well as alternative moderation approaches that provide users with greater agency in platform governance decisions. It is currently not clear from existing literature to what degree any differences between moderation decisions made by appointed moderators and those made by users might impact polarization; our study explores this question. However, it is worth noting that in work that compared user perceptions of both types of systems, users tended to prefer systems that were more democratic. Systems whose affordances enable collaboration and more democratic engagement with governance can amplify social cohesion and mitigate polarization Schirch (2023).

2.1 Definitions of polarization

We distinguish between three different types of polarization that have been identified in prior work. *Ideological polarization* refers to the difference in ideological self-placement, e.g. on a liberal–conservative scale Iyengar & Hahn (2009). *Affective polarization* refers to how positively or negatively partisans feel about members of an in-party versus an out-party Suhay *et al.* (2018). *Social polarization* refers to how likely partisans are to socially self-segregate from members of an out-party Suhay *et al.* (2018). As these dimensions of polarization may be impacted differently depending on the content users interact with, we examine each of them in our study as well.

2.2 Political polarization and social media

Both major U.S. political parties have polarized steadily since the 1980s Bonica *et al.* (2013), and today the median Democrat and Republican are more ideologically divided than ever before Pew Research Center (2017). Partisanship has become closely linked to social identity, and perceived ideological differences have generated remarkable levels of hostility between members of opposing parties Wilson *et al.* (2020). Such identity politics can drive people to support their party's policy stances out of disdain for the opposition, rather than ideological agreement on issues. High polarization can worsen this phenomenon Druckman *et al.* (2013), Robison & Mullinix (2016), Tucker *et al.* (2018), Wilson *et al.* (2020).

Social media has the ability to amplify divisive political rhetoric, allowing political leaders to immediately disseminate information and opinions directly to their bases of support. Several studies have suggested that a tendency for social media users to self-segregate into echo chambers or filter bubbles Garimella (2018), where they are only exposed to information that reinforces their political views and are isolated from those with opposing views, is to blame for increasing polarization Barberá (2020), Spohr (2017). This leads to increased opportunities for enclave deliberation, where conversations only occur among like-minded people. While not inherently negative Conroy *et al.* (2012), members of homogeneous groups tend to adopt more extreme positions after discussions with their peers, either because the diversity of arguments is limited or because they are more likely to voice popular opinions in order to obtain the approval of as many other members as possible Barberá (2020). However, other studies have shown that cross-cutting interactions on social media are more frequent than commonly believed Barberá (2020), Barberá *et al.* (2015), Tucker *et al.* (2018), with exposure to ideologically diverse news and opinions more common online than from either in-person networks or traditional media consumption. One way to reconcile these seemingly contradictory observations is that social media platforms allow users to maintain contact with "weak ties": classmates, coworkers and other acquaintances. It is through weak ties that people are exposed to novel information Granovetter (1973), and their views are more likely to differ from one's own than those of close friends and family members.

Some work has assumed that such exposure to opposing views would decrease polarization Garimella (2018). Rather than decreasing polarization, however, cross-cutting interactions with political content on social media may exacerbate it instead. Bail *et al.* (2018) showed that exposure to messages with opposing political views can increase ideological polarization. This seems to indicate the presence of a backfire effect, whereby exposure to opposing views caused participants to double down on their existing views. Suhay *et al.* (2018) similarly found that participants who read news articles with negative comments that were critical of either party were more affectively and socially polarized than those who read negative nonpartisan comments, indicating that criticism of partisan identities, rather than opinions about specific issues, could be driving polarization on social media. Interestingly, this effect was stronger among Republicans than Democrats.

While not all content that is polarizing is inherently harmful or extreme (voicing differing perspectives about issues such as climate change, income inequality, and so on can lead to productive discussions and social change), content that is harmful, such as misinformation, hate speech and political propaganda, has been shown to contribute to polarization Schirch (2023). Such content can aggravate existing social divisions and negatively impact social cohesion by fragmenting public discourse on issues and undermining digital governance and norms González-Bailón & Lelkes (2023). The engagement-driven framework designed to maximize ad-based profits that underlies social media platforms amplifies content that is sensational and divisive, leaving users more polarized and vulnerable to political influence campaigns Schirch (2023). This polarization can limit the ability of societies to respond effectively to complex problems, lead to mistrust of social groups and public institutions, and reduce belief in civic engagement as an effective route to change Zuckerman (2021).

Platforms have several tools at their disposal to tackle these issues. Major platforms already employ automated moderation systems to remove some harmful content Gorwa *et al.* (2020), Koebler & Cox (2018), though their application may be opaque or uneven. For the content that remains, users have the ability to report content for a wide array of infractions for platform community standards Gillespie (2018). In the case where content is misinformative or untrustworthy, Facebook Meta (2021), X (formerly Twitter) Roth & Pickles (2020) and Instagram Instagram (2019) have employed fact-checking operations that apply labels warning users of the issue. Several researchers have either developed their own extensive credibility indicators Zhang (2018), or assessed the impact such labeling had on user perceptions. Fake news flags were found to have no influence on user judgments of truth, with users likely to believe news that aligned with their political opinions regardless of labeling Moravec *et al.* (2018). Users were also more likely to trust headlines they had seen before, regardless of whether they were flagged Pennycook *et al.* (2018), and attaching warnings to fake news articles also increased trust in articles without warnings Pennycook *et al.* (2020). Despite this, crowdsourced fact checking by lay users was shown to be strongly correlated with ratings from expert fact-checkers, with users tending to rate mainstream sources as more reliable than hyper-partisan or fake sources regardless of their political affiliation Pennycook & Rand (2019). This indicates that aggregate efforts by social media users are effective at assessing the trustworthiness of news sources, and could be used to inform content ranking algorithms to better prioritize trustworthy sources. Twitter, now called X, deployed the Community Notes (formerly Birdwatch) system Coleman (2021), which implements a community-based approach to labeling misinformative content, allowing users to provide context for tweets and reach a consensus as a community on which context is helpful. In a study, users largely found the user-added context notes helpful, and were less likely to agree with misinformative content that had these notes Coleman (2022). However, platforms may also change their policies and practices periodically (e.g., X's changes to its moderation staff and practices after its change in ownership Conger *et al.* (2022) and Reddit's changes to its API pricing Bell (2023)), so the effectiveness of Community Notes may evolve and change in the future.

Technologies and platform features that have implemented more democratic affordances for users fall under the umbrella of *peacetech*, which broadly aims to foster prosocial behavior, improve social cohesion and reduce polarization Schirch (2020 2023). Several tools and systems have been developed to facilitate civil and productive conversations between users of different groups, including eBay's online dispute resolution (ODR) system Rule & Schmidtz (2018), which allows buyers and sellers to resolve disputes using a web-based forum that provides scaffolding for fair discussions to ensure resolutions satisfactory to both parties; and Reddit's ChangeMyView subreddit Jhaver *et al.* (2017), a heavily moderated online community that incentivizes building bridges between different worldviews by gamifying conversations, where users whose views have been changed can reward those who made compelling arguments with a delta symbol (Δ). Other tools, such as Community Notes, are designed to leverage democratic participation to build consensus and trust. Another example is Polis Smith *et al.* (2020), a crowd-powered survey platform allowing users to host and take part in large-scale digital citizen assemblies, in which users can submit ideas related to a central topic and vote on the ideas of others, thereby mapping out the spectrum of perspectives of large groups on various issues and allowing for shared understanding. Although these systems offer insights into design strategies that are able to support social cohesion and mitigate the impact of online toxicity, few have reached the level of adoption necessary to combat widespread polarization. We chose to focus our attention on content moderation, as this is already a core feature present on all platforms influencing the content users interact with, while keeping in mind how incorporating more democratic affordances for users could have a positive impact.

2.3 Content moderation

According to Gillespie *et al.* (2020), *content moderation* refers to "the detection of, assessment of, and interventions taken on content or behavior deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labor and technologies required and the institutional mechanisms of adjudication, enforcement and appeal that support it". Social media platforms engage in moderation to safeguard their users and foster environments they will engage with, while navigating the legal and political dynamics of speech online. While there is little existing work examining the direct impact of moderation decisions on polarization Haimson *et al.* (2021), Shen & Rose (2019), we chose to explore the possibility of a relationship between them in our study because such moderation can influence the selection of content that users might see, which in turn has been shown to have such an effect (Section 2.2). Here we outline existing moderation approaches, their shortcomings and user perceptions and potential alternative approaches.

2.3.1 *Existing Approaches*

Broadly speaking, platforms employ two main moderation philosophies (sometimes in tandem): moderation can be centralized, whereby enforcement of the platform's content policies is managed by platform employees, teams of external contractors and the platforms themselves in the form of automated machine learning algorithms, (Facebook, X, YouTube); or decentralized, with moderation driven by platform users (Wikipedia, Reddit, Nextdoor communities). In the case of the former, while specific policies may differ between platforms, human moderators they contract or employ view posts that have been flagged by algorithms or users and decide whether they are permitted on the platform Seering (2020). Depending on the severity and frequency of offenses, moderator actions might vary from merely removing offending posts for first-time infractions, to temporarily suspending or permanently banning user accounts in the case of repeat severe offenders. Crucially, this decision-making process lacks civic participation from users.

In the case of the latter, users play a more active role in developing community-specific rules and making and enforcing moderation decisions. Wikipedia, for example employs a decentralized structure that emphasizes open deliberation for delegating tasks and resolving disputes, though it has been criticized as bureaucratic and confusing for newcomers Im *et al.* (2018). On Reddit, any user can create a community, or "subreddit," but only a subreddit's moderators may create and enforce community rules, rather than members of the subreddit as a whole. Instead, all users are able to influence the visibility of content by upvoting or downvoting posts or comments. Platforms employing such hierarchical community-based moderation approaches are superseded by site-wide community standards. Enforcement of moderation decisions may occur either before content is visible to users, or after it is already publicly available Veglis (2014).

A key difference between moderation on Reddit and other platforms is that moderators, while beholden to Reddit's rules, have a high degree of autonomy and are typically active members of the communities they moderate Gilbert (2020), and thus may be able to take advantage of sociocultural knowledge of their communities to provide nuanced perspectives on any issues that may arise Dosono & Semaan (2020). By enabling them to prioritize certain content and interact directly with group members, Reddit gives moderators the agency to shape collective action and protect marginalized communities, e.g., women, LGBTQIA individuals and members of ethnic groups such as Asian Americans and Pacific Islanders (AAPI) from brigading and erasure, as well as set positive examples by widening participation and modeling civil discourse Dosono & Semaan (2020).

While efforts by moderators can go a long way toward maintaining healthy and supportive digital spaces, the decisions they make can put them at odds with their communities as well. It is ultimately moderators, rather than community members, who wield control over which content and users to allow. It is possible for longstanding senior moderators to arbitrarily remove other moderators and ban users regardless of whether others agree with their decisions. Efforts by moderators to enforce rules and limit certain types of speech leads users unsupportive of these efforts to perceive moderation as censorship Gilbert (2020), Vaccaro *et al.* (2020). Indeed, platforms employing centralized and volunteer-driven moderation alike are incentivized to prioritize removing content that is deemed offensive, even if such removals stand at odds with notions of freedom of speech and the priorities of First Amendment legal doctrine Langvardt (2017). It is a matter of ongoing debate exactly where boundaries should be drawn, which moderation decisions are appropriate, and how these decisions should be conveyed to users Common (2020), Thach *et al.* (2022), Vaccaro *et al.* (2020).

Thus while moderators can remove polarizing content, the removal of such content can also be polarizing, and the legitimacy of moderation efforts ultimately depends on those impacted by these decisions Matias (2019). Moderators' political ideology also influences their perceptions of their own roles—liberals are more likely use metaphors evoking parental nurturing and fairness, while conservatives more likely to use metaphors evoking discipline and rules enforcement Seering *et al.* (2022). It remains an outstanding question whether the experience of Reddit moderators and context they are aware of can impact levels of polarization differently than the collective knowledge of community members, and what role political ideology may play in any differences.

