Effects of Digital Jury Moderation on the Polarization of Social Media Users

by

Christopher Micek

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

June 2022

APPROVED:

Associate Professor Erin T. Solovey, Major Thesis Advisor

Associate Professor Gillian M. Smith, Thesis Reader

Associate Professor Craig A. Shue, Head of Department

Abstract

As polarization among political officials and the public at large has increased dramatically in recent years, the social media landscape has followed suit. The increased prevalence of disinformation, inflammatory rhetoric, and harassment online has augmented polarization in turn, propelling a feedback loop resulting in the erosion of democratic norms. Effective content moderation can mitigate this problem, but many existing major social media platforms employ moderation systems that are autocratic, with decisions about what sorts of content are acceptable shaped by the platforms themselves, potentially disregarding community knowledge and cultural nuances when making these decisions. We conducted a two-phase study to compare how a social media platform employing a traditional, top-down moderation system would impact outcome measures of the polarization of its users in comparison to a peer-based *digital jury* moderation system, which promises to make platform users active participants in digital governance. While 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 digital jury moderation system has the benefit of providing users agency in platform governance without adversely impacting user experience.

Acknowledgements

I would like to first and foremost thank my advisor, Professor Erin Solovey, who has been a source of tremendous guidance, insights, edits, and resources, and an overall outstanding mentor. I would also like to thank Professor Gillian Smith, the reader for this thesis, who additionally helped shape the project in its early stages. Additional thanks goes to Zijian Guan, who helped to collect posts, develop survey questions, and get the ball rolling. Finally I would like to thank Shruti Mahajan, Jonathan Coelho, and Mikaela Milch, who helped rate the content that was shown to our participants.

Contents

1 Introduction

When the major social networking and file sharing websites (Facebook, Twitter, YouTube, Reddit, etc.) were introduced in the early 2000s, many were hopeful that these platforms would enable the creation and sharing of content and ideas in an inclusive, participatory, and global way [40]. Company executives and researchers were optimistic that, with enough adoption worldwide, social media could provide "dramatic new possibilities for pluralizing flows of information and widening the scope of commentary, debate, and dissent" [27]. In many respects, these hopes have come to fruition: as of March 2022, Facebook has a global user base of 2.9 billion people [7], and the two largest social media platforms in the US, YouTube and Facebook, were used by 81% and 69% of US adults, respectively [10]. News can break on Twitter before it appears in traditional mass media [47]. Indeed, Internet-based communication and social media platforms have been used so far throughout the 21st century by citizens of repressive regimes to coordinate protest activities and to communicate effectively, in some cases circumventing government interference. The ability to quickly disseminate news and mobilization information played critical roles in the impeachment and removal of President Joseph Estrada of the Philippines in 2001 [96]; the revolutions of the Arab Spring in the early 2010s [46]; the Ukrainian Revolution in 2014 [16]; and the Hong Kong protests of 2019–20 [91].

Despite these successes, social media has also played a role in amplifying and disseminating sensational divisive content [93, 108]. 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 [35], the competing #BlackLivesMatter and #AllLivesMatter social media campaigns and the associated protests and counter-protests [36], and a partisan divide in compliance with public health measures to mitigate the effects of the COVID-19 pandemic. (According to Milosh et al. [67], out of several covariates, partisanship was the most significant predictor for local mask use, with Republican voters significantly less likely to wear masks than Democrats.) How is it that tools for fostering connections could be misused to sow division instead?

To reconcile these disparate outcomes, it is useful to first clarify what, exactly, we mean when we refer to "social media platforms." Adopting the definition from Gillespie's *Custodians of the Internet* [40], social media *platforms* "are online sites or services that

a) host, organize, and circulate users' shared content or social interactions for them,

b) without having produced or commissioned (the bulk of) that content;

c) are built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit; and

d) platforms do, and must, moderate the content and activity of users, using some logistics

of detection, review, and enforcement."

Platforms, therefore, are systems designed with the needs of various stakeholders (e.g., end users, advertisers, moderators, lawmakers, the public, and the platforms themselves) in mind. Above all, platform employees are tasked with creating and maintaining systems people will want to use, while also ensuring their venture is profitable. From the context of affordance theory [39, 98], the design decisions platforms make to this end *afford* these stakeholders the means for navigating this complex array of incentives. Such affordances influence the construction and negotiation of authority and trust on platforms [98], shaping social norms and behaviors that continue to evolve over time. Platforms have an incentive to ensure they have a sizeable user base (to create and share content, and entice advertisers with a large audience) and sustained user engagement (so user behavior can be measured and advertisers can target specific user demographics). The ability to follow other users, easily share content, and engage with content and communities that users find meaningful are consequently common features on several platforms [90, 98, 73]. Platforms, in turn, may curate personalized feeds of content to keep users engaged, and moderate this content to maintain a safe, welcoming environment for users and to preclude legal issues or regulatory action [105]. While these design elements arguably well-intentioned, malicious actors have nonetheless been able to circumvent or weaponize them for their own ends.

The scale and ease with which harmful content can spread poses challenges for several platforms. Facebook, for instance, has deployed a large-scale fact-checking operation that appears to be at least somewhat effective at mitigating the spread of misinformation [75], though it is not without its limitations [76]. A common theme of existing social media moderation structures on most mainstream platforms such as Facebook, Instagram, or Twitter 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 [40]), with employees or contractors performing moderator duties. However, content may require context-specific knowledge and information about local sociocultural norms to be moderated effectively [59, 54], and enforcement of moderation policies can be uneven: an independent civil rights audit [69] 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, content users perceive as toxic is still prevalent [108], and moderation decisions may still lack transparency [56]. Given the shortfalls of existing governance structures on social media platforms, could other more democratic methods prove more effective?

Digital jury moderation, a system piloted by Fan & Zhang [32] which places the responsibility for moderation in the hands of end users, presents a promising alternative. Instead of moderators existing as separate entities from day-to-day users, a digital "jury of one's peers" is selected among a group of users to place content to be moderated "on trial" and reach

a consensus on what, if any, consequences should be employed. Digital juries are perceived as more transparent and procedurally just than existing practices, while also ensuring that members of online communities play an active role in moderating the content they interact with. However, while Fan & Zhang explored the design considerations for such a system and assessed how moderators perceived it, no existing work compares the attitudes of end users interacting with political social media communities moderated with an implemented digital jury versus traditional, top-down moderation.

This thesis examines 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. We aim 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, 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, while 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.

1.1 Definitions of Polarization

We distinguish between three different types of polarization:

Ideological polarization The difference in ideological self-placement, e.g. on a liberal–conservative scale [50].

Affective polarization How positively or negatively partisans feel about members of an in-party versus an out-party [100].

Social polarization How likely partisans are to socially self-segregate from members of an out-party [100].

2 Related Work

2.1 Political Polarization and Social Media

The current level of extreme polarization in the US has been decades in the making [102]. Since roughly 1980, both major US political parties have polarized steadily [17] as they

developed distinct and cohesive policy agendas in the wake of the civil rights movement, and today the median Democrat and Republican are more ideologically divided than ever before [1]. Partisanship has become closely linked to social identity, and perceived ideological differences have generated remarkable levels of hostility between members of opposing parties [108]. 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. One study [30] asked participants to evaluate policy positions bolstered by either weak or strong arguments. When participants were told that political parties were unpolarized on the issues, they based their policy support on the strength of the arguments; when instead they were told that the parties were highly polarized on the issue, they favored the position of their own party, regardless of the strengths of the arguments they were shown. Other work [94] has shown that people with high dispositional empathic concern tend to be more affectively polarized, as people tend to display more empathy toward their in-group than their out-group. This is especially true for those who strongly identify with a particular party. The opposite was true for social polarization, however, as empathic participants were more likely to feel comfortable with outparty contact. There is also evidence that the behavior of party elites may drive polarization among the electorate [102], where politicians' inflammatory rhetoric demonizing their opponents and the absence of bipartisan cooperation, coupled with the tendency of partisan media to select for polarizing content, contribute to their constituents' unfavorable views of the out-party and exaggerating perceptions of ideological differences [108]. When traditional news reporting frames elite polarization as problematic, however, partisans are more likely to hold favorable views of bipartisan cooperation, despite little effect on their preferred policy positions [83]. Independents, however, reported significantly less political interest, trust, and perceived government efficacy when polarization was made salient, regardless of how it was framed.

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 [37], where they are only exposed to information that reinforces their political views and are isolated from those with opposing views, are to blame for increasing polarization [14, 97]. This "cyberbalkanization" leads to increased opportunities for enclave deliberation, where conversations only occur among like-minded people. While not inherently negative (political group membership on Facebook, for example, was correlated with offline political engagement [25]), 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 [14]. However, other studies have shown that cross-cutting interactions on social media are more frequent than commonly believed [102, 13, 14], with exposure to ideologically diverse news and opinions more com-

mon 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 [43], 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 [37]. Rather than decreasing polarization, however, cross-cutting interactions with political content on social media may exacerbate it instead. Bail et al. [12] showed that exposure to messages with opposing political views from political elites can increase ideological polarization. The authors recruited a large sample of Republicans and Democrats, who were tasked with following a Twitter bot that would share messages tweeted by political leaders from the opposing party. After one month, post-experiment surveys showed that Republicans had become significantly more conservative, while Democrats became slightly more liberal (though this difference was not statistically significant). This seems to indicate the presence of a backfire effect (at least among partisans who use Twitter), whereby exposure to opposing views caused participants to double down on their existing views. Suhay et al. [100] 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 also stronger among Republicans than Democrats. Leaving social media could potentially reverse these effects: Allcott et al. [8] surveyed people who had deactivated their Facebook accounts, and found that their ideological polarization had decreased. However, their knowledge of current events also decreased, which could indicate that lower exposure to political news overall could explain the reduction in polarization.