Because of the large volumes of user-generated content posted to social media platforms (Facebook users, for instance, create billions of posts per day Koebler & Cox (2018)), they have increasingly turned to using automated algorithmic moderation solutions to effectively scale what would otherwise be an intractably large undertaking by human moderators. This algorithmic content moderation is used at scale by Facebook, YouTube, X (formerly called Twitter), and others to classify user-generated content through either pattern matching or prediction, employing perceptual hashing Niu & Jiao (2008), machine learning classification and other techniques to judge whether such content is appropriate or prohibited Gorwa *et al.* (2020). Such moderation techniques have proven effective at detecting spam, violence and nudity, but are less adept at detecting inappropriate use of copyrighted content, hate speech or other toxic speech, with AI tools lacking context awareness or knowledge of cultural nuances necessary for classifying such instances with high accuracy Gorwa *et al.* (2020), Koebler & Cox (2018). Sometimes these tools can be easily evaded by malicious actors, as Gerrard reports in a case study on the use of hashtag moderation on Instagram, Pinterest, and Tumblr to detect and remove content with pro-eating disorder hashtags Gerrard (2018), illustrating limits to language-based approaches. Furthermore, while some platforms release regular reports outlining the prevalence of different types of content violations Meta (2023), Reddit (2023), , these do not share transparent information about how moderation algorithms are trained and how their decisions are implemented. Thus, while useful for alleviating the workflows of human moderators, deploying these tools makes moderation decisions less transparent, obfuscates accountability, re-obscures the political nature of speech decisions by platforms Gorwa *et al.* (2020) and risks undermining free speech and equitable information access Oliva (2020).

Even though the moderation actions of platforms are largely successful at removing the most harmful content in a timely manner, the evenness of their application across different demographics and how users perceive them can vary. Users whose content is removed are often left wondering why, and a lack of transparency on the part of moderators and their decisions has left users distrustful of moderator decisions West (2018). On Reddit, this is especially true for users with new accounts, whose posts may be removed by Automoderator (a bot subreddit moderators can configure to automate various moderation tasks) for violating unlisted account age or thresholds for karma (points indicating a user's reputation, related to the number of upvotes their posts and comments have received) Squirrell (2019). 69% of analyzed content removals were not accompanied by any moderator feedback Juneja *et al.* (2020). A survey of Reddit users whose content was removed showed most disagreed with these removals, and several were confused and angry about these decisions Jhaver *et al.* (2019). Similarly, Haimson *et al.* (2021) assessed perspectives of Black, transgender and conservative social media users whose content was removed by moderators. Content from conservative users was generally removed for being offensive, containing misinformation or hate speech Jiang *et al.* (2019), Shen *et al.* (2019), while content from Black and transgender users was related to them expressing their marginalized identities, but being labeled as racism or "adult" content, respectively, even if no rules were violated. Inconsistent moderation can limit the ability of community members to understand the bounds of acceptable behavior, contributing to incivility Dosono & Semaan (2019).

2.3.2 *Alternative Approaches to Platform Moderation and Governance*

Researchers have developed several tools and platforms that give users more agency in governance decisions and facilitate discourse between users with different beliefs. These include PolicyKit Zhang *et al.* (2020), a framework for users to develop and enforce their own governance structures; Crossmod Chandrasekharan *et al.* (2019), a Reddit moderator tool that recommends actions based on sitewide norms as well as those of similar communities; and several others Cambre *et al.* (2017), Kulkarni *et al.* (2015), Matias & Mou (2018). This focus on the empowerment of users to shape the governance of their platforms stands in contrast to many existing approaches, where users are beholden to the decisions of platforms. Existing automated systems for performing or assisting with moderation may be effective in several instances, but they can exhibit biased decision making, overreliance on efficiency, and inconsistent enforcement of decisions Common (2020). While more user-centric approaches do not necessarily eliminate these issues, sharing decision-making power has the potential to increase transparency and accountability and incentivizes building up and maintaining online communities Seering (2020).

Several studies also developed tools for and tested the effectiveness of platform users engaging directly in moderation. Hettiachchi & Goncalves (2019) analyzed how 28 participants recruited from Amazon Mechanical Turk moderated political content from Twitter (now X), and found that Tweets that were labeled as inappropriate contained profanity, hate speech, grammatical errors or were off-topic, similar to removal reasons from Reddit's subreddit moderation Fiesler *et al.* (2018). Jahanbakhsh *et al.* (2022) developed Trustnet, a social media platform that allowed users to share their assessments of the accuracy of posts and other users, and filter content by accuracy. In a user study, they found that participants leveraged the features of the platform to assess the veracity of content posted by other users and themselves, and used this information in various ways to filter content appearing on their homepage feeds, supporting the value in placing more filtering control in users' hands.

Vashistha *et al.* (2015) created Sangeet Swara, a community-moderated voice forum for users in rural India with limited internet access. Users were highly engaged with the platform, but notably the content was uncontroversial, as the domain was entertainment rather than politics. Squadbox Mahar *et al.* (2018), by contrast, allowed users to moderate more controversial content via *friendsourced moderation*, where recipients of online harassment via email could organize a squad of friends to monitor their email inbox, allowing them to filter, reject, redirect, and organize email messages.

Fan & Zhang (2020) explored peer-based moderation in the context of a more traditional social media environment by assessing how users perceived "digital juries," a moderation system where platform users would be actively involved in making moderation decisions, and on which this work is predicated. These juries would place potentially rule-breaking content "on trial," and jurors would need to reach a consensus about any punitive actions to be taken. Online juries have been shown able to make consistent and repeatable moderation decisions Hu *et al.* (2021), and are seen by users as more legitimate than algorithmic moderation Pan *et al.* (2022). Overall, digital juries have the potential to overcome problems with existing moderation systems because they are viewed as democratically legitimate and directly empower user stakeholders in the governance of social media platforms. However, further research is necessary to evaluate whether implementing jury moderation can influence the relationship between platform users and the content they see, and to what degree this can ultimately impact their political polarization.

3 Digital Jury Moderation Study

The main goal of the study was to begin exploring the impact of implementing a digital jury moderation system on the polarization of social media end users relative to top-down moderation by appointed moderators. Furthermore, we also wanted to see whether the subjective perceptions of the content users saw would differ depending on which moderation system the content they viewed had passed through. To realize these goals, we employed a mixed-methods approach to our experiment design and analysis, leveraging quantitative measures of polarization and perceptions of the jury moderation process in order to allow us to make statistical conclusions, along with qualitative analysis of open-ended responses to gain a nuanced understanding of participant experiences moderating and interacting with content and uncover common themes, similar to approaches in prior work Fan & Zhang (2020), Haimson *et al.* (2021), Jhaver *et al.* (2019). Though the sample size and duration of our experiment was limited by operational constraints, we discuss these limitations as well as ways areas that might be improved in future work in Section 6.2.

We first recruited participants to act as jurors for our implementation of a digital jury, and additionally sought to replicate the analysis conducted by Fan and Zhang assessing how moderators viewed the democratic legitimacy of the system, and to gather any insights or considerations they noted from their moderation experience for how a digital jury might be deployed on a real social media platform. We then recruited a new set of participants to act as end users to engage with content moderated either via traditional moderation or by the digital juries. We explored **RQ1** by comparing responses to survey questions assessing polarization before and after the experiment. We analyzed responses to questions assessing how either jurors or users perceived their portions of the experiment and jurors' recommendations for improving and deploying the system to explore **RQ2** and **RQ3**, respectively.

3.1 Study design

We conducted a two-phase study. The first phase channeled social media content through two different moderation workflows and then propagated the results of each to a researcher-controlled platform. In the second phase, "end user" participants interacted with content from one of the two moderation approaches.

Thus, there were two study conditions: (i) a control condition, where users interact with posts from a community employing "status quo," top-down supervised moderation (simulated by Reddit moderation, where appointed moderators enforce moderation decisions), and (ii) a digital jury condition. The digital jury platform from Fan & Zhang (2020) serves as the basis for our digital jury. Fan and Zhang piloted multiple conditions, and feedback indicated an immersive jury, where jurors deliberate via online chat, can reduce juror disagreement; additionally, prior research indicates jurors report higher satisfaction when the final ruling arises from unanimous (vs. majority) agreement Fan & Zhang (2020), Nemeth (1977). Therefore, these are conditions we adopted for our study.

The goal was to simulate the effect of community moderation: how would moderators with similar interests as the users whose content they are moderating affect the attitudes of users subjected to such moderation? Thus, we recruited politically engaged users of social media on Reddit (Section 3.4), and placed them into three overarching groups: two user groups (one per experimental condition), and one jury group. Additionally, participants were grouped by political interests/affiliations, as users with different political beliefs tend to view different content on social media platforms Garimella (2018).

The moderated content was real content from political subreddits on Reddit (e.g., user submissions and their associated discussion threads), the majority of which were subjected to moderation by subreddit moderators. This selection was pruned by researchers to ensure that the content was both topical and not overly toxic (did not contain nudity, pornography, explicit graphic violence or content designed to incite violence; see Section 3.3).

The study design was reviewed and approved by our institution's institutional review board (IRB) under protocol IRB-20-0810. A diagram of the study is shown in Figure 1.

3.2 Measures and data analysis

3.2.1 *Measures*

The measures and survey instruments used in the experiment are summarized below. See Sections A and B of the Supplementary Material for further details.

(i) Political affiliation and social media engagement, for participant screening and demographics: Questions as in Supplementary Materials Section 2.2 of Bail *et al.* (2018), e.g., "What do you consider your political affiliation? (Republican / Democrat / Independent/ Libertarian / Other / Not Sure)"; "Do you visit a social media site at least three times a week in order to read messages/posts?"

(ii) Political interest, adapted from Keeter & Igielnik (2016), for participant screening: "How engaged do you consider yourself with US politics? (Very Disinterested / Disinterested / Neutral / Interested / Very Interested)"

(iii) Polarization, to answer **RQ1**: These are assessed before/after the experiment for both moderators and users via several Likert scale questions. Three types: ***ideological*** Bail *et al.* (2018), ***affective*** Suhay *et al.* (2018) and **social** Simas *et al.* (2020), Suhay *et al.* (2018). Using the method from Bail *et al.* (2018), we reverse-coded responses to questions that indicated favorability toward liberal positions/Democrats, then normalized and averaged responses to each group of questions. This yields scales ranging from −1 (most liberal) to +1 (most conservative) for each type of polarization. Our score for social polarization is the mean of three sub-scores measuring partisan **marriage preference** (i.e., the degree a relative marrying either a Republican or Democrat would be acceptable), a desire to self-segregate and maintain **social distance** from out-partisans, and the desire to live with other members of a **like-minded community** comprised of the partisan in-group.

(iv) Subjective impressions of participant experiences with either the digital jury moderation system or the content they viewed, obtained after the experiment via questionnaire, to answer **RQ2** and **RQ3**. Specifically,

  (a) Likert-scale questions assessing moderators' perceptions of the digital jury moderation process (e.g., "To what degree do you think that the moderation process was fair?") and how users perceived the content they saw (e.g., "On average, to what extent did you agree or disagree with the content you read during the course of the experiment?"), as well as open-ended responses from users, to answer **RQ2**.

  (b) An open-ended question for moderator participants "Are there ways you think the moderation system could be improved?" to answer **RQ3**.

(b) While participant responses to these questions were based on their experiences with our implementation of a digital jury, we discuss in Section 6.1 the degree to which themes that arise relate to the implementation and use of digital juries more broadly.

(v) The Secondary Traumatic Stress Scale for Social Media Users (STSS-SM) in Appendix C of Mancini Mancini (2019), to assess if participants experienced emotional distress during the course of the study.

We additionally collected jurors' voting data (their toxicity ratings for case components, which actions were taken, and whether voting was unanimous) from the moderation phase of the study, which informs **RQ2**.

3.2.2 *Data Analysis*

We collected data from two groups of participants in Phase I (liberal and conservative jurors), and four groups from Phase II (liberal and conservative users, who interacted with content from either top-down or digital jury moderation). We conducted exploratory statistical analyses of quantitative data collected from our pre- and post-surveys to measure differences in responses *within* groups (differences in measures of polarization before versus after the experiment for each) and *between* groups (changes in polarization, as well as differences in Likert scale responses from item iv.a above, between participant groups from each phase of the study). To measure within-group differences, we used Wilcoxon signed-rank tests to compare ordinal data (raw

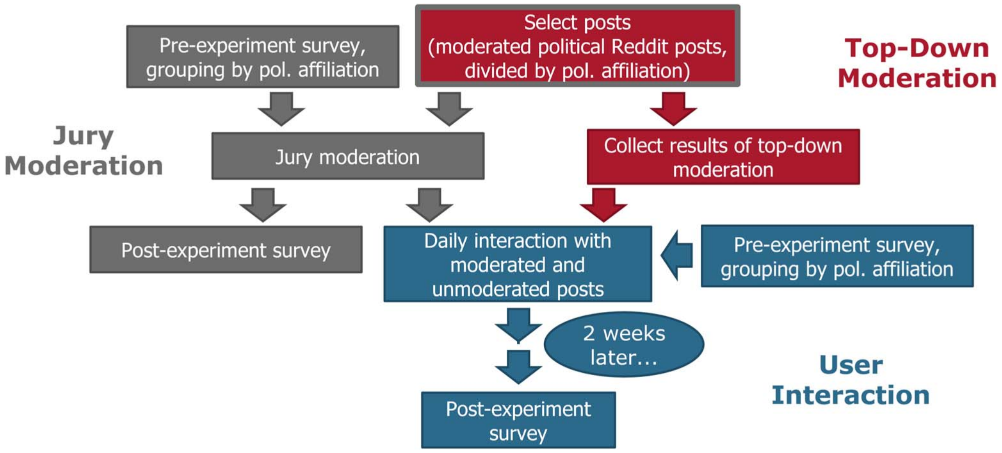

Pre-experiment survey, grouping by pol. affiliation
Select posts (moderated political Reddit posts, divided by pol. affiliation)
Top-Down Moderation
Jury Moderation
Jury moderation
Collect results of top-down moderation
Post-experiment survey
Daily interaction with moderated and unmoderated posts
Pre-experiment survey, grouping by pol. affiliation
2 weeks later…
User Interaction
Post-experiment survey

FIGURE 1. Flow diagram of each portion of the study. ***Phase I: Moderation***. First, political posts from Reddit were selected by researchers. Posts that had been moderated were sent to two different moderation systems. In jury moderation (gray, left), jurors were presented with a series of posts and the associated discussions, and deliberated to choose what, if anything, to moderate, and what the associated consequences would be—whether to ban the author, delete the post, alert authorities, etc. "Status quo," top-down moderation (red, top right) for these same posts is simulated by the researchers, using the results of the moderation that occurred on Reddit (i.e., whether posts were removed by moderators). ***Phase II: User interaction***. A separate group of participants taking the role of end users read the moderated posts for each of their respective conditions (blue, bottom right). This continued for two weeks with each day's content selection from both moderated content from Phase I as well as content that was unmoderated.

Likert scale responses) and paired *t*-tests to compare continuous data (our outcome measures for each type of polarization). To measure between-group differences from Phase I, we used Mann-Whitney *U* tests to compare ordinal data between liberal and conservative moderators and unpaired *t*-tests to compare continuous data; we did the same for data from Phase II when comparing liberal and conservative users within each experiment group. We additionally use multivariate ordinary least squares linear regression to assess the impact of experiment condition on scores the three types of polarization while controlling for demographic data. All regression models are provided in Section C of the Supplementary Material.

To analyze the qualitative data from our open-ended response questions, we employed thematic analysis Braun & Clarke (2006) to categorize participant responses. We first used inductive open coding to describe elements of participant responses, identifying similar relevant components, and then grouped these into themes (Sections 5.1.2 and 4.1.3) without aiming to develop axial and selective codes. Any concerns, questions, comments, or suggestions relating to the content that was moderated or seen, the design of the study websites, the moderation process and workflow or nature of online speech were considered thematically relevant.

3.3 Content selection

We selected our content from the PushShift archive of Reddit submissions and comments Baumgartner *et al.* (2020). Specifically, it stores posts in their initial state shortly after submission, allowing us to see any posts or comments that might have been removed as they were prior to removal.

We first curated a list of ideologically diverse political subreddits with at least 10,000 members (listed in Table 1) by examining existing Reddit submissions that listed contemporary political subreddits and their ideological leanings, verifying that their content was political in nature (either directly related to political figures and events, or related to a broader social issue such as abortion rights, the coronavirus pandemic, gun control, racism, police violence, economic and foreign policy, etc.), and additionally examined subreddits to which their users had cross-posted similar content. We then scraped their archives for posts created between December 28, 2020 and August 3, 2021 using PushShift's Python API, and then curated two sets of posts, one set each for liberal or conservative participants. Each set contained a total of 210 posts: 100 that had faced moderation by subreddit moderators (i.e., had a comment from a moderator describing the reason for moderation and met our inclusion criteria below), and 110 that did not. We included unmoderated posts so user participants would still have content to engage with in case the moderated content was removed. 60% of posts for each set were obtained from either liberal or conservative subreddits, and the other 40% were obtained from subreddits not from either affiliation (based roughly on Knobloch-Westerwick & Meng (2009), who found that approximately 40% of content users encountered on social media did not align with their views).

Content was selected in an iterative fashion. For all groups of posts (moderated or unmoderated for liberal or conservative participants), a random post from the archived set of posts from the relevant subreddits would be presented without replacement to a researcher for approval. Posts that had broken or missing links or images, contained violence or nudity, or had faced moderation (i.e., had a comment from a moderator indicating as such) for a reason unrelated to the nature of the post's content (was removed for formatting reasons or for being off-topic, and not for trolling, harassment or similar offenses) would be rejected. Additionally, posts that were accepted were rated for their perceived toxicity. Researchers were presented with the following instructions (similar to that used by Fan and Zhang):

> "Toxic content" can have many definitions, **including hateful, aggressive, or disrespectful comments** that may make it likely to **encourage violence, exacerbate derogatory views towards a group of people, or make a reader feel emotional or psychological harm**. Toxicity measures the degree that speech may have the potential to harm people, much like a toxic poisonous substance could cause harm. This may include slurs, epithets, profanity, insults, political dogwhistling (coded messages), or explicit/implicit threats. Some content may be seen as racist, sexist, homophobic, xenophobic, etc. Regardless of what you

TABLE 1. Subreddits used as sources of content for the study. Note that the current state of some subreddits may be different from when the study was conducted (e.g., r/The_Donald has since been banned, and r/worldpolitics is no longer actually related to world politics).

Liberal	**r/Liberal, r/progressive, r/SandersForPresident, r/HillaryForPresident, r/socialism, r/neoliberal, r/democrats, r/VoteBlue, r/BernieSanders, r/hillaryclinton, r/ChapoTrapHouse, r/esist, r/Political_Revolution, r/Biden2020, r/DemocraticSocialism, r/LibertarianLeft, r/obama, r/ElizabethWarren, r/Pete_Buttigieg, r/EnoughTrumpSpam, r/Impeach_Trump, r/Fuckthealtright, r/Anticonsumption**
Conservative	r/Conservative, r/The_Donald, r/TheNewRight, r/Republican, r/sjwhate, r/Anarcho_Capitalism, r/debatealtright, r/debatefascism, r/altright, r/new_right, r/MensRights, r/romney, r/progun, r/CollegeRepublicans, r/prolife, r/WatchRedditDie
Neutral/Other	r/politics, r/worldpolitics, r/Libertarian, r/LateStageCapitalism, r/ukpolitics, r/Enough_Sanders_Spam, r/GoldandBlack, r/PoliticalHumor, r/PoliticalDiscussion, r/eupolitics, r/uspolitics, r/geopolitics, r/COMPLETEANARCHY, r/conspiracy, r/collapse, r/AmericanPolitics, r/Anarchism, r/moderatepolitics, r/NeutralPolitics, r/GaryJohnson

> think should be done with the content, please use the following benchmark to select your **personal opinion** on how toxic the content is:
>
> - 0-2: **OK**, unlikely to cause harm
> - 3-7: **Borderline**, ambiguous or hard to say, with the potential to cause harm
> - 8-10: **Toxic**, likely to be perceived as aggressive, hateful, or with potential to cause harm.

Posts rated as 'Borderline' were considered moderation gray areas, and used for the moderation portion of the study; posts rated as 'OK' were accepted as-is; and posts rated as 'Toxic' (i.e., unambiguously harmful) were not shown to participants, to limit the potential for any mental distress from participation. We enforced a ratio of borderline to non-borderline posts, requiring 70 moderated and 35 unmoderated posts to be borderline (we assumed that posts that had faced moderation were more likely than posts that had not to fit into this category). This yielded a collection of 210 posts each for liberals and conservatives, half of which were borderline and shown to moderators.

Once an initial post selection was made, these posts were rated by at least one other researcher for their toxicity. The final toxicity score for a post was the maximum of all scores provided by researchers, and any post with a rating of 8 or higher (in the 'Toxic' category) was removed. This process was repeated for the remaining posts until an acceptable final set was obtained. Up to ten top-level comments from each post were also obtained and reviewed following the same procedure.

We reviewed a total of 817 posts, and kept a total of 417 for our final set (three posts from the neutral/other category occurred in the sets for both liberals and conservatives). Any usernames or identifiable text present in the selected content was altered or removed by researchers.

3.4 Recruitment

We recruited participants on Reddit using a method similar to Jhaver *et al.* (2019). Recruitment efforts began in August 2021 and were iterative, until March 2022. We chose Reddit as our recruitment platform because it allowed us to cheaply and easily tailor our recruitment efforts to active social media users across the political spectrum. (At the time the study was conducted, the Reddit and PushShift APIs were free to use and publicly available.) Targeting politically engaged social media users, we contacted a random subset of users who had commented on either the r/Liberal, r/Conservative or r/politics subreddits within the past month of each recruitment blast (determined using the PushShift API) with information about the study, as well as researcher and institutional review board contact information, and addressed any questions and concerns participants had about the study procedure, motivations, participant anonymity and data security practices. We kept the fact that moderators would be grouped with others of similar partisan ideology hidden so as not to potentially bias future moderation decisions.

We accepted participants who were ages 18 or older, from the USA or Canada, were active on social media for three or more days per week, and were engaged with US politics. We also included two questions ("What made you decide to participate in this study?" and "Imagine you are working on a group project, and one of your group members isn't doing an equal share of the work. How might you resolve such a situation?") in an effort to avoid recruiting any malicious actors who would not engage with the experiment; potential recruits who did not make a meaningful effort to answer these questions were rejected. Our screening questionnaire also included questions for our measure of ideological polarization, which assessed participants' policy positions on several issues and which we used to sort participants into either liberal or conservative groups. Participants who had an ideological score of zero were rejected, as this indicated no partisan leaning. The rationale for forming partisan juries mirrors that for using partisan content aligning with participants' views: since political communities on social media tend to be partisan, having partisan juries simulates selecting users to serve as jurors from these communities. While we were recruiting for the moderation phase of our study, we also asked participants to indicate their availability for time slots in the upcoming two weeks set aside for jury moderation. If participants were unavailable but still interested in participating, or they were ideologically aligned with a partisan affiliation that had already completed the moderation phase, they were instead assigned to the user cohort.

4 Phase I: Moderation

Once participants were screened and recruited, we obtained informed consent as well as responses to the pre-experiment survey (1, 2, and 3 from Section 3.2.1 above) before providing them access credentials to our study website, which hosted the moderation platform. We used participants' screening questionnaire responses to assign them to groups of 2–5 jurors of the same political affiliation and availability (Figure 2). (Only one group had two members, due to a last-minute scheduling conflict and inability to postpone the first session). Each group was assigned two dates and times to log in to the study website for synchronous deliberation sessions of one hour each. There were five liberal and five conservative groups, and each was responsible for moderating 11 cases on their first day, and

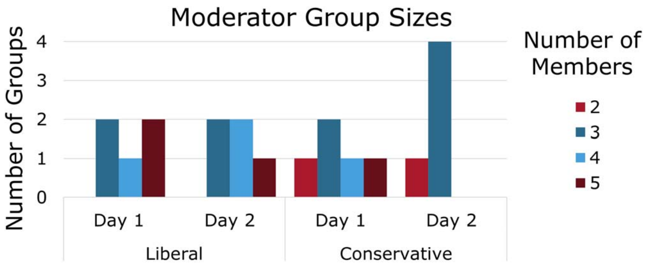


FIGURE 2. The makeup of each of the liberal and conservative juries for each day of the moderation phase of the study. The legend indicates the number of jurors in each group. Five liberal and conservative groups completed each day of the study, though some groups were smaller on the second day due to participant attrition.