It is interesting to note the asymmetric partisan divide in the measures of polarization observed by Bail et al. [12] and Suhay et al. [100] between Democrats and Republicans. While such partisan differences have been noted elsewhere [1, 35, 67, 102], the specific causes for these asymmetries are varied and not clearly defined. Suhay et al. attribute them to the current political climate and power distribution in Congress, differences in the strength of party identity, and cultural and historical factors. Personality traits that may be correlated with party identity, such as openness or conscientiousness, as well as the asymmetry of clickbait and hyper-partisan content (i.e., its greater prevalence on the right than the left) may also play a role [14].

This thesis focuses mainly on political polarization in the United States, but it is also worth mentioning that polarization is a global phenomenon. The types of content users find objectionable and their relative perceived severity differ from country to country [54], and media diets and cultural and political landscapes vary widely. While the US has a two-party system with plurality electoral rules, other nations may have multi-party political systems

with different voting rules (e.g., majoritarian, proportional, or mixed). In an analysis of the Twitter audiences for the political parties of several different democratic countries, Urman [103] found that polarization was highest in countries with two-party systems with plurality electoral rules, and lowest in countries with multi-party systems and proportional voting. Research on polarization on social media in the US is thus not necessarily generalizeable to other nations.

2.2 Influences of Platform Activity on User Behavior Dynamics

Given that certain types of content can exacerbate polarization on social media platforms, how is it that end users come in contact with polarizing content? Stewart et al. [99] model interactions on social media as "influence networks," whereby the flow of information is constrained by collective decisionmaking: individuals continuously revise their perspectives by weighing their views against the views of others until a consensus is reached. The authors developed a voting game to assess how the set of perspectives that were visible to players would influence those they adopted, simulating real-world networks of political discourse. The assortment of influence throughout the network had a substantial impact on the eventual winner. "Information gerrymandering," which occurs when the assortment of influence is asymmetric between teams, favored teams that were tightly connected, or had more relative influence on the opposing team. The researchers also found that including bots, which supported their own team regardless of others' choices, increased their team's odds of winning if their influence was asymmetric. Analysis of several real-world political social media networks, as well as US Congress and EU legislatures, revealed such influence gaps, indicating they play a role in existing partisan interactions. Therefore, parties and malicious actors are incentivized to create such influence gaps, and employ strategic use of bots or encourage a zero-sum worldview on social media platforms, as this can provide disproportionate advantages.

We can see this illustrated in the sorts of posts that are widely circulated on platforms: language referring to political opponents in posts from news sources and politicians on Facebook and Twitter was a substantially better predictor for engagement and whether content would be shared than the presence of emotional language [81] (though negative language is also a strong predictor of engagement and virality [95], and posts referring to opposing parties tended to be negative). An investigation into which factors contributed to toxic online disinhibition and influenced users' likelihood of engaging in hostile flaming behavior when engaging in discussions online found that lack of eye contact with other users was the chief contributing factor, though the anonymity and invisibility platforms can afford users also influenced negative behaviors [61].

However, as alluded to in Stewart et al.'s work, evidence shows that in addition to individuals, collective campaigns by political parties, hyper-partisan media outlets, conspiracy

groups, and state actors take advantage of these phenomena, at times deploying networks of bots to distort narratives and propagate misinformation. While the efforts of state-sponsored groups in Russia and China are well-known, political parties in several democratic countries have also used bots to inflate follower counts, promote certain hashtags, and amplify certain narratives over others [18]. A study analyzing the deployment of political bots on Twitter prior to the 2020 US election found that just a few thousand bot accounts were able to generate spikes in conversation about real-world events comparable to the volume of activity of humans, and exacerbated content consumption produced by users with similar political views [33]. The bots were more likely to share right-leaning or conspiracy-related content; such "attention hacking," coupled with sharing content that is sensational and novel, exploits the capacity for such content to drive engagement, contributing to decreased trust of mainstream media, increased misinformation, and radicalization by far-right extremist groups [65].

Platforms have several tools at their disposal to tackle these issues. Major platforms such as Facebook, Twitter, and YouTube already employ automated moderation systems to remove some harmful content [59, 42], 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 [40]. In the case where content is misinformative or untrustworthy, Facebook [5], Twitter [86], and Instagram [2] have employed fact-checking operations that apply labels warning users of the issue. Several researchers have either developed their own extensive credibility indicators [110], 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 (though users did spend more time considering the validity of flagged content) [68]. Users were also more likely to trust headlines they had seen before, regardless of whether they were flagged [76], and attaching warnings to fake news articles also increased trust in articles without warnings [75]. 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 hyperpartisan or fake sources regardless of their political affiliation [77]. 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.

2.3 Content Moderation

Social media platforms engage in content moderation to safeguard their users and foster welcoming environments they will enjoy and engage with, while navigating the legal and political dynamics of speech online. Here we outline existing moderation approaches, their shortcomings and user perceptions, and potential alternative approaches.

2.3.1 Existing Approaches

Social media platforms are essentially middlemen, coordinating interactions between speakers and potential audiences and retaining valuable user data in exchange [40]. As Tarleton Gillespie notes in *Custodians of the Internet*, they can also set the terms of this exchange: "the required technical standards, what counts as a commodity, what is measured as value, how long content is kept, and the depth and duration of the relationship." They may offer content creators "a share of the advertising revenue or not, and [get] to decide how much, to whom, and under what conditions... [This] requires excluding some to serve others: those who provide unwanted goods, those who game the system, those who disrupt the entire arrangement."

Thus to this end, all the major platforms have rules (e.g., policies on hate speech, violence, harassment, misinformation, pornographic content, spam, and copyright infringement) [59] and the means of enforcing them. Broadly speaking, platforms employ two main moderation philosophies (sometimes in tandem): moderation is either 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, Twitter, 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 [89]. Depending on the severity and frequency of offenses, moderators might resort to 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 process lacks civic participation from users when making these decisions.

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 [49]. 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. 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 [105].

Because of the large volumes of user-generated content posted to social media platforms (Facebook users, for instance, create billions of posts per day [59]), 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 used at scale by Facebook, YouTube, Twitter, and others to classify

user-generated content through either pattern matching or prediction, employing perceptual hashing [72], machine learning classification, and other techniques to judge whether such content is appropriate or prohibited [42]. 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 [42, 59]. 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 [38], illustrating limits to language-based approaches. 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 [42], and risks undermining free speech and equitable information access [28].

Reddit The majority of content on Reddit is publicly visible, and its free developer API makes it easy for researchers to retrieve and analyze this content, so it is widely studied. As our study's real-world data was collected from Reddit, we devote this next section to understanding user behavior and moderation on Reddit, specifically.

Reddit is organized into several interest-based communities, or "subreddits," where users may post text, links, images, or videos that other users may comment on. Community moderators bear the responsibility of creating and enforcing subreddit rules, as well as Reddit's sitewide content policy, and have wide discretion to remove content and ban users from participating in their subreddit, while regular users may report rule violations to moderators [34]. They may also modify the subreddit's appearance by changing its banner, icons, and other UI components, and can assign labels to users and content with customized flairs. A key feature of Reddit's ecosystem is the karma system, which allows users to upvote and downvote posts or comments to influence their visibility to other users [98], purportedly incentivizing users to meaningfully contribute to discussions, but in practice contributing to a bandwagon effect, whereby popular opinions are most visible and unpopular opinions can be suppressed.

An analysis of subreddit rules shows they fall into two categories: prescriptive (what users should do), and restrictive (what users should not do), and most employ a combination of both [34]. Prescriptive rules include formatting requirements for posts, instructions to follow the Reddiquette (a list of informal dos and don'ts created by site users that embody common values of the site) [6], and a reminder to "be civil"; restrictive rules include rules forbidding spam, harassment, hate speech, and trolling, and are most common on politics subreddits. In the case of egregious rule violations, e.g., for excessive hateful or toxic speech or calls to violence, subreddits may be quarantined or banned by Reddit administrators. This community-level moderation has shown to decrease the activity of such communities

even after migrating to different platforms (in the case of r/TheRedPill, r/The_Donald and r/Incels), even though the radicalization of their users was undiminished [21, 82].

Like the other platforms described, Reddit employs automated moderation tools to alleviate the workload of human moderators. Aside from automatic spam detection, which operates at the site level, Automoderator is a tool that subreddit moderators can configure to perform certain moderation actions automatically. Moderators can adapt the tool to the needs of their particular subreddits, e.g., automatically removing posts with certain keywords or website domains, removing posts from users with too little karma or accounts that are too new, and enforcing formatting requirements for posts [52]. Automoderator can also be used to automatically provide reasons for content removals, which educate users about community norms and reduce the likelihood of future removals and rule violations [53].

2.3.2 Perceptions of Moderation

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.

Cook et al. [26] explored the differences in user perceptions of toxicity and transparency on platforms that were commercially moderated versus platforms moderated by user volunteers. They found that there was no significant difference in the perceived toxicity of commercially moderated versus user-moderated platforms, although they were significantly more understanding of moderation practices on user-moderated platforms; participants who were more confident reported engaging in more toxicity management behaviors (e.g., reporting toxic posts) themselves, although they perceived these efforts as ultimately ineffective. On commercially moderated platforms, users wanted the company to take more responsibility for moderation; on user-moderated platforms, users wanted moderators and the posters of content to have responsibility. Pan et al. [74] 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.

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 [70]. On Reddit, this is especially true for users with new accounts, whose posts may be removed by Automoderator for violating unlisted account age or karma requirements. 69% of analyzed content removals were not accompanied by any moderator feedback [56]. A survey of Reddit users whose content was removed showed most disagreed with these removals, and several were confused and angry about these decisions. 37% did not understand why the removals had occurred, and 29% of respondents were frustrated by the removals and be-

lieved them to be unjust [51]. Similarly, Haimson et al. [44] 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 [55, 92], 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.

Values on Reddit differ between users and moderators, with a survey of moderators showing they desired communities to be 56% more undemocratic than users did, but also had a greater desire to improve the safety and trustworthiness of their communities [107]. While moderators have large latitude to make moderation decisions that safeguard their communities, platforms lack tools for formalizing input from users when making these governance decisions. A key takeaway is that most platforms could benefit from more transparent moderation practices that take user perspectives into account.

2.3.3 Alternative Approaches to Platform Moderation and Governance

Researchers have developed several other tools and platforms that give users more agency in governance decisions and facilitate discourse between users with different beliefs.

Cambre et al. [19] piloted a variant of the Talkabout platform [60], which allowed users to engage in synchronous online political discussions in small groups via video calls. Though the authors envisioned creating groups of users with diverse political beliefs, in practice only 2 of the 30 participants who indicated their political affiliation when signing up were conservatives, and there was a large degree of attrition after recruitment, leading to the creation of many small groups where participants held similar views. Nonetheless, participants found the discussions valuable and informative, though future iterations could be improved with discussion moderators or more specific discussion prompts.

A tool developed by Matias & Mou, called CivilServant, aids Reddit moderators in designing, conducting, and analyzing data from field studies evaluating community phenomena and policy implementations (e.g., to what extent newcomers follow subreddit norms, or whether pinned moderator messages warning users to downvote untrustworthy news articles would influence their spread) [66]. To provide moderators on Reddit with a more flexible automated moderation solution, Chandrasekharan et al. [20] developed an open source AI tool (Crossmod) to recommend moderation actions. The tool is a collection of classifiers trained both on the moderation decisions from 100 different subreddits, plus meta-rule classifiers to detect violations of Reddit sitewide norms. The result was an ensemble of classifiers trained with cross-community learning that based its recommendations on an aggregation of prior moderation decisions, rather than manually defined rules, that was easy for moderators of other communities to adapt to their needs. Similarly, Zhang et al. [109] developed a framework (PolicyKit) that could be deployed on multiple platforms, allowing users to develop

and enforce their own governance structures. PolicyKit represents governance as a set of procedures to be followed, and could be used to express a variety of models, from deliberative juries, to elections, reputation systems, content filters, and constitutions governing when policies can be changed.

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 decisionmaking, over-reliance on efficiency, and inconsistent enforcement of decisions [24]. While more user-centric approaches do not necessarily eliminate these issues, sharing decisionmaking power has the potential to increase transparency and accountability, and incentivizes building up and maintaining online communities [89].

Crowdsourced Moderation Several studies also developed tools for and tested the effectiveness of platform users engaging directly in moderation. Hettiachchi & Goncalves [45] analyzed how 28 participants recruited from Amazon Mechanical Turk moderated political content from Twitter. A collection of tweets with the "Obama" keyword with a uniform distribution of sentiments were presented to participants over 20 days for moderation, in which participants were tasked with determining whether they were appropriate for a general audience. Tweets that were labeled as inappropriate had low sentiment scores, and contained profanity, hate speech, grammatical errors, or were off-topic, similar to removal reasons from Reddit's subreddit moderation [34].

Vashistha et al. [104] created Sangeet Swara, a community-moderated voice forum for users in rural India with limited internet access. Users could interact with the platform exclusively via phone call audio and keypad presses, and submit voice recordings of songs, poems, and other cultural content. Submitted content could then be played back in a personalized feed, and users could upvote or downvote it to influence its ranking. Users were highly engaged with the platform, but notably the content was uncontroversial, as the domain was entertainment rather than politics.

Squadbox [63], 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. Users were appreciative of the collaborative nature of the moderation effort and found the system easy to navigate and flexible enough to accommodate their use preferences, though found that the moderation of itself was a lot of work.

Fan & Zhang [32] 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 the study of this thesis 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. Fan & Zhang present a five-stage model outlining such a digital jury system, comprising jury selection (who should be a juror, how large should the jury be), onboarding (what incentives should be offered to jurors, what preparation is necessary), the case trial (what types of contents and evidence jurors can see, as well as how it is presented and how a verdict is delivered), consensus (how/if jurors should communicate with each other, what the consensus method should be, how long jurors can deliberate), and finally enforcement (what records of the trial are available, and how a verdict is enforced).

A related study assessing whether digital juries could make consistent, repeatable decisions [48] by obfuscating juror identities for repeated collaboration for similar moderation cases found that individual jurors made similar decisions for similar cases, but that this was only true if juries were allowed to deliberate. However, this consistency arose in part from group polarization, whereby initial majority opinions swayed the final decisions of the group. This may pose challenges for real-world implementation, but other work [41] aims to model juries of different compositions (e.g., juries containing a certain proportion of Black or transgender members) via supervised machine learning, potentially providing future researchers a tool for assessing the effects of different jury compositions on decision outcomes.

Fan & Zhang's study assessed user attitudes toward two possible configurations of the system and compared them to user attitudes toward traditional platform-conducted moderation. 82 participants from Amazon Mechanical Turk were recruited and grouped into juries of approximately six members each (15 juries total). Each jury viewed the same three cases, which were "written by the first author and designed to be contextually nuanced and borderline with respect to violating standards." Juries were shown each case once, in a random ordering of three conditions: a control condition of "status quo" moderation, where the jurors were shown the case and the moderation decision rationale are shown without any user input; and two conditions requiring user interaction, where each juror is required to assign a toxicity score to the case (with 0 as least likely to cause harm, and 10 as the most likely), and choose from possible punishments for the content and the user who posted it. These interactive conditions were a "scalable" condition, where users submitted votes and decisions asynchronously, without deliberating, and an "immersive" condition, where jurors deliberated synchronously via chat before submitting their decisions. After each condition, participants were surveyed about the condition's democratic legitimacy—whether they thought the process was fair, and how satisfied they were with the outcome. Additionally, after all three conditions, participants were asked whether they thought jury decisions should merely be recommendations or enforced by platforms. Participants expressed a greater sense of procedural justice in the jury conditions compared to the control condition, and most preferred the immersive condition (57.5%) versus the scalable condition (35%); the control condition was the least preferred and was generally regarded as unfair. Thoughts were mixed about deliberating, with some participants distrusting the opinions and motivations of other ju-

rors and noting the potential for "groupthink" when making decisions. However, others believed that exposure to other perspectives was valuable. There was no group preference for verdicts being recommendations versus enforced, although individuals strongly favored one over the other, depending on whether they were more trusting of platforms or other users, respectively. Overall, while further research is necessary to explore other variations of the system, it is promising because it was viewed as democratically legitimate by a majority of participants, and would directly empower user stakeholders in the governance of social media platforms.

3 Methodology

The goals of our study were threefold: Ultimately, we wanted to assess the impact of implementing a digital jury moderation system on the polarization of social media end users relative to traditional top-down moderation, which is our main novel contribution. 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, our secondary novel contribution. To realize these goals, 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 & 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.

Thus the goals this thesis addresses can be summarized in 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?

3.1 Study Design

We conducted a two-phase study to examine how employing a digital jury moderation system would impact the polarization of moderators and end users, and how they would perceive such a system, in comparison with the top-down moderation approach used by several existing platforms. The main premise was to channel the same content through two different

moderation workflows, and then propagate the results of both to a researcher-controlled platform where "end user" participants could interact with content from an assigned moderation approach for a period of time.

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 (moderation decisions pre-selected without user input), and (ii) a digital jury condition, where the community would employ digital jury moderation instead. The digital jury platform from Fan & Zhang [32] serves as the basis for our digital jury. Fan & 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 (instead of majority) agreement. Therefore, these are conditions we adopted for our study as well.

Our **participants** were politically engaged users of social media, and placed 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 [37]. The hope 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?

The **content** that is moderated 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).

Participants were recruited on an ongoing basis for the two phases of the study:

Phase I: Moderation 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 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, or allowed to remain).

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. 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. Polarization for each group was measured before and after the experiment (see Measures below), in addition to subjective responses to answer **RQ2** and **RQ3**.

3.2 Measures

The measures and survey instruments used in the experiment are summarized below. See Appendix A for further details.

1. Measures of political affiliation and social media engagement, used to record demographics and screen subjects: Questions similar to those in Supplementary Materials Section 2.2 of [12], 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?"