10 on their second (for the requisite 105 cases per political affiliation that researchers had initially rated as having borderline toxicity). 19 liberals (median age, 25–34; 17 male, 1 female, 1 non-binary) and 12 conservatives (median age, 25 – 44; 10 male, 2 female) completed both days of the experiment (one liberal and three conservatives dropped out after the first, however, and one conservative had uncaught errors when taking the pre-experiment survey, leaving only demographic data available). Jury vote statistics presented in our results use votes from all jurors available; as those who withdrew from the study did not complete the post-experiment survey, all other results omit data from these participants.

At the scheduled start time, participants logged into the study website, and were shown a web page with onboarding instructions for their task (an explanation of the task workflow, what constitutes toxic content—analogous to the one researchers received in Section 3.3—as well as explanations for the possible punitive actions to take against content and users, along with a statement reminding jurors to keep their personal views in mind, and to aim to balance freedom of expression with maintaining a safe, comfortable browsing experience) and an instructional video for how to use the moderation platform. Deliberation was synchronous; once all jurors were in attendance, deliberation could begin. The site interface is shown in Figure 3. Each case consisted of one post and up to ten associated comments, collectively the "case components." An image of the post an any of its comments is shown as it would appear in our site for the user portion of the experiment. Any links or multimedia from case components could be clicked and viewed in the panel on the left. The grey bottom banner was open by default to show the voting interface, and could be toggled open and closed. Participants communicated via text-based chat with the other jurors, initiated by entering a user name in the chat box on the lower left.

For each case component all jurors needed to provide three assessments (slightly modified from Fan and Zhang):

- A toxicity score of content, defined as likelihood to cause harm (0-2 OK, 3-7 Borderline, 8-10 Toxic).
- Punishment for the content, if any (`unlist` from users' feeds, `delete` from the site, `report` to authorities).
- Punishment for the user, if any (`warn`, `ban` for 1 week, and `permanently ban`).

The dropdown on the bottom right allowed users to select components for the case. Jurors rated the toxicity of each post using the slider below the dropdown menu, and could select punitive actions to take for posts rated 3 or greater. All ratings and actions selected per component were saved as soon as jurors made them, and could lock in their vote once they were finished.

To constrain the time required for each day of the experiment and facilitate scheduling each moderation session, we enforced a six-minute time limit per case during which jurors could deliberate and vote on each component. (We felt this was an appropriate compromise between the unlimited duration of Fan and Zhang's study and the small amount of time spent per case in real moderation, while still giving jurors time to deliberate and reach consensus.) Jurors were advised that all of their votes and actions for each component needed to be unanimous, or the jury would be "hung" and the case tabled for later. A unanimous component is one where all jurors rated toxicity in the same bin ('OK,' 'Borderline' or 'Toxic') and selected the same actions. Any case not rated by jurors defaulted to a toxicity score of zero and no actions taken. If all jurors locked in their votes before the timer elapsed, jurors could immediately proceed to the next case; otherwise, all juror's votes from when the timer elapsed would be locked in. Once all votes were submitted, jurors could see the votes and actions selected by all members in the panel on the right.

Once the jurors had completed both days of moderation, they completed a post-experiment survey (items 3 and 4 from Measures above) and were debriefed. Jurors were provided $7.50 gift card credit for each day of participation. At the conclusion of the experiment, out of an abundance of caution to ensure that none of the content that participants saw had been harmful, participants also completed the Secondary Traumatic Stress Scale for Social Media Mancini (2019). If any participants scored higher than 27 (moderate secondary traumatic stress), we followed up and provided information about mental health resources.[1] [2] [3] Five juror participants were contacted.

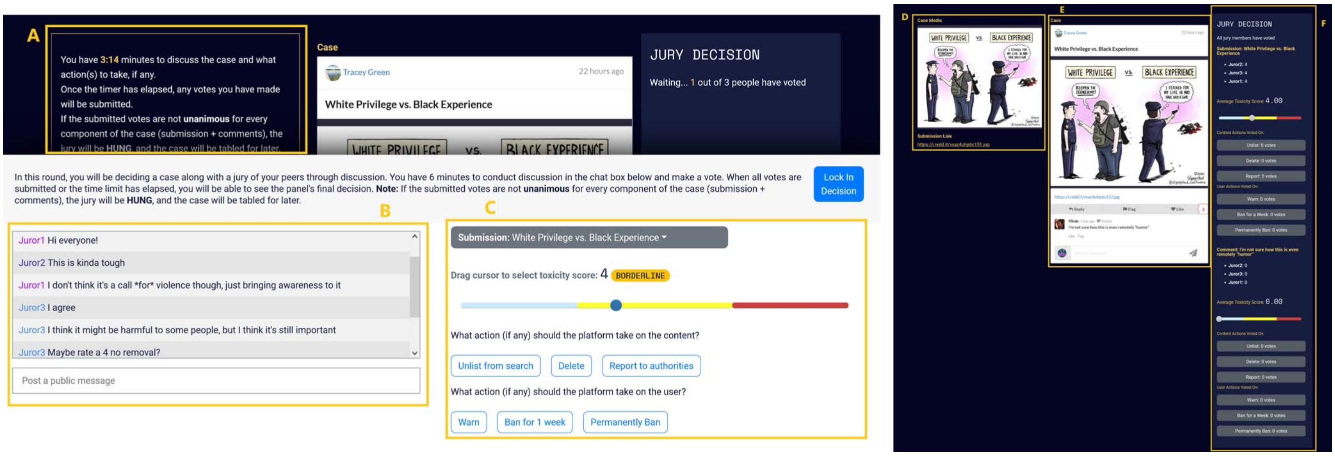
4.1 Phase I results

4.1.1 *RQ1: Changes in Moderator Polarization*

We assessed polarization using the measures described in Measures above. We used the raw responses to the Likert scale questions from our pre- and post-surveys to create continuous outcome measures for ideological, affective and social polarization. Our ideological, affective, and social scores (the mean of the marriage preference, social distance and like-minded community preference components), as well as the social score's components, range from ±1, with −1 the most aligned with or favorable to liberals and +1 the most aligned with or favorable to conservatives. As both our moderator and user participants were separated according to their political leanings, the quantitative results from both phases of our study are separated for liberals and conservatives as well. Results for statistical analyses are outlined below; our full regression models can be found in Section C of the Supplementary Material.

Our main goal was to investigate the polarization of end users, but to explore any differences that might exist between jurors, we analyzed polarization among our moderator participants as well. Paired *t*-tests were conducted to compare the continuous outcome measures before and after the experiment. We observed no significant changes in ideological, affective or combined social scores for either liberals or conservatives, though the social distance component of the combined social score had significantly decreased for conservatives (pre-experiment $M = 0.068$, $SD = 0.162$; post-experiment $M = -0.021$, $SD = 0.167$; $t(10) = 2.887$,

[1] https://www.crisistextline.org/
[2] https://www.nami.org/help
[3] https://www.iasp.info/resources/Crisis_Centres/



You have 3:14 minutes to discuss the case and what action(s) to take, if any.
Once the timer has elapsed, any votes you have made will be submitted.
If the submitted votes are not **unanimous** for every component of the case (submission + comments), the jury will be **HUNG**, and the case will be tabled for later.

JURY DECISION

Waiting... 1 out of 3 people have voted

In this round, you will be deciding a case along with a jury of your peers through discussion. You have 6 minutes to conduct discussion in the chat box below and make a vote. When all votes are submitted or the time limit has elapsed, you will be able to see the panel's final decision. **Note:** If the submitted votes are not **unanimous** for every component of the case (submission + comments), the jury will be **HUNG**, and the case will be tabled for later.

Juror1 Hi everyone!
Juror2 This is kinda tough
Juror1 I don't think it's a call *for* violence though, just bringing awareness to it
Juror3 I agree
Juror3 I think it might be harmful to some people, but I think it's still important
Juror3 Maybe rate a 4 no removal?

Post a public message

Submission: White Privilege vs. Black Experience

Drag cursor to select toxicity score: 4 BORDERLINE

What action (if any) should the platform take on the content?

What action (if any) should the platform take on the user?

Submission Link
https://i.redd.it/vxaz4uhp6c151.jpg

https://i.redd.it/vxaz4uhp6c151.jpg

Hiran
I'm not sure how this is even remotely "humor"

JURY DECISION

All jury members have voted

Submission: White Privilege vs. Black Experience
- Juror2: 4
- Juror3: 4
- Juror1: 4

Average Toxicity Score: 4.00

Content Actions Voted On:
Unlist: 0 votes
Delete: 0 votes
Report: 0 votes
User Actions Voted On:
Warn: 0 votes
Ban for a Week: 0 votes
Permanently Ban: 0 votes

Comment: I'm not sure how this is even remotely "humor"
- Juror2: 0
- Juror3: 0
- Juror1: 0

Average Toxicity Score: 0.00

Content Actions Voted On:
Unlist: 0 votes
Delete: 0 votes
Report: 0 votes
User Actions Voted On:
Warn: 0 votes
Ban for a Week: 0 votes
Permanently Ban: 0 votes

FIGURE 3. The Phase I moderation interface during deliberation (left) and once voting is complete (right). Box A: Timer and instructions are visible. Box B: Chat interface for jury deliberation. Box C: Voting interface to select toxicity score and any punitive actions for the component chosen in the dropdown. Box D: Any case links or media. Box E: Post and comments as they would appear to end users. Box F: Juror votes for toxicity score and any actions to take for each component.

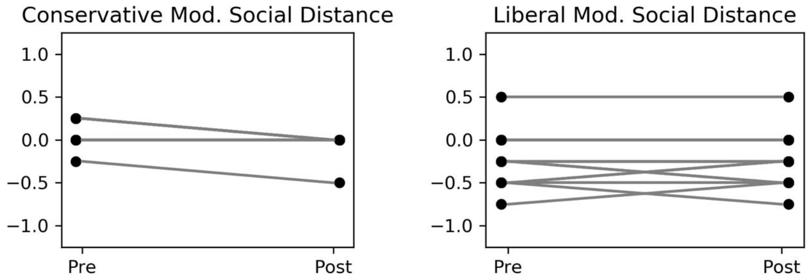


FIGURE 4. (Left) Slopegraph showing a significant difference in social distance for conservative moderators ($N = 12$) before vs. after the experiment ($p < 0.05$). (Right) Slopegraph showing social distance for liberal moderators ($N = 19$) before vs. after the experiment, with no statistically significant difference. In both figures dots are data from single participants, and the lines show how the scores for each participant changed after the experiment.

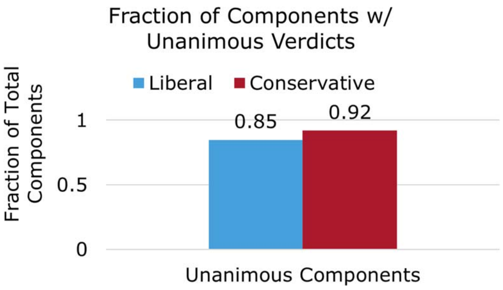


FIGURE 5. Fraction of case components with unanimous verdicts for liberal and conservative moderators.

TABLE 2. Fraction of case components with unanimous verdicts for liberal and conservative moderators, divided by toxicity category (Top) and action type (Bottom).

	Toxicity Category		
	OK (0 – 2.5)	Borderline (3 – 7)	Toxic (7.5 – 10)
Liberal	0.918	0.065	0.016
Conservative	0.995	0.005	0
	Actions		
	Content	User	Any
Liberal	0.033	0.033	0.039
Conservative	0.005	0.003	0.005

$p = 0.016$, 95% CI[0.026, 0.201]), but not for liberals (Figure 4). This is reflected in a significant difference between the change in social distance of liberal ($M = 0$, $SD = 0.144$) and conservative ($M = -0.114$, $SD = 0.131$) moderators, assessed via an unpaired t-test ($t(28) = -2.149$, $p = 0.040$, 95% CI[0.005, 0.222]).

4.1.2 *RQ2: How Moderators Used and Viewed Jury Moderation*

In addition to measuring its impact on polarization, we were interested in understanding how moderators used and perceived our implementation of a digital jury moderation system.

We analyzed the vote data from all of the juries to see if there were any differences between liberal and conservative juries, and whether these decisions were any different from those made by Reddit moderation, analysis dimensions unavailable in Fan and Zhang's study. We found that jurors of both affiliations consistently achieved unanimous verdicts, with liberals slightly less likely to do so than conservatives (Figure 5). This was largely attributable to the fact the vast majority of case components (over 90% for both cohorts) were marked as 'OK' for toxicity, with scores between 0 and 2.5. Liberals were slightly more likely than conservatives to rate posts as either 'Borderline' or 'Toxic,' and were six times more likely to take punitive actions against content or users than conservatives (Table 2). However, in absolute terms, the downstream impact on content was minimal: the actions liberal jurors chose resulted in the removal of 6 posts, and the actions conservative jurors chose resulted in the removal of only 2, compared to Reddit moderators' removal of 76 and 72 posts from the same respective collections . Thus our participants were

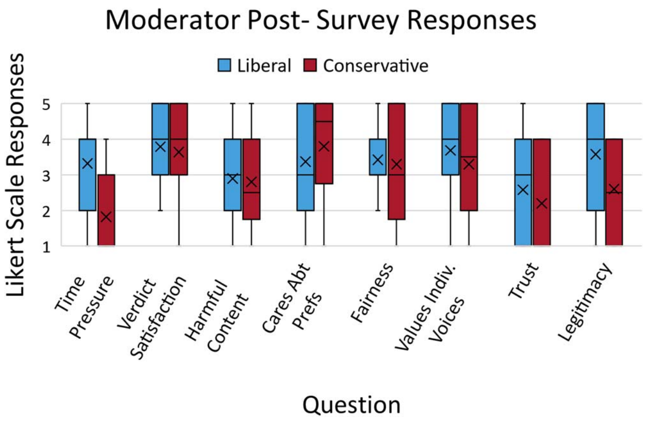


FIGURE 6. Moderator post-survey responses assessing the legitimacy of the moderation process, as well as satisfaction in verdict outcomes and subjective time pressure.

much less likely to remove content compared to Reddit moderators overall.

We also replicated part of Fan and Zhang's study to analyze the perceived democratic legitimacy of digital jury moderation. As in their work, we surveyed jurors on six criteria assessing their perception of the moderation process's legitimacy (legitimate exercise of power, trust, equal valuing of individual voices, fairness, care of personal preferences and efficacy in moderating content), as well as satisfaction in verdict outcomes and subjective sense of time pressure (Figure 6). We largely corroborate their findings, and additionally are able to separate results based on participants' political affiliations. Jurors of both affiliations believed the moderation process was fair, valued jurors' individual voices and preferences and achieved satisfactory outcomes. Measures of trust in the platform and platform efficacy in removing harmful content was lower for both affiliations than was observed in Fan and Zhang's study (five jurors noted there was, in their view, little objectionable content that was encountered). Mann-Whitney U tests showed liberals experienced significantly more time pressure than conservatives (liberal median = Somewhat (4), conservative median = Not at all (1), $U = 36.5$, $p = 0.003$, 95% CI$[-3.0, -1.0]$), which may be related to their increased likelihood to rate posts as having a higher toxicity and take punitive action, which requires more engaged deliberation to achieve unanimity (rather than leaving components at a score of zero). We also observed that conservative jurors were much less likely to view the moderation process as a legitimate exercise of a social media platform's power, though this difference did not reach statistical significance (liberal median = Somewhat (4), conservative median = A little (2), $U = 56.5$, $p = 0.075$, 95% CI$[-2.0, 0.0]$).

4.1.3 *RQ3: Drawbacks and Improvements*

Moderators provided the most insight regarding the shortcomings of our implementation, as well as suggestions for how it could be improved. We categorized participant responses and outline some common trends here.

Desire for clearer rules. Of the 30 moderators who completed the post-experiment survey, eight expressed that clearer rules would have been useful when making moderation decisions: "*Moderation is a very difficult situation to get correct. Your best bet is to set out clear and concise rules and get thick-skinned people to enforce them without bias.*"

Concerns about moderator bias. Six participants expressed concerns over how moderator bias could impact verdicts: "*The problem is that, more typically than not, you have an arbitrary number of moderators who all think exactly the same.... That is when speech can be impeded because it does not align within the moderator groups' political spectrum.*" Ensuring that there was "*more input from a variety of mods with different perspectives*" and that "*...the moderators are not politically motivated to censor social media*" was seen by some participants as key to ensuring an effective and fair moderation experience. There is merit to these comments as our juries were ideologically homogeneous to better reflect the nature of partisan communities online, but this can be contrasted with jury trials in the United States.

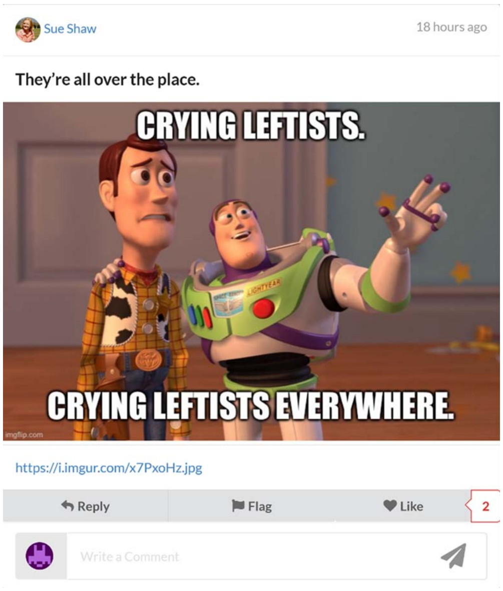
Time and social pressures. Eight remarked there was time pressure and expressed a desire for more time to deliberate, stating they would have liked "*less pressure to come to an agreement within a specific time limit*" and "*longer time to research/fact check.*" Two also commented on the presence of social pressure to change their vote when deliberating with other jurors: "*Once I voted that one meme was not toxic, I felt pressured to say zero again... Even though I read the instructions, I felt conflicted... because their standard was whether words on a screen could cause bodily harm.*" One user liked the unanimity requirement ("*Probably the biggest advantage of this system is that it helps reduce upward creep of scores. When everyone has to be unanimous, it brings down the scores that are just slightly higher than average.*") though others did not like that jurors in the minority could hold deliberation hostage ("*I can respectfully say that I think something is toxic but not everyone agrees which is why there were a lot of hung juries. People weren't willing to negotiate/budge on their rating.*")

Cases were easy. Despite the prevalence of time pressure, five jurors also commented on the relative ease of the moderation cases they encountered: "*I felt that you could have included edgier, more problematic content. The things I see on the internet are overtly racist, sexist, homophobic, etc., and I feel like the collection of posts we were given were mostly morally ambiguous.*"

Possible improvements. Several participants had suggestions for novel improvements to the system. Four expressed interest in the ability to see how other jurors had voted before submitting their votes: "*Being able to see each person's votes/selections before locking in, rather than only being able to chat about them before lock in, would make it easier and faster to compare approaches.*" One suggested adding an indicator of the jury's status during deliberation: "*If the UI gave an active meter of the decision as it's being made, it would help inform the discussion. Often we didn't know it was going to be hung until after it was too late to discuss.*" Others suggested changes to the voting system ("*[It] could be interesting to implement ranked choice voting for content actions*"; "*Discussion might be reserved for content that has already been elevated from a larger pool of asynchronously voting moderators who can make quick/gut decisions. Averaging many opinions together can do a good job where a few will be stochastic and noisy.*"), mechanisms for resolving disagreements ("*A mechanism for resolving disagreements would be helpful, instead of just tabling them.*"), or using different jury sizes ("*Larger juries would help.*")

4.1.4 *Summary*

We explored the polarization of moderator participants and found that conservative moderators became less socially polarized than liberal moderators. We found that liberal and conservative jurors moderated content differently, with liberals tending to remove more content and reaching slightly fewer unanimous decisions than conservatives while experiencing significantly more time pressure. Conservative moderators also perceived the moderation system as less legitimate than liberals, though both groups were satisfied with the juries' decisions and believed it valued their



Sue Shaw

18 hours ago

They're all over the place.

https://i.imgur.com/x7PxoHz.jpg

FIGURE 7. Sample post as it appeared on the simulated social media platform used for the user interaction portion of the study.

individual voices. While moderation was successful overall and participants enjoyed the experience of deliberating with other jurors, the broad interpretability of our moderation rules, moderator bias, and time and social pressures were all common concerns. These are issues a real-world implementation of a digital jury, and peer-based moderation systems more broadly, would need to address in order to be seen as just and trustworthy while also being scalable and effective.

5 Phase II: User Interaction

Once Phase I was complete, user groups began Phase II in January 2022. First, informed consent and pre-experiment survey responses were obtained. Users were divided into either liberal or conservative group based on their screening questionnaire responses, and were further divided into a moderation strategy condition (jury or top-down) (2 partisan affiliations × 2 experiment conditions = 4 groups). A total of 24 participants completed this phase—7 liberal users in the top-down condition (median age, 25–34; 6 male, 1 female), 10 liberal users in the jury condition (median age, 25–34; 9 male, 1 other), 4 conservative users in the top-down condition (median age 25–34; 2 male, 2 female), and 3 conservative users in the jury condition (median age 35–44; 2 male, 1 female).

The pools of potential content shown to either liberal or conservative participants were the same for each experiment condition; these pools were then reduced depending on which, if any, posts and comments either moderation system had removed. In the *top-down condition*, participants were shown unmoderated posts, as well as the results of Reddit's moderation of the moderated posts for their respective political affiliations. In the *jury moderation condition*, participants also saw the unmoderated posts, in addition to the results from the Phase I jury moderation. Any case components that jurors unanimously chose to unlist, remove or report to authorities were not shown to users, and any banned authors would be banned for the duration the jury had chosen if their content appeared elsewhere in the post selection (real author usernames were not shown to users). Cases that had been "tabled for later" were accepted as-is for the user group.

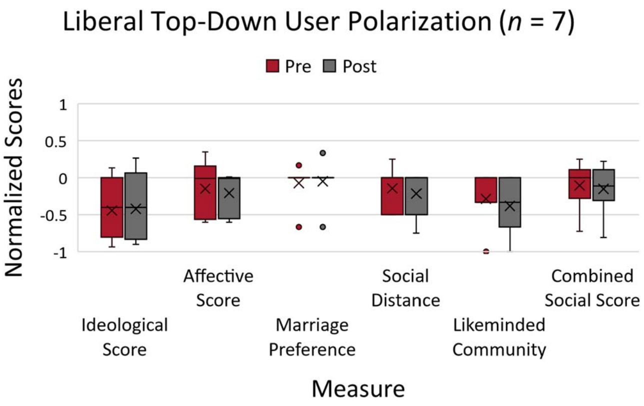


FIGURE 8. Continuous measures of polarization for liberal moderators from the top-down condition. The like-minded community component has been negated to facilitate comparison with the other data (more negative indicates more favorable toward liberals). A paired *t*-test indicated a significant decrease in the combined score for social polarization.

To host the content, we built a platform called Agora (Figure 7), on top of the Truman platform DiFranzo *et al.* (2018). Our implementation was simplified; the ability of participants to make posts of their own had been removed, as there were no scripted interactions with actors.

Phase II took place asynchronously over fourteen days for each participant. We evenly divided the post collections for each group into 14 sets (of 6–14 posts, depending on the group), to be shown on each day. From the moment they logged in, users had 24 hours to view the day's content before the content cycled to the next day. At this point, participants would be sent a link to a daily engagement question, asking about one of the posts from the prior day. Participants were paid $13 in gift card credit for the first day of the experiment, plus $0.50 per day they answered the daily engagement question (a maximum of $20). After each week, participants completed the STSS-SM, and were sent information about mental health resources if they scored 27 or above (moderate secondary traumatic stress; four user participants were contacted). Once the two-week period elapsed, participants completed the post-experiment survey.