2. Measure of political interest, adapted from Keeter & Igielnik [57]: "How engaged do you consider yourself with US politics? (Very Disinterested / Disinterested / Neutral / Interested / Very Interested)"

3. Measures of polarization: These are assessed before/after the experiment for both moderators and users. Three types:

   (a) **Ideological**: Questions about policy positions on a seven-point Likert scale, (e.g. stances on immigration, government regulations, homosexuality, etc.), as in [12].

   (b) **Affective**: Feeling thermometer from Appendix A of [100]. (How do you feel about Democrats/Republicans on a scale from 0 to 100?)

   (c) **Social**: Assessments of social distance, such as marriage/community preference questions in Appendix B of [100], or social distance questions after Table B1 in Appendix B of [94].

4. Subjective impressions of their experiences with the digital jury moderation system, obtained after the experiment via questionnaire, e.g. "To what degree do you think that the moderation process was fair?"; "Are there ways you think the moderation system could be improved?"

5. The Secondary Traumatic Stress Scale for Social Media Users (STSS-SM) in Appendix C of Mancini [64], to assess if participants experienced emotional distress during the course of the study.

3.3 Content Selection

We selected our content from the PushShift archive of Reddit submissions and comments [15] as the source of our content, since it is publicly available, easy to query, and designed with researchers in mind. Specifically, it generally stores posts in their initial state shortly after submission, allowing us to see any posts or comments that might have been removed in the state they were in prior to removal.

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

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).

We scraped the archives of subreddits listed in Table 1 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, 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 [58], who found that approximately 40% of content users encountered when browsing 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 & 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 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' would be considered moderation gray areas, and used for the moderation portion of the study; posts rated as 'OK' would be accepted as-is; and posts rated as 'Toxic' would not be shown to participants. 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 were 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 of post selection and rating 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.

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

Once we had completed our post selection, we began recruiting participants using a method similar to Jhaver et al. in [51]. We chose Reddit as our recruitment platform of choice, as messaging there or followup via email would allowed us to develop rapport with participants, address any questions or concerns about the study, and facilitate scheduling synchronous deliberation sessions among groups of participants for the moderation portion of the study more easily. Recruitment efforts began in August 2021 and were iterative, until March 2022 when our last user was enrolled.

We contacted a random subset of users who had commented on either the r/Liberal, r/Conservative, or r/politics subreddits (as we were looking to recruit participants who were regular social media users and politically engaged) 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 any future moderation decisions. Participants interested in participating completed our screening questionnaire (Appendix A.1).

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 "honeypot" questions in our screening questionnaire ("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. (See Appendix B for how the ideological score is calculated.) Participants who had an ideological score of zero were rejected, as this indicated no partisan leaning. 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.

3.5 Phase I: Moderation

Once participants were screened and recruited according to the criteria above, we obtained informed consent (participants were provided consent forms via email, and were able to ask researchers any questions they might have before submitting the signed form) as well as responses to the pre-experiment survey (see Appendix A.2) 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. (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 groups per political affiliation (five liberal and five conservative groups), and each was responsible for moderating 11 cases on their first day, and 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 (four conservatives dropped out after the first, however, and one had uncaught errors when taking the pre-experiment survey, leaving only demographic data available). Jury vote statistics in Section 4.2.1 use votes from all jurors available.

At the scheduled start time, participants logged into the study website, and were shown a web page with onboarding instructions for their task and how to use the moderation interface, as well as 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 as shown in Figure 1. 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 (see Section 3.6). Any links or multimedia from any of the 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, and could be initiated by entering a user name in the chat box on the lower left.

For each case component each juror needed to provide three assessments (slightly modified from Fan & Zhang):

- A toxicity score of content, defined as likelihood to cause harm (0-2 OK, 3-7 Borderline, 8-10 Toxic); participants were provided the same directions for scoring that researchers were in Section 3.3.
- Punishment for the content, if any (`unlist` from users' feeds, `delete` from the site, and `report` to authorities).

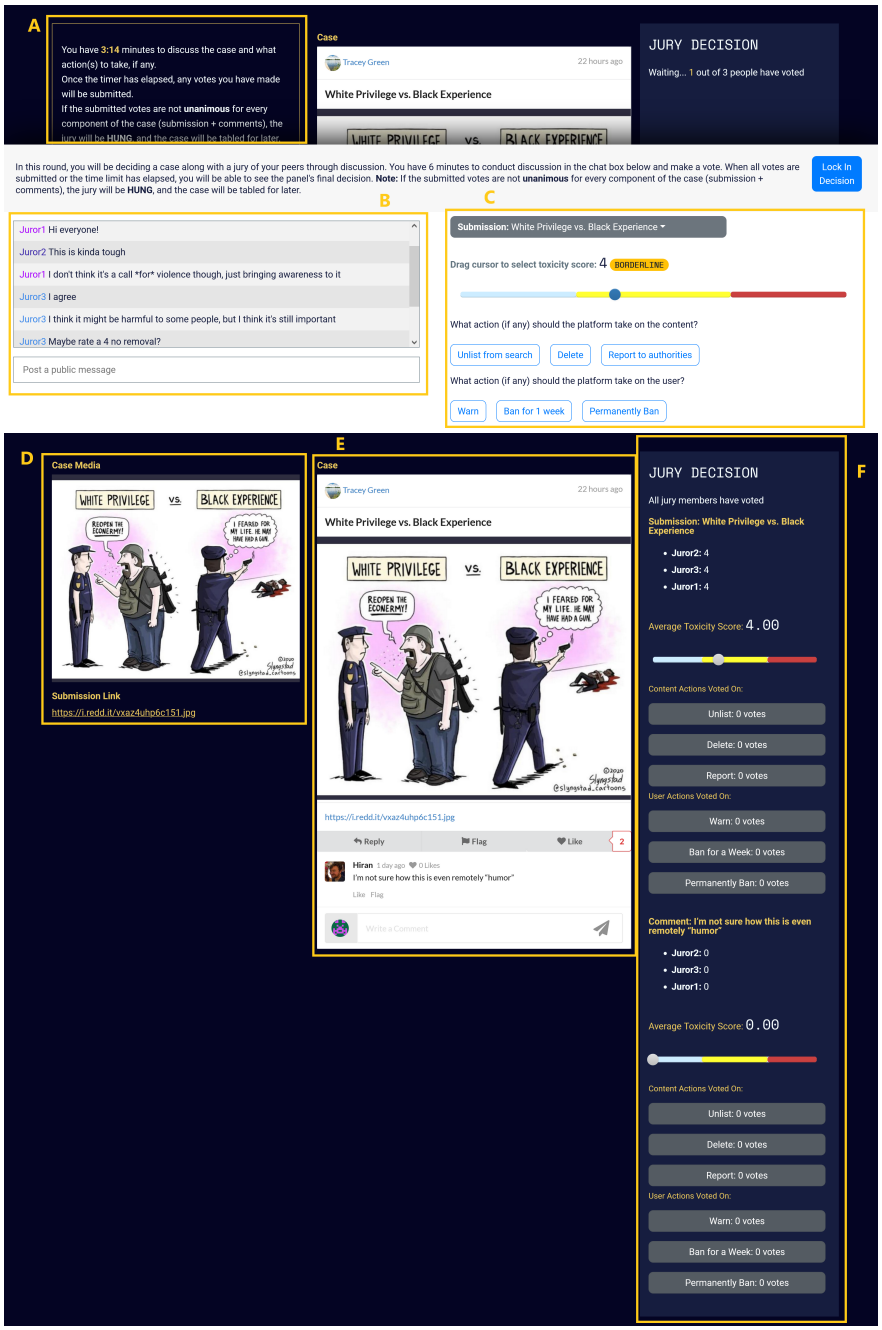

Figure 1: The moderation website user interface. (Top) The voting panel, which participants can toggle open or closed. The timer and instructions are visible in box A, and the chat interface jurors can use to deliberate is shown in box B. The voting interface, where users can select a toxicity score as well as any punitive actions for the component chosen in the dropdown menu, is shown in box C. (Bottom) The website layout once voting is complete. Any case links or media are shown in box D, and an image of the post and any comments as they would appear on a social media website are shown in box E, regardless of whether voting is complete. (The voting panel can obscure these if it is open.) The votes made by jurors for the toxicity score and any actions to take for each component are in box F.

- 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.

Jurors had six minutes per case to deliberate and vote on each component. 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 before proceeding.

Once the jurors had completed both days of moderation, they completed a post-experiment survey (see Appendix A.3) and were debriefed on the full nature of the experiment and the expected results. 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 (Appendix A.4. If any participants scored higher than 27 (moderate secondary traumatic stress), we followed up and provided information about the following mental health resources:

- The National Alliance on Mental Illness (NAMI) Helpline[1]
- The Crisis Textline[2]
- The International Association for Suicide Prevention (IASP) list of crisis centers and intervention services[3]

Five juror participants were contacted.

3.6 Phase II: User Interaction

After the juror groups had completed moderation for all of their content, the user groups began their portion of the experiment beginning January 2022. As in the prior portion, informed consent and pre-experiment survey responses were obtained for each participant

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

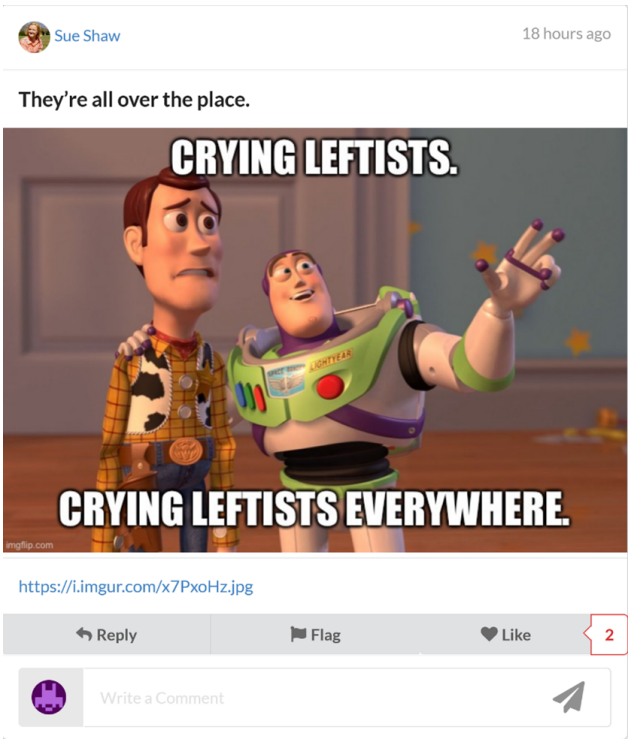

Figure 2: Sample post as it appeared on the Agora platform used in the user interaction phase of our study.

prior to the start of the experiment. Users were once again divided into either liberal or conservative group based on their screening questionnaire responses, and were further divided into groups that saw the results of either top-down moderation from Reddit, or the results of our jury moderation (2 partisan affiliations $\times$ 2 experiment conditions = 4 groups total). We recruited 24 participants who completed this portion of the experiment—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).

Participants in the top-down conditions were shown the current state of posts that were not moderated, as well as the results of Reddit's moderation of the moderated posts (e.g., if Reddit moderators elected to delete a post, it was not shown) for their respective political affiliations. Participants in the jury moderation condition saw the unmoderated posts, in addition to the results from our 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 in fact accepted as-is for the user group.

To host the content, we built a platform called Agora (Figure 2, built on top of the

Truman platform developed by DiFranzo et al. [29]. The Truman platform serves as a simulated social media feed, where researchers can developed customized content feeds for study participants. Researchers are able to assign posts and comments to different fake users, or "actors," control when they appear, and even create scripted interactions between actors. Our implementation of the Truman platform was simplified; the ability of participants to make posts of their own had been removed, as there were no scripted interactions with actors.

The user interaction portion of the experiment took place asynchronously over fourteen days for each participant, in order to ensure they engaged with the results of each type of moderation for an appreciable amount of time. We evenly divided the post collections to be shown for each group into fourteen sets (of six to fourteen posts, depending on the group), to be shown on each day of the experiment. We also developed engagement questions related to the content of one of the unmoderated posts shown to liberals or conservatives for each day of the experiment that participants would be required to answer, to ensure they engaged with the content.

Before they began, users were told that they would need to log into Agora for 5 – 10 minutes each day for two weeks to view a series of social media posts, and would need to answer questions about the content they saw each day. From the moment they logged in, users would have 24 hours to view the content for that day of the experiment; once 24 hours elapsed, the content would cycle 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 were also asked to complete the STSS-SM, as the jurors were above, and sent information about mental health resources if they scored 27 or above (indicating moderate secondary traumatic stress; four user participants were contacted). Once the two-week period elapsed, participants completed the post-experiment survey.

4 Results

4.1 RQ1: Changes in Polarization

We assessed polarization for both moderators and users using the measures described in Section 3.2. In addition to the raw responses to the Likert scale questions from our pre- and post-surveys (see Appendix A), we also used these responses to create continuous outcome measures for ideological, affective, and social polarization (see Appendix B). Our ideological, affective, and combined social scores (as well as the combined social score's marriage preference, social distance, and likeminded community 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

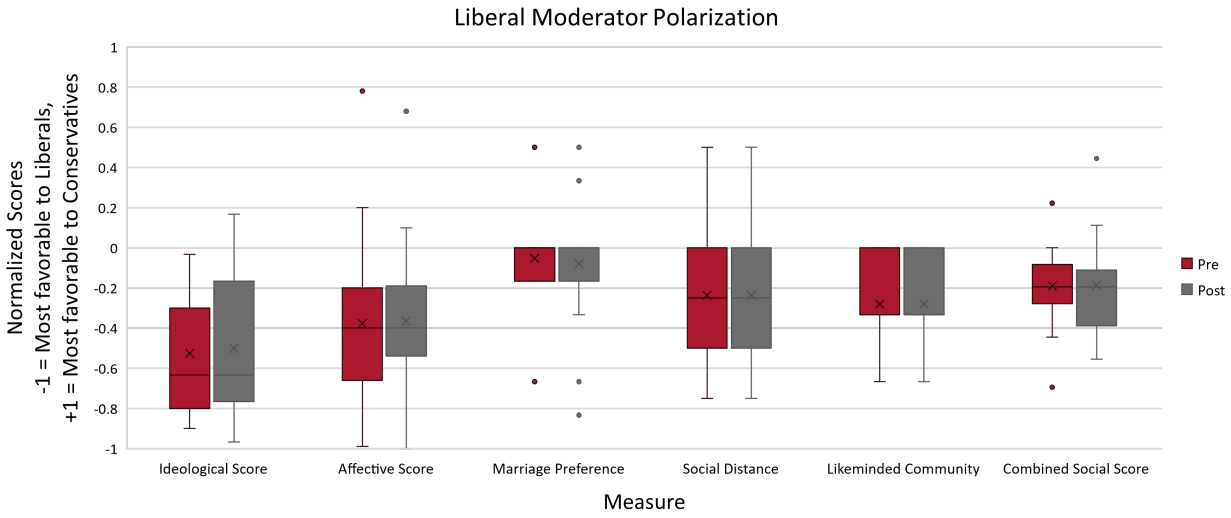

Figure 3: Continuous measures of polarization for liberal moderators. The Likeminded Community component has been negated to facilitate comparison with the other data (more negative indicates more favorable toward liberals).

their political leanings, the quantitative results from both phases of our study are separated for liberals and conservatives as well.

4.1.1 Phase I: Moderation

Our main goal was to investigate the polarization of end users, but to explore any differences that might exist between jurors, we conducted an exploratory analysis of polarization among our moderator participants as well. For the moderation portion of the study, Wilcoxon signed-rank tests comparing Likert responses from before and after the experiment for members of each affiliation yielded no significant changes for either liberals or conservatives. Paired t-tests were conducted to compare our 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 (Figures 3, 4), 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 $M = -0.021$, $SD = 0.167$; $t(16) = 2.287$, $p = 0.016$), but not for liberals (Figure 5). 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 ($t(28) = -2.149$, $p = 0.040$).

4.1.2 Phase II: User Interaction

After the user interaction portion of the experiment concluded, we performed similar analyses to those above for our main contribution, to determine to what degree polarization had

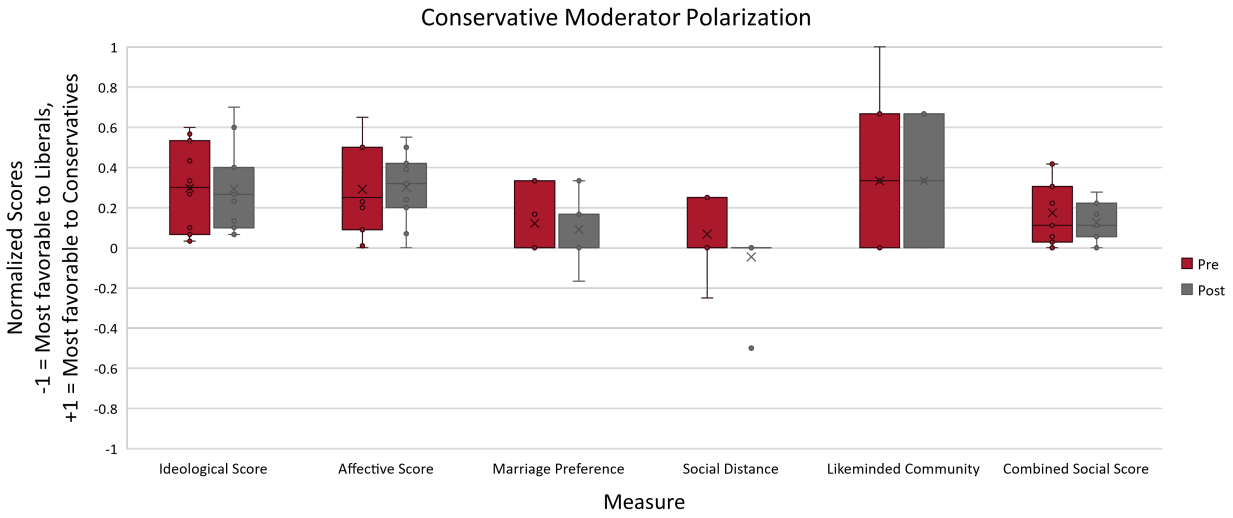

Figure 4: Continuous measures of polarization for conservative moderators.