5.1 Phase II results

5.1.1 *RQ1: Changes in User Polarization*

We assessed polarization for users using the same procedure as for moderators, described in Section 4.1.1. Paired *t*-tests comparing the continuous outcome measures before and after the experiment for all four groups (liberal top-down, liberal jury, conservative top-down and conservative jury users) yielded a significant decrease in the combined social score for liberal users in the top-down condition (pre-experiment $M = -0.103$, $SD = 0.318$; post-experiment $M = -0.151$, $SD = 0.338$; $t(6) = 2.828$, $p = 0.030$, 95% CI$[0.006, 0.089]$) (Figure 8).

Wilcoxon signed-rank tests to compare the raw Likert survey responses from before and after the experiment for all four groups yielded no significant differences.

We conducted a Type II two-way ANOVA to determine the effects of partisan affiliation and experiment condition on the changes (post−pre differences) observed for each of the continuous outcome variables. No significant main effects or interactions were observed for the ideological score or affective score, nor for the social distance or like-minded community components of the combined social score. However, there was a significant

TABLE 3. Changes in combined social score after the experiment for end user participants from each partisan affiliation and experiment condition. A two-way ANOVA indicated a significant interaction between affiliation and condition ($F(1, 20) = 6.112$, $p = 0.023$, $\omega_p^2 = 0.18$).

Affiliation	Condition	Mean	SD
Liberal	Top-down	-0.048	0.045
Conservative	Top-down	0.069	0.083
Liberal	Jury	0.003	0.103
Conservative	Jury	-0.167	0.313

interaction between partisan affiliation and experiment condition for the combined social score ($F(1, 20) = 6.112$, $p = 0.023$, $\omega_p^2 = 0.18$). Table 3 shows the mean score changes for each group of participants. Users who interacted with content from the top-down condition became more polarized, while users who interacted with content from the jury condition became less polarized, on average. Unpaired t-tests comparing changes in continuous outcome measures between experiment conditions for either liberals or conservatives yielded no significant differences.

To assess the impact of experiment condition on our scores for ideological, affective, and social polarization while controlling for demographic data, we performed multivariate ordinary least squares regression. This yielded three linear models (ideological, affective, and combined social polarization scores) for predicting each post-experiment score, with the experiment group, pre-treatment score, and seven other covariates as predictors. For both the ideological and affective scores, the pre-experiment scores were the only significant predictors of the post-experiment scores ($p = 0.003$ and $p = 0.004$, respectively). There were no significant interactions between experiment group and the post-experiment polarization scores, although the interaction between experiment condition and the combined social score approached significance ($p = 0.064$).

5.1.2 *RQ2: User Perceptions*

We next examined how the choice of moderation system impacted the experiences of end users. Once Phase II concluded, we surveyed participants about their overall impressions of how toxic the content they saw was, as well as to what extent they agreed with the content they saw to determine if there were significant differences in the perceptions of content between the two moderation systems (Figure 9). Given that the content in jury moderation was moderated by an aggregate of users' peers, rather than individual moderators, it is possible that moderation decisions would favor content that the majority of users agreed with. None of the responses indicated that such an effect existed, or that users were aware of the moderation system they were assigned. Liberal users in the top-down condition were slightly more likely to rate the content they saw as borderline (mean toxicity rating 4.14), compared to liberal users in the jury condition (2.8), conservative users in the top-down condition (1.5) or conservative users in the jury condition (2.33). Overall, users who viewed content from the jury moderation condition rated content more closely than those who viewed content from the top-down condition. Users of all groups appeared to have similar agreement with the content they saw, neither agreeing nor disagreeing. Ordinal logistic regression to determine if there were any differences mediated by partisan affiliation, experiment condition, or the interaction between the two indicated that there were no statistically significant differences between these responses for all groups. Thus participants viewed the content from both moderation systems similarly.

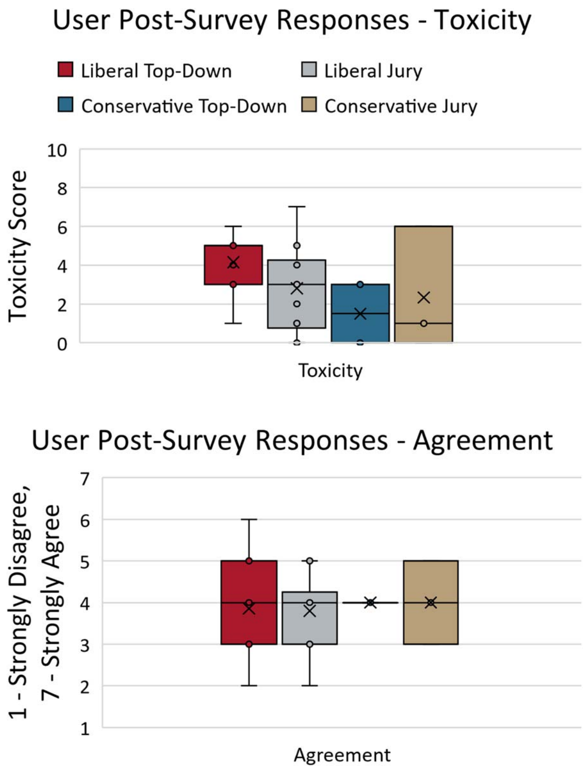


FIGURE 9. (Top) User post-survey responses assessing how toxic they believed the content they saw was. (Bottom) User post-survey responses assessing the extent users agreed with the content they saw.

Users disliked content. Two users viewed the content unfavorably. According to one user: "*Many of the posts admittedly were what I could only describe as boring or at least uninteresting posts seemingly made by people with little grasp of what they were endeavoring to pontificate upon.*" The other user stated that "*the majority of the comments, and some of the articles, simply seemed naive and ignorant.*"

Content was outdated. Two other responses expressed disappointment that the content was outdated: "*The experiment didn't evoke much emotion given that most of the articles were based on news stories from a year ago I was already aware of and had processed.*"

Lack of engagement. Two users were disappointed that they did not engage more with the content ("*I thought I would engage more.*")

Users liked content There were also two users who appreciated the content: "*I enjoyed reading the daily posts because they were not overly biased.*"

Content was unrealistic One user believed that the content was fake: "*It also felt very surreal and made it tough to genuinely take seriously because I saw the same "people" who had used their full name and a headshot as their profile picture. This made it feel like I was just reading fake/curated takes intended to try and evoke an emotional response more than a genuine take.*"

5.1.3 *Summary*

When we examined the polarization of user participants, our main focus, we did not observe a clear pattern indicating significant changes in ideological, affective, or social polarization mediated by experiment condition or partisan affiliation (although significant changes in some components of ideological and social polarization show users who viewed jury-moderated content became slightly less polarized on average). We did observe a significant interaction between partisan affiliation and experiment condition for the combined social score for social polarization, whereby users who viewed top-down moderated content became slightly more polarized, and users who viewed jury-moderated content became slightly less polarized; however, this difference was not significant in a pairwise comparison within each partisan affiliation. All groups of users had similar views of the content they encountered regardless of partisan affiliation or experiment condition, indicating that there were no observable differences between the moderation systems that users were aware of.

6 Discussion

Our use of digital jury moderation did not appear to have a significant impact on the political polarization of moderators or users. However, there were some effects on measures related to social polarization—social distance for conservative moderators decreased after the experiment, and the combined social score for liberal users in the top-down condition also decreased—that indicate partly reduced polarization for conservative moderators, but *increased* polarization for liberal users who were shown content moderated by Reddit instead of digital jury moderation. A significant interaction between experiment group and partisan affiliation for users for changes in the combined social score may relate to this as well, though pairwise comparisons between experiment conditions within the same affiliation did not yield meaningful differences.

Nonetheless, taken together these observations also support the trend that both liberals and conservative users who interacted with jury-moderated content became slightly less polarized, and liberal users who interacted with top-down moderated content became slightly more polarized (changes for conservative users in the top-down condition were mixed), with changes in social polarization the most pronounced. This was somewhat surprising because participants in the jury moderation condition viewed more content overall than those in the top-down condition, since jurors from both affiliations chose to remove less content than Reddit moderators; prior work has shown that exposure to toxic partisan content can increase social polarization Simas *et al.* (2020), Suhay *et al.* (2018). However, in our experiment, users in the jury moderation condition tended to rate the toxicity of the posts they saw lower in their post-survey responses, with their median rating lower than users in the top-down condition of the same partisan affiliation (though these differences were not significant). We computed Spearman's correlation to investigate this further and found a weak positive correlation between the absolute value of users' combined social scores (assessing whether they were more or less polarized, regardless of affiliation) after the experiment with how toxic they perceived the posts they saw ($r(22) = 0.26$, $p = 0.11$; H_1 = positive correlation). One conjecture as to the cause of this phenomenon is a salience bias in perceived toxicity; since users in the jury moderation condition saw more posts than those in the top down group, content with relatively higher toxicity could be more normalized and thus actually perceived as *less* toxic Beres *et al.* (2021). However, as the small sample sizes for our user groups limit the power of any of these claims, this should be explored further in future work (Section 6.2).

Despite the study's limitations, we note that digital jury moderation did not *increase* the polarization of users, even if it largely did not reduce it. The fact there were no significant differences in user perceptions of content when interacting with either system lends credence to the idea that using a digital jury system would be acceptable. Furthermore, many of our moderator participants who used the system enjoyed the experience of deliberating with other participants, were largely satisfied with the outcomes of their juries, and generally considered the system fair and procedurally just, corroborating Fan and Zhang's findings. Implementing such a system could therefore empower platform users to moderate their own communities and mitigate conflicts between the necessity of moderation and the profit motives of platforms Aswad (2018).

Differences in how liberals and conservative moderators used and perceived the system were also notable. Conservatives tended to have a narrower view of what qualified as toxic, valued free speech, and were wary of the power moderators could hold to censor dissenting voices. This is consistent with work showing that Republicans are much more likely than Democrats to disapprove of social media companies labeling posts from elected officials or ordinary users as inaccurate or misleading Vogel *et al.* (2020). They were therefore much less likely to remove content or take actions against it, so it is plausible that they felt less time pressure than their liberal counterparts because several groups had decided at the outset that content was not toxic and no actions would be taken. It was also interesting to note that liberals viewed the legitimacy of the system higher than conservatives, who in general were leery of moderators making decisions that aligned with existing biases. This reflects a partisan divide over perceived bias in the moderation of commercial social media platforms and attitudes toward censorship and the regulation of speech noted in existing literature, with conservatives much more likely than liberals to believe that platforms are more likely to censor conservative viewpoints and make moderation decisions that favor preserving liberal political content at the expense of content that is more conservative Buckley & Schafer (2022), Vogel *et al.* (2020). Despite the fact that conservative participants were only placed on juries with other conservatives, and were responsible for moderating content that was largely conservative, they nonetheless viewed the system as less legitimate than liberals because some participants were against the regulation of online speech in principle.

While we did not ask our moderator participants about the legitimacy of traditional moderation in our study, Pan et al. Pan *et al.* (2022) compared the perceived legitimacy of different moderation practices. Out of moderation by paid contractors, algorithms, expert panels, or user juries, expert panels were seen as the most legitimate. However, user agreement with the moderation decisions that were made were more important than the moderation process when determining legitimacy. The requirement for collective decision-making among peers a digital jury necessitates, while nonetheless providing greater agency to users, can thus be seen as a drawback when considering the legitimacy of a platform moderation system, especially if jurors disagree about what decisions to make. Indeed, in our study as in Fan and Zhang's, participants expressed concern about the motivations and opinions of fellow jurors and the ability for minority opinions to hold deliberations hostage. The lower trust in our moderation system we observed may be related to this as well. Overall, there is likely not a single one-size-fits-all solution to platform moderation that perfectly balances user agency, fairness, and legitimacy while keeping polarization in check, and the degree to which such a balance is possible may depend on the partisanship of platform users. The most effective solutions would therefore be catered to different user populations or communities, so the degree of expert involvement, algorithmic moderation and user input might be allowed to vary (Section 6.2).

6.1 Broader impacts and ethical considerations

Our work explores a paradigm shift in how social media platforms function—affecting which types of content are acceptable; how users interact with content, the platform, and each other; and the role of users in platform governance and moderation. Because social media platforms play a major role in day-to-day discourse worldwide, the broader impacts of our work are potentially far-reaching.