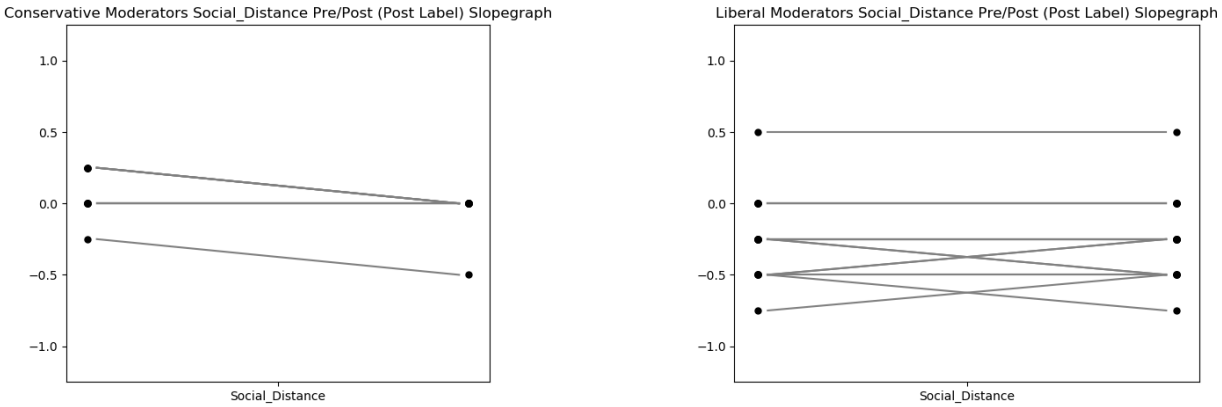

Figure 5: (Left) Slopegraph showing a significant difference in social distance for conservative moderators before vs. after the experiment ($p < 0.05$). (Right) Slopegraph showing social distance for liberal moderators before vs. after the experiment. There was no statistically significant difference.

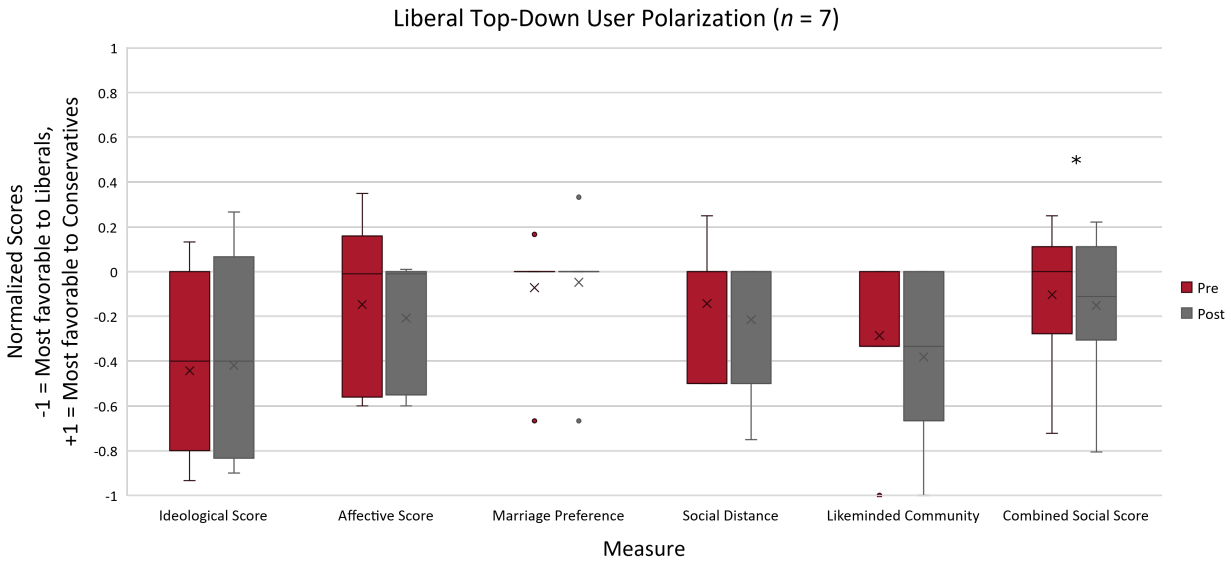

Figure 6: Continuous measures of polarization for liberal moderators from the top-down condition. The likeminded community component has been negated to facilitate comparison with the other data (more negative indicates more favorable toward liberals). A paired t-test indicated a significant decrease in the combined social score for social polarization.

changed and whether there were any significant differences between the different experiment groups. 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 $M = -0.151$, $SD = 0.338$; $t(6) = 2.828$, $p = 0.030$) (Figure 6). No significant differences were observed in outcome measures for liberal users in the jury condition (Figure 7), conservative users in the top-down condition (Figure 8), or conservative users in the jury condition (Figure 9). Changes in the continuous outcome measures for liberal and conservative users are shown in Figure 10 and Figure 11, respectively.

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 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 likeminded community components of the combined social score. However, there was a significant interaction between partisan affiliation and experiment condition for the combined social score ($F(1, 20) = 6.112$, $p = 0.023$). Table 2 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

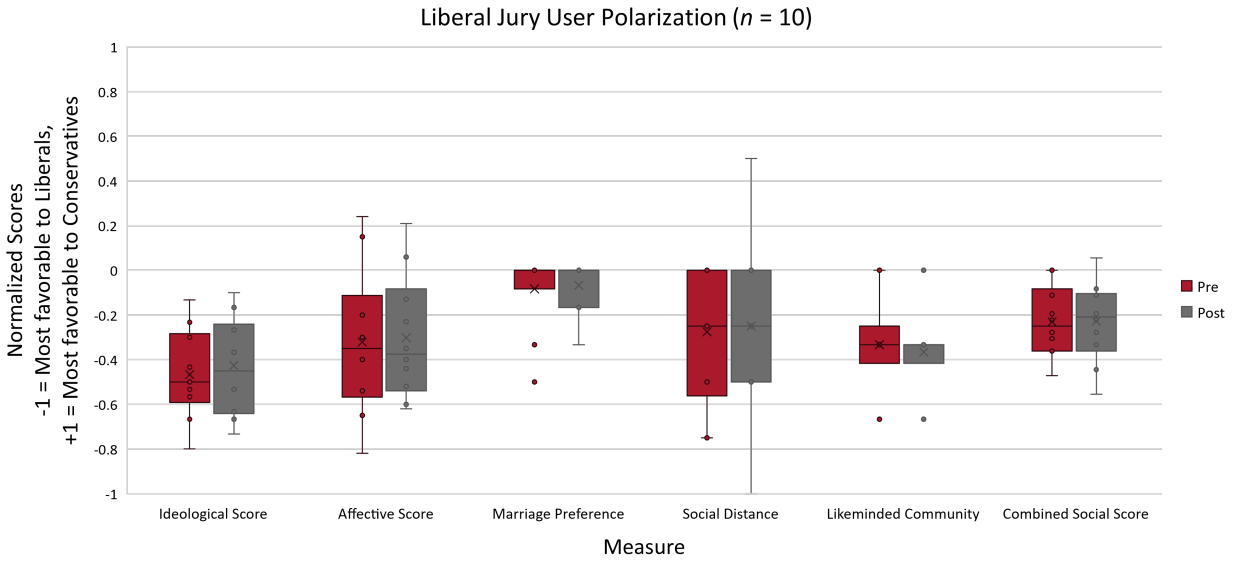

Figure 7: Continuous measures of polarization for liberal moderators from the jury condition. The likeminded community component has been negated to facilitate comparison with the other data (more negative indicates more favorable toward liberals).

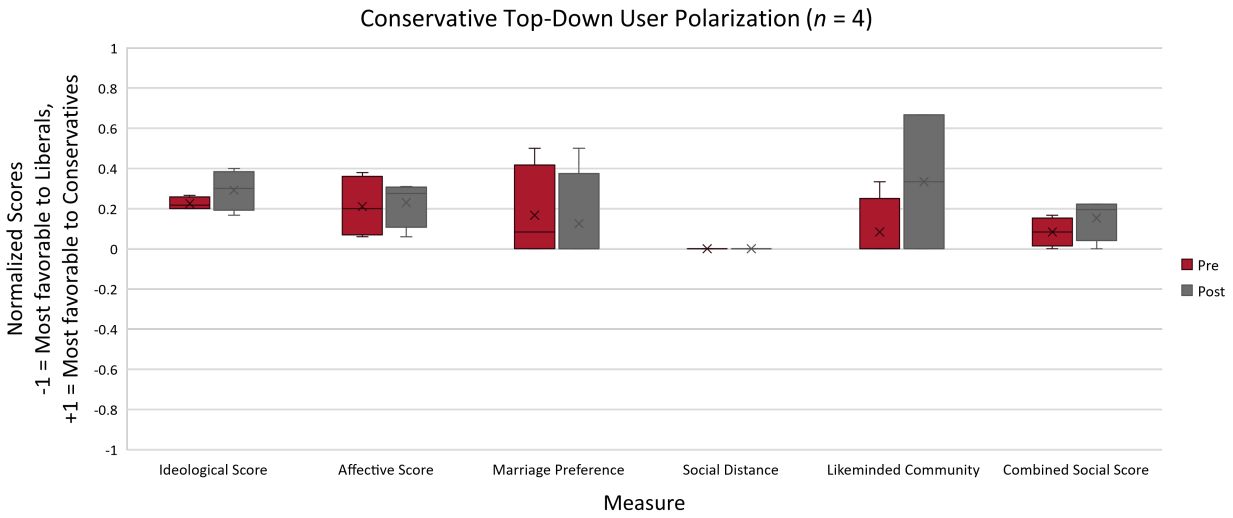

Figure 8: Continuous measures of polarization for conservative moderators from the top-down condition.

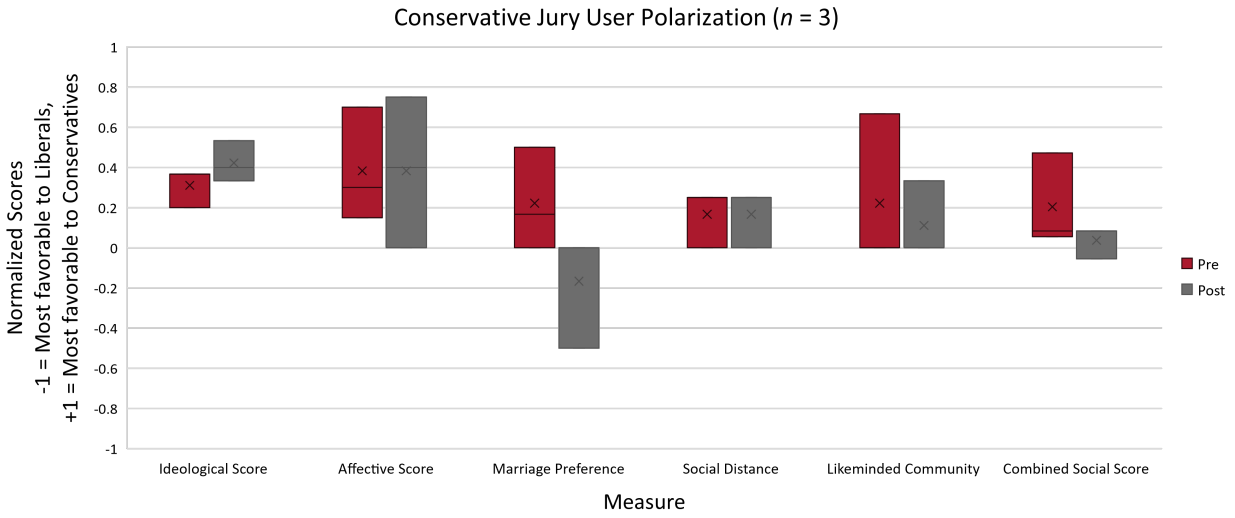

Figure 9: Continuous measures of polarization for conservative moderators from the jury condition.

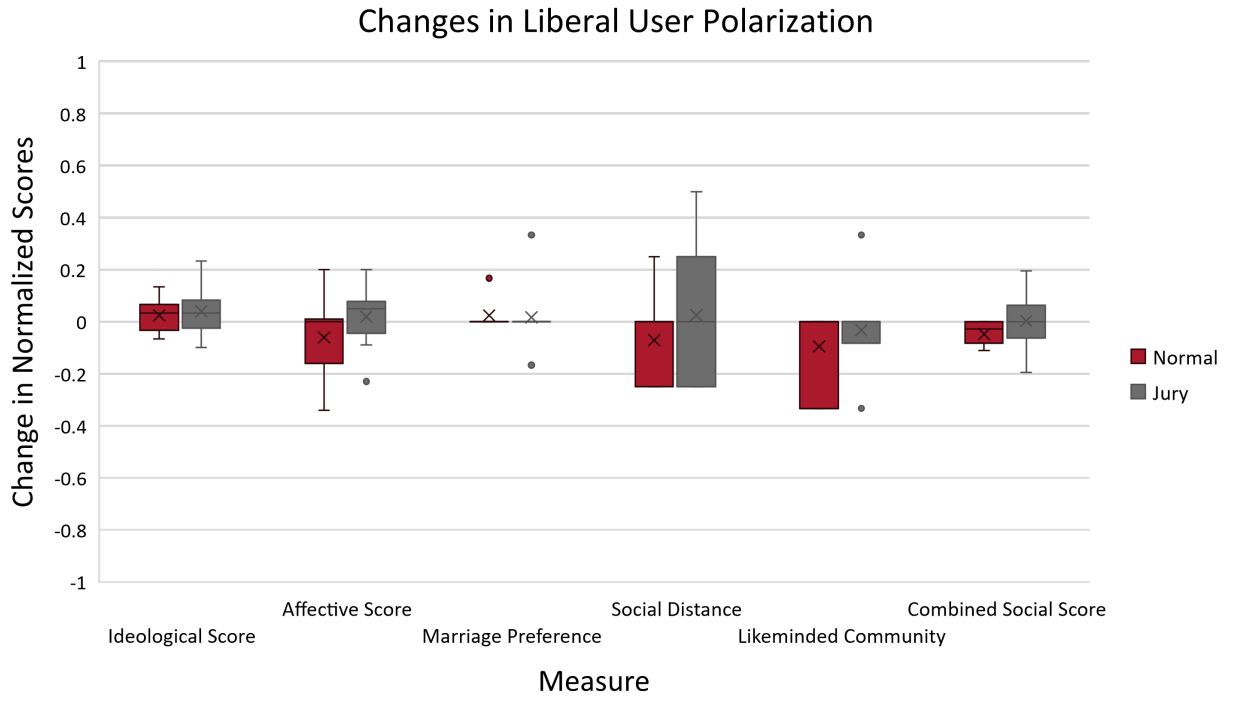

Figure 10: Changes in continuous measures of polarization for liberal users from both the top-down and jury groups. The Likeminded Community component has been negated to facilitate comparison with the other data (more negative indicates more favorable toward liberals).

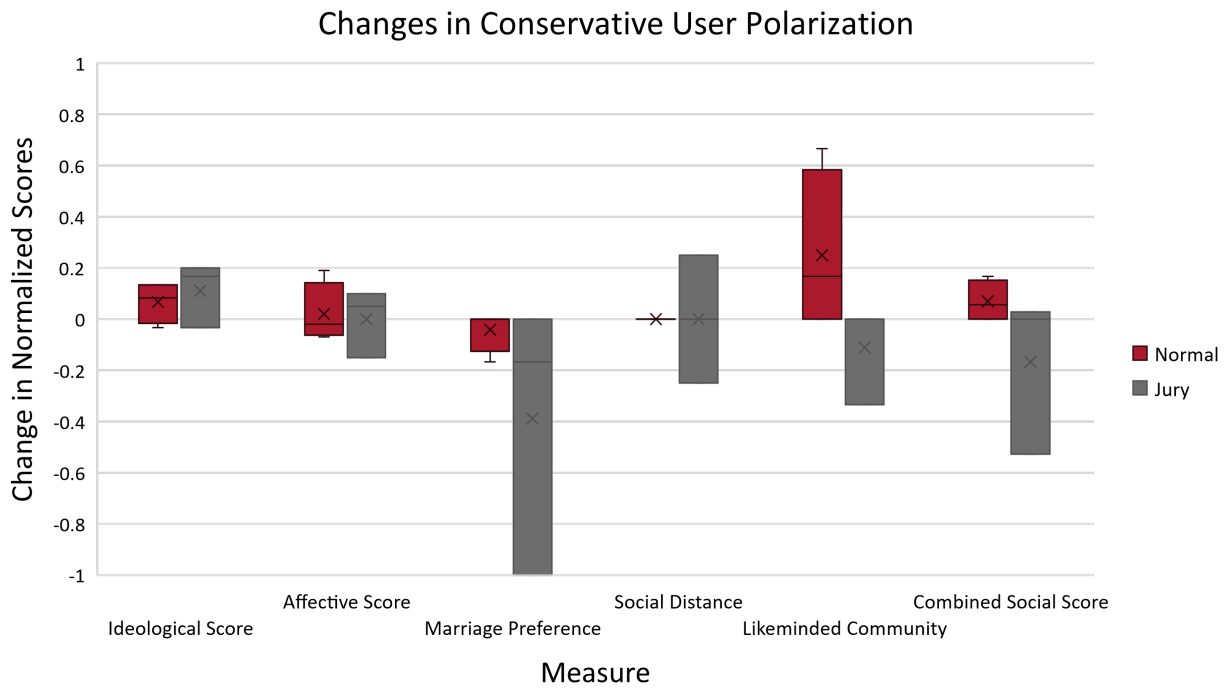

Figure 11: Changes in continuous measures of polarization for conservative users from both the top-down and jury groups.

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. Mann-Whitney U tests comparing changes in Likert survey responses between experiment conditions for either liberals or conservatives likewise did not yield any significant differences.

We also performed ordinal logistic regression to determine the effects of partisan affiliation and experiment condition on the changes observed for each of the Likert scale survey responses (analogous to the ANOVA above). We observed a significant interaction between partisan affiliation and experiment condition for responses to question 5a in our pre-

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

Table 2: 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 ($F(1, 20) = 6.112$, $p = 0.023$).