Results from our experiment do not indicate that introducing a digital jury moderation would necessarily reduce polarization online. Nonetheless, regardless of whether a jury moderation system alters the digital landscape and user attitudes for the better, it may still be worthwhile to implement because of the values it upholds. As outlined by Fan and Zhang, a digital jury can provide platform users with greater agency, and empower them to exercise their right to self-government as digital citizens. Participants rated the legitimacy, equality, trustworthiness, fairness, and care of the jury moderation process favorably compared to the top-down moderation currently used by most major social media platforms, and after the experiment described the system as supporting the democratic values of popular sovereignty, equality and justice, as well as the humanistic value of trust in humans. Therefore, implementing digital juries is worth considering due to the way they positively transform the relationship between social media platforms and their users.

A digital jury moderation system should be a peer-based system that promotes and protects the human rights of social media users. Such a system could empower users and enhance their self-governance, agency and civil rights on these platforms, enabling them to more easily determine for themselves which content is acceptable, ensure their voices are heard, and exert their own control over the platform ecosystem. These rights would need to be fostered by juries that are transparent in their decision-making and accountable for their verdicts. How transparency and accountability are ensured would rely on precisely how the moderation system is implemented, but clear community guidelines, consequences, and feedback for rule violators, as well as jury decisions that are publicly available, would be necessary. Finally, accidental misuse can be guarded against by instilling competence in jurors via training. The most effective form such training might take is still a matter of open debate, but text-based directions, instructional videos or online training sessions with example cases, along with feedback from existing jurors, are all potential candidates.

6.2 Limitations and future work

There were several limitations to our study that future work could address. While this study explored differences in how liberal and conservative participants perceived digital jury moderation and its outcomes, as well as demonstrated that digital juries are largely seen as fair and effective across the political spectrum, a key limitation that limits the statistical power and generalizability of our results is our sample size. While our choice to recruit participants from Reddit allowed us to develop rapport with participants, easily address questions or concerns, and made coordinating the logistics of our study easier, our overall recruitment pool was rather small. Of the 309 Redditors that expressed interest in our study, only 215 completed our screening questionnaire, and of those only 59 participated in the study. This coupled with how our sample was divided into six total groups meant that our largest group contained only 19 participants, and the smallest only three. Furthermore, our ongoing recruitment was a slow process, and the long amount of time between when our content selection was finalized to when the study was completed meant that some of our posts (e.g., those about the aftermath of the US 2020 election) were outdated by the time our study began. Future studies engaging in similar work would benefit from recruiting a larger sample more quickly, and ensuring that the content used is contemporary.

Additionally, while the open availability and convenience of the PushShift archive for filtering relevant content and querying the state of posts at the time of submission made Reddit an attractive platform to compare with our digital jury implementation, we note that moderation as it occurs on Reddit, where moderators are community members with a high degree of agency, is an imperfect comparison with the largely employee-driven moderation on Facebook or X. Although Reddit moderation is "top-down" in the sense that subreddit moderators are able to unilaterally impose their decisions on members, conducting a similar study on other platforms where moderation is more centralized might yield different results, though the closed nature of these platforms may pose a challenge to researchers. Conducting the study directly on an existing platform such as Reddit with actual moderation taking place could also improve validity and user engagement. Directly contrasting results when testing jury systems on different platforms, such as Reddit vs. Facebook or X, could also reveal key differences.

Our implementation of a digital jury was only one potential implementation, and there are several potential configurations that could be created in the future. Some might make more sense for platforms that are smaller and less diverse, and others for large platforms whose users span the globe. It may make more sense for a small forum focused on a specific interest (e.g., cycling) to select jurors from its entire user base regardless of geography, but large platforms have greater latitude to select groups of jurors who are representative cross-samples of the site population, and neither over- nor under-represent particular groups or classes. Having separate pools of jurors per region would also make sense in this case, as the breadth of different content would make it more likely that local, contextual knowledge is required to perform effective moderation. Large platforms may also find it more feasible to financially compensate jurors (or otherwise reward them) and provide more in-depth training. Participants could be allotted more (or less) time for deliberation, or the deliberation requirement could be eliminated altogether: one participant acknowledged the potential scalability issues of the deliberative approach, and suggested a hybrid model, where "*discussion might be reserved for content that has already been elevated from a larger pool of asynchronously voting moderators who can make quick/gut decisions.*" Because of the volume of content that must be moderated on platforms such as Facebook, we believe that jury moderation would be best used as a complementary form of moderation, with algorithms filtering out the content that is easiest to moderate, such as nudity or violence, though further work is necessary to investigate this.

Another point raised by participants was a mechanism for resolving disagreements, rather than tabling cases for later. We envisioned such cases as simply being shown to a different jury in the future, but other methods for resolving hung juries might exist, such as allowing juries to choose a "default action" to take in the event of a tie. One point participants noted that could also serve as a solution to this problem would be relaxing the unanimity requirement. While the boon of the unanimity requirement was that it gave all perspectives during deliberation equal weight, and prevented the majority from superseding the minority without engaging in any efforts to sway their opinion, in practice it would be easy for one malicious actor to hold a group hostage by refusing to alter their choices no matter the efforts made to convince them (which did occur in one of our conservative groups). Maintaining unanimity also becomes more difficult as jury sizes increase. Thus a simple or weighted majority vote might also be a way to counter these issues, as might a vetting system for potential jurors. We also noticed that our participants were much less likely to remove content relative to Reddit moderators, and future work could investigate this phenomenon.

Another avenue for future exploration could be to explore the effect of ideological composition on the decisions that digital juries make. Our study only included juries composed of members with similar ideological leanings (either fully liberal or fully conservative), potentially contributing to the large degree of agreement between jurors in each group. An interesting addition would be the inclusion of juries with mixed ideological compositions, examining the ease with which they make decisions and whether those decisions differ meaningfully from their ideologically uniform counterparts. This might also be extended to the types of content that are shown to jurors and platform end users: in our experiment, we only showed users content that aligned with their own views or was ideologically neutral; we never showed users content from sources that were from an opposing ideology. Thus examining whether juries made different decisions based on the ideological positions of the content they moderated may also merit further investigation. Examining different jury selection models, such as lottery selection, which could generate juries more representative of different user populations, would also be worth exploring.

Finally, users in our study were blind to the type of moderation system they interacted with (either Reddit moderation or our digital jury). Future work could also explore if revealing that content was moderated by a jury reduces polarization and increases trust in the platform and other users; knowledge of the democratic nature of the system and the fact that every user can play a part in governance may have a additional positive impact on user attitudes and behavior than differences in moderation decisions alone. Work could also explore whether there are changes in user perceptions and behavior before vs. after serving as a juror.

7 Conclusion

In this work, we investigated whether democratizing content moderation on social media platforms would impact the polarization of end users. We compared measures of polarization for participants who interacted with content moderated via an implementation of a peer-based digital jury moderation system versus traditional, top-down moderation and found that the moderation system used did not significantly impact the polarization of participants. However, we replicated findings from Fan and Zhang's work showing that digital juries had high perceived democratic legitimacy, efficacy, and procedural justice. Additionally, users had similar perceptions of the content they saw regardless of the moderation system used, indicating that deploying a digital jury moderation system would have the benefit of providing users agency in platform governance without adversely impacting user experience. There are several limitations (sample size, study duration, content that was easy to moderate) and potential variations and improvements (as discussed above) that future work could address. Despite these limitations, it also revealed several important considerations that would be important to explore in future work. Ultimately, peer-based moderation systems such as digital juries represent promising participatory mechanisms that are seen as just, legitimate, and effective by platform users, and can enable their civic involvement in social media ecosystems.

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

Aswad, E. M. (2018) The future of freedom of expression online. *Duke L. & Tech. Rev.*, **17**, 26.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Haohan Chen, M. B., Hunzaker, F., Lee, J., Mann, M., Merhout, F. and Volfovsky, A. (2018) Exposure to opposing views on social media can increase political polarization. *PNAS*, **115**, 9216–9221. https://doi.org/10.1073/pnas.1804840115.

Barberá, P. (2020) *Social Media, Echo Chambers, and Political Polarization, Book Section 3*, pp. 34–55. Cambridge University Press.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. and Bonneau, R. (2015) Tweeting from left to right: Is online political communication more than an echo chamber? *Psychol. Sci.*, **26**, 1531–1542. https://doi.org/10.1177/0956797615594620.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. and Blackburn, J. (2020) The pushshift reddit dataset. *In Proc. AAAI ICWSM*, **14**, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347.

Bell, K. (2023) *Reddit CEO Steve Huffman Defends API Changes in AMA*.

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L. and Klarkowski, M. (2021) Don't you know that you're toxic: Normalization of toxicity in online gaming. *Proceedings of the 2021 CHI conf. human factors in computing systems*, 1–15.

Bonica, A., McCarty, N., Poole, K. T. and Rosenthal, H. (2013) Why hasn't democracy slowed rising inequality? *J. Econ. Perspect.*, **27**, 103–124. https://doi.org/10.1257/jep.27.3.103.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Res. Psychol.*, **3**, 77–101. https://doi.org/10.1191/1478088706qp063oa.

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

Cambre, J., Klemmer, S. R. and Kulkarni, C. (2017) Escaping the echo chamber: Ideologically and geographically diverse discussions about politics. *Proc. ACM human-computer interaction Extended Abstracts, CHI EA '17*, 2423–2428.

Chandrasekharan, E., Gandhi, C., Mustelier, M. W. and Gilbert, E. (2019) Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM human-computer interaction*, **3**: Article 174, 1–30. https://doi.org/10.1145/3359276.

Coleman, K. (2021) *Introducing Birdwatch, A Community-Based Approach to Misinformation*. Twitter.

Coleman, K. (2022) *Building a better birdwatch*. Twitter.

Common, M. K. F. (2020) Fear the reaper: how content moderation rules are enforced on social media. *Int. Rev. Law, Comput. Technol.*, **34**, 126–152. https://doi.org/10.1080/13600869.2020.1733762.

Conger, K., Mac, R. and Isaac, M. (2022) Confusion and frustration reign as Elon Musk cuts half of Twitter's staff. *The New York Times*.

Conroy, M., Feezell, J. T. and Guerrero, M. (2012) Facebook and political engagement: A study of online political group membership and offline political engagement. *Comput. Hum. Behav.*, **28**, 1535–1546. https://doi.org/10.1016/j.chb.2012.03.012.

Cook, C. L., Patel, A. and Wohn, D. Y. (2021) Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Front. Hum. Dynam.*, **3**. https://doi.org/10.3389/fhumd.2021.626409.

De Gregorio, G. (2020) Democratising online content moderation: A constitutional framework. *Comput. Law Security Rev.*, **36**, 105374. https://doi.org/10.1016/j.clsr.2019.105374.

Oliva, T. D. (2020) Content moderation technologies: Applying human rights standards to protect freedom of expression. *Hum. Rights Law Rev.*, **20**, 607–640. https://doi.org/10.1093/hrlr/ngaa032.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D. and Bazarova, N. N. (2018) Upstanding by design: Bystander intervention in cyberbullying. *Proc. ACM human-computer interaction, CHI '18*, 1–12.

Dosono, B. and Semaan, B. (2019) Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proc. 2019 CHI conf. human factors in computing systems, CHI '19*, pp. 1–13. Association for Computing Machinery, New York, NY, USA.

Dosono, B. and Semaan, B. (2020) Decolonizing tactics as collective resilience: Identity work of AAPI communities on Reddit. *Proc. ACM human-computer interaction*, **4**, 1–20. https://doi.org/10.1145/3392881.

Druckman, J. N., Peterson, E. and Slothuus, R. (2013) How elite partisan polarization affects public opinion formation. *Amer. Pol. Sci. Rev.*, **107**, 57–79. https://doi.org/10.1017/S0003055412000500.

Fan, J. and Zhang, A. X. (2020) Digital juries: A civics-oriented approach to platform governance. *Proc. ACM human-computer interaction, CHI '20*, 1–14.

Fiesler, C., Jiang, J., McCann, J., Frye, K. and Brubaker, J. (2018) Reddit rules! Characterizing an ecosystem of governance. *Proc. of the international AAAI conf. on web and social media*, **12**. https://doi.org/10.1609/icwsm.v12i1.15033.

Frankovic, K. (2016) Belief in conspiracies largely depends on political identity. *YouGov*, **27**, 17–20.

Gallagher, R. J., Reagan, A. J., Danforth, C. M. and Dodds, P. S. (2018) Divergent discourse between protests and counter-protests: BlackLivesMatter and AllLivesMatter. *PLoS One*, **13**, e0195644. https://doi.org/10.1371/journal.pone.0195644.

Garimella, K. (2018) *Polarization on social media*. Ph.d. dissertation, Aalto University.

Gerrard, Y. (2018) Beyond the hashtag: Circumventing content moderation on social media. *New Media Soc.*, **20**, 4492–4511. https://doi.org/10.1177/1461444818776611.

Gilbert, S. A. (2020) "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a public scholarship site on Reddit: A case study of r/askhistorians. *Proc. ACM human-computer interaction*, **4**, 1–27. https://doi.org/10.1145/3392822.

Gillespie, T. (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A. and West, S. M. (2020) Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Rev.*, **9**. https://doi.org/10.14763/2020.4.1512.

González-Bailón, S. and Lelkes, Y. (2023) Do social media undermine social cohesion? A critical review. *Social Issues Policy Rev.*, **17**, 155–180. https://doi.org/10.1111/sipr.12091.

Gorwa, R., Binns, R. and Katzenbach, C. (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*.

Granovetter, M. S. (1973) The strength of weak ties. *Amer. J. Sociol.*, **78**, 1360–1380. https://doi.org/10.1086/225469.

Haimson, O. L., Delmonaco, D., Nie, P. and Wegner, A.. (2021) Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM human–computer interaction*, **5**, 1–35. https://doi.org/10.1145/3479610.

Hettiachchi, D. and Goncalves, J. (2019) Towards effective crowd-powered online content moderation. *Proc. ACM Aus. human-computer interaction, OZCHI'19*, 342–346.

Hu, X. E., Whiting, M. E. and Bernstein, M. S. (2021) *Can Online Juries Make Consistent, Repeatable Decisions?* p. Article 142. Association for Computing Machinery.

Im, J., Zhang, A. X., Schilling, C. J. and Karger, D. (2018) Deliberation and resolution on wikipedia: A case study of requests for comments. *Proc. ACM Hum.-Comput. Interact.*, **2**, 1–24. https://doi.org/10.1145/3274343.

Instagram. (2019) Combatting misinformation on Instagram.

Iyengar, S. and Hahn, K. S. (2009) Red media, blue media: Evidence of ideological selectivity in media use. *J. commun.*, **59**, 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x.

Jahanbakhsh, F., Zhang, A. X. and Karger, D. R. (2022) Leveraging structured trusted-peer assessments to combat misinformation. *Proc. ACM human-computer interaction*, **6**.

Jhaver, S., Appling, D. S., Gilbert, E. and Bruckman, A. (2019) "Did you suspect the post would be removed?": Understanding user reactions to content removals on Reddit. *Proc. ACM human-computer interaction*, **3**, 1–33. https://doi.org/10.1145/3359294.

Jhaver, S., Vora, P. and Bruckman, A. (2017) Designing for civil conversations: Lessons learned from Changemyview. *Georgia Institute of Technology*.

Jiang, J. A., Scheuerman, M. K., Fiesler, C. and Brubaker, J. R. (2021) Understanding international perceptions of the severity of harmful content online. *PLoS One*, **16**, e0256762. https://doi.org/10.1371/journal.pone.0256762.

Jiang, S., Robertson, R. E. and Wilson, C. (2019) Bias misperceived: The role of partisanship and misinformation in Youtube comment moderation. *Proc. of the international AAAI conf. on web and social media*, **13**, 278–289.

Juneja, P., Subramanian, D. R. and Mitra, T. (2020) Through the looking glass: Study of transparency in Rreddit's moderation practices. *Proc. ACM human-computer interaction*, **4**, Article 17.

Keeter, S. and Igielnik, R. (2016) Can likely voter models be improved? *Report, Pew Research Center*, **7**, 2016.

Knobloch-Westerwick, S. and Meng, J. (2009) Looking the other way:selective exposure to attitude-consistent and counterattitudinal political information. *Commun. Res.*, **36**, 426–448. https://doi.org/10.1177/0093650209333030.

Koebler, J. and Cox, J. (2018) The impossible job: Inside Facebook's struggle to moderate two billion people. In *Vice Motherboard*.

Kulkarni, C., Cambre, J., Kotturi, Y., Bernstein, M. S. and Klemmer, S. R. (2015) Talkabout: Making distance matter with small groups in massive classes. *Proc. ACM human-computer interaction, CSCW '15*, 1116–1128.

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

Mahar, K., Zhang, A. X. and Karger, D. (2018) Squadbox: A tool to combat email harassment using friendsourced moderation. *Proc. ACM human-computer interaction, CHI '18*, 1–13.

Mancini, M. N. (2019) *Development and validation of the secondary traumatic stress scale in a sample of social media users*. M.s. thesis,. Cleveland State University.

Matias, J. N. (2019) The civic labor of volunteer moderators online. *Social Media + Society*, **5**, 2056305119836778.

Matias, J. N. and Mou, M. (2018) Civilservant: Community-led experiments in platform governance. *Proc. ACM human-computer interaction, CHI '18*, 1–13.

Meta (2021) How Meta's third-party fact-checking program works.

Meta. (2023) Community standards enforcement report, Q3 2023.

Milosh, M., Painter, M., Van Dijcke, D. and Wright, A. L. (2020) *Unmasking partisanship: How polarization influences public responses to collective risk*. University of Chicago, Becker Friedman Institute for Economics Working Paper.

Moravec, P., Minas, R. and Dennis, A. R. (2018) Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper*. https://doi.org/10.2139/ssrn.3269541.

Murphy, L. and Cacace, M. (2020, July 8) Facebook's civil rights audit – final report.

West, S. M. (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media Soc.*, **20**, 4366–4383. https://doi.org/10.1177/1461444818773059.

Nelimarkka, M. (2019) A review of research on participation in democratic decision-making presented at SIGCHI conferences. toward an improved trading zone between political science and HCI. *Proc. ACM human-computer interaction*, **3**, 1–29. https://doi.org/10.1145/3359241.

Nemeth, C. (1977) Interactions between jurors as a function of majority vs. unanimity decision rules. *J. Appl. Soc. Psychol.*, **7**, 38–56. https://doi.org/10.1111/j.1559-1816.1977.tb02416.x.

Niu, X.-м. and Jiao, Y.-н. (2008) An overview of perceptual hashing. *ACTA ELECTONICA SINICA*, **36**, 1405.

Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., and Bernstein, M. S.. (2022) Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proc. ACM human–computer interaction*, **6**, 1–31. https://doi.org/10.1145/3512929.

Pennycook, G., Bear, A., Collins, E. T. and Rand, D. G. (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Sci.*, **66**, 4944–4957. https://doi.org/10.1287/mnsc.2019.3478.

Pennycook, G., Cannon, T. D. and Rand, D. G. (2018) Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol.*, **147**, 1865–1880. https://doi.org/10.1037/xge0000465.

Pennycook, G. and Rand, D. G. (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl Acad. Sci.*, **116**, 2521–2526. https://doi.org/10.1073/pnas.1806781116.

Pew Research Center. (2017) The partisan divide on political values grows even wider.

Reddit. (2023) Transparency report: January to June 2023.

Robison, J. and Mullinix, K. J. (2016) Elite polarization and public opinion: How polarization is communicated and its effects. *Pol. Commun.*, **33**, 261–282. https://doi.org/10.1080/10584609.2015.1055526.

Roth, Y. and Pickles, N. (2020) *Updating Our Approach to Misleading Information*. Twitter.

Rule, C. and Schmidtz, A. J. (2018) *The New Handshake: Online Dispute Resolution and the Future of Consumer Protection*. American Bar Association.

Schirch, L. (2020) 25 *spheres of digital peacebuilding and peacetech*. Technical report. Toda Peace Institute and Alliance for Peacebuilding.

Schirch, L. (2023) The case for designing tech for social cohesion: The limits of content moderation and tech regulation. *Yale J. Law Humanities, forthcoming*. https://doi.org/10.2139/ssrn.4360807.

Seering, J. (2020) Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM human-computer interaction*, **4**, Article 107.

Seering, J., Kaufman, G. and Chancellor, S. (2022) Metaphors in moderation. *New Media & Society*, **24**, 621–640. https://doi.org/10.1177/1461444820964968.

Shen, Q. and Rose, C. (2019) The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proc. of the 3rd workshop on abusive language online*, pp. 58–69. Association for Computational Linguistics, Florence, Italy.

Shen, Q., Yoder, M. M., Jo, Y. and Rosé, C. P. (2019) Perceptions of censorship and moderation bias in political debate forums. In *Twelfth international AAAI conf. on web and social media*.

Silverman, C. (2016) *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook*. BuzzFeed News.

Simas, E. N., Clifford, S. and Kirkland, J. H. (2020) How empathic concern fuels political polarization. *Amer. Pol. Sci. Rev.*, **114**, 258–269. https://doi.org/10.1017/S0003055419000534.

Smith, J., O'Brien, T., Carr, H., Crowe, P., and Rice, M.. (2020) *Polis and the Political Process*. Demos.

Spohr, D. (2017) Fake news and ideological polarization:filter bubbles and selective exposure on social media. *Business Information Rev.*, **34**, 150–160. https://doi.org/10.1177/0266382117722446.

Squirrell, T. (2019) Platform dialectics: The relationships between volunteer moderators and end users on Reddit. *New Media Soc.*, **21**, 1910–1927. https://doi.org/10.1177/1461444819834317.

Suhay, E., Bello-Pardo, E. and Maurer, B. (2018) The polarizing effects of online partisan criticism: Evidence from two experiments. *The Int. J. Press/Politics*, **23**, 95–115. https://doi.org/10.1177/1940161217740697.

Thach, H., Mayworm, S., Delmonaco, D. and Haimson, O. (2022) (In) visible moderation: A digital ethnography of marginalized users and content moderation on twitch and Reddit. *New Media Soc.*, **26**, 4034–4055. https://doi.org/10.1177/14614448221109804.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. and Nyhan, B. (2018) Social media, political polarization, and political disinformation: A review of the scientific literature.

Vaccaro, K., Sandvig, C. and Karahalios, K. (2020) "At the end of the day Facebook does what it wants": How users experience contesting algorithmic content moderation. *Proc. ACM human-computer Interaction*, **4**, 1–22. https://doi.org/10.1145/3415238.

Vashistha, A., Cutrell, E., Borriello, G. and Thies, W. (2015) Sangeet swara: A community-moderated voice forum in rural india. *Proc. ACM human-computer interaction, CHI '15*, 417–426.

Veglis, A. (2014) Moderation techniques for social media content. In *Int. conf. on social computing and social media*, pp. 137–148. Springer.

Vlachokyriakos, V., Comber, R., Ladha, K., Taylor, N., Dunphy, P., McCorry, P. and Olivier, P. (2014) Postervote: expanding the action repertoire for local political activism. *Proc. of the 2014 conf. on designing interactive systems*, 795–804.

Vogel, E., Perrin, A., and Anderson, M.. *Most americans think social media sites censor political viewpoints*. Report, Pew Research Center, August 19, 20202020.

Wilson, A. E., Parker, V. A. and Feinberg, M. (2020) Polarization in the contemporary political and media landscape. *Curr. Opin. Behav. Sci.*, **34**, 223–228. https://doi.org/10.1016/j.cobeha.2020.07.005.

Zhang, A. X., Hugh, G. and Bernstein, M. S. (2020) Policykit: building governance in online communities. *Proc. of the 33rd annual ACM symposium on user interface software and technology*, 365–378.

Zhang, A. X. *et al.* (2018) A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proc. of the the web conference 2018, WWW '18*, pp. 603–612. International World Wide Web Conferences Steering Committee.

Zuckerman, E. (2021) *Mistrust: Why Losing Faith in Institutions Provides the Tools to Transform Them*. WW Norton & Company.