Affiliation	Condition	Mean	SD
Liberal	Top-down	-0.286	0.951
Conservative	Top-down	1.0	0.817
Liberal	Jury	0.1	0.568
Conservative	Jury	-1.0	1.732

Table 3: Changes in Likert responses to pre-experiment question 5a (Please rate the degree to which you agree or disagree with the following statements from 1 (Strongly disagree) to 7 (Strongly agree): Stricter environmental laws and regulations cost too many jobs and hurt the economy) for end user participants from each partisan affiliation and experiment condition. Ordinal logistic regression indicated a significant interaction between affiliation and condition ($p = 0.013$) as well as a significant main effect from condition ($p = 0.012$).

Affiliation	Condition	Mean	SD
Liberal	Top-down	0.286	0.488
Conservative	Top-down	0.75	0.957
Liberal	Jury	0.1	0.568
Conservative	Jury	-0.333	0.577

Table 4: Changes in Likert responses to pre-experiment question 11 (Imagine for a moment that you are moving to another community... In deciding where to live, how important would it be to live in a place where most people held political views similar to your own? (1 Not important, Somewhat important, Moderately important, 4 Very important)) for end user participants from each partisan affiliation and experiment condition. Ordinal logistic regression indicated a significant main effect from condition ($p = 0.033$).

experiment survey ($p = 0.013$) (used to calculate the ideological score; see Appendix A.2), as well as a significant main effect from experiment condition ($p = 0.012$). Changes are summarized in Table 3; see Table 5 for model parameters and statistics. We additionally observed a significant main effect of experiment condition on responses for question 11 ($p = 0.033$), which determines the likeminded community component of the combined social score. See Table 4 for the mean response changes for each group of participants; model parameters and statistics can be found in Table 6.

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 (one each for the 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. See Tables 7, 8, and 9 for model information. 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 ($p = 0.064$).

4.2 RQ2: Perceptions of Moderators and Users

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

4.2.1 Phase I: How Moderators Used and Viewed Jury Moderation

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 & 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 12. 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 (Figure 13). 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 (Figure 14). Thus our participants were much less likely to remove content compared to Reddit moderators overall.

We also replicated part of Fan & 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 15; see Appendix A.3 for questions). We largely corroborate the findings from their immersive condition, and additionally are able to separate results based on participants' political affiliations. Jurors of both affiliations believing 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 & 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

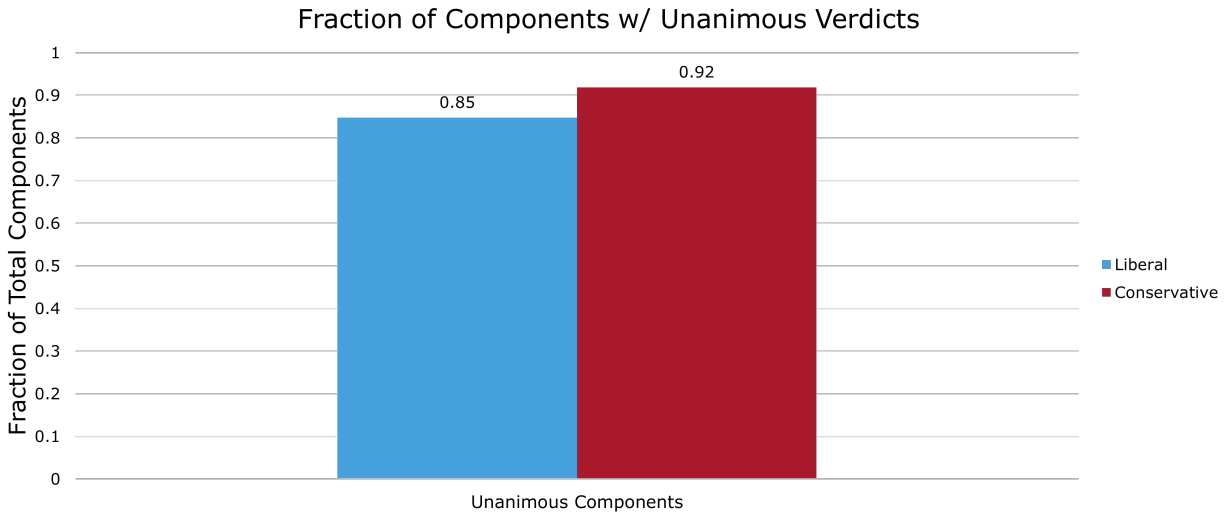

Figure 12: Fraction of case components with unanimous verdicts for liberal and conservative moderators. Components without unanimous verdicts received mixed votes from juries that were hung, and were allowed to remain unchanged when shown in the user interaction portion of the study.

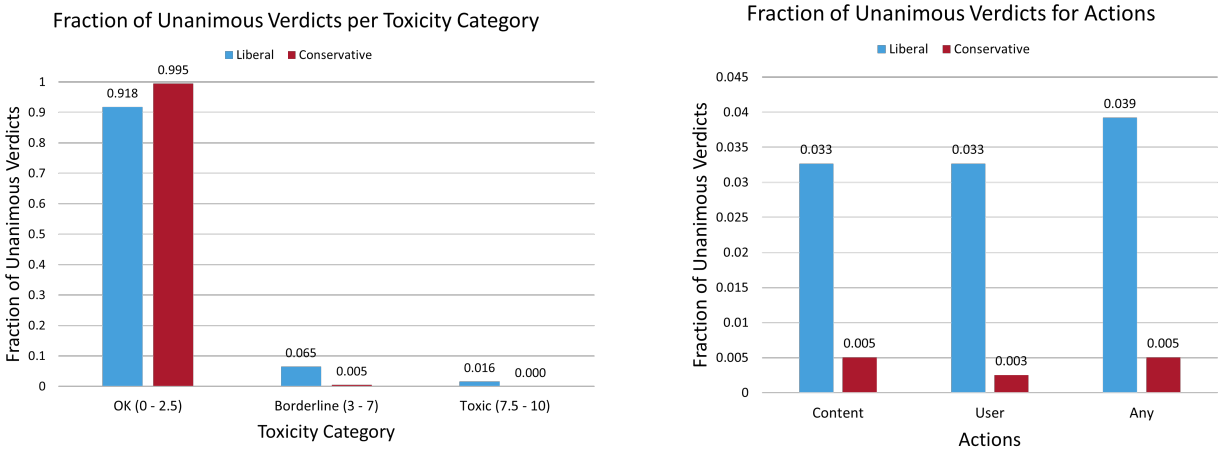

Figure 13: (Left) Fraction of case components with unanimous verdicts for liberal and conservative moderators, divided by toxicity category. (Right) Fraction of case components with unanimous verdicts for liberal and conservative moderators, divided by action type.

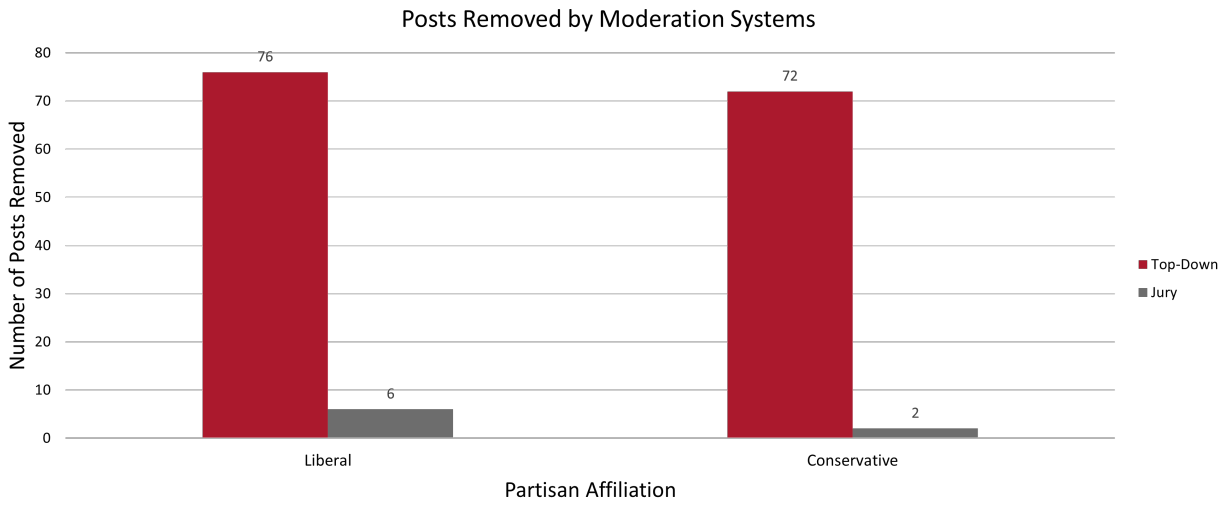

Figure 14: The number of posts removed by each moderation system. Reddit moderators removed 76 liberal and 72 conservative posts for our top-down condition, while our juror participants removed considerably fewer, with 6 liberal and 2 conservative posts removed in the jury condition.

median = Somewhat (4), conservative median = Not at all (1), $U = 36.5$, $p = 0.003$), 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), $U = 56.5$, $p = 0.075$).

4.2.2 Phase II: User Perceptions

After the moderation phase of the study was complete and content had been allocated for each user group, we looked to establish our secondary novel contribution and see whether there were any significant differences between the respective experiences of users in each group. Once the user portion had concluded, we surveyed participants about how toxic the perceived the content they saw was (Figure 16, as well as to what extent they agreed with the content they saw (Figure 17. 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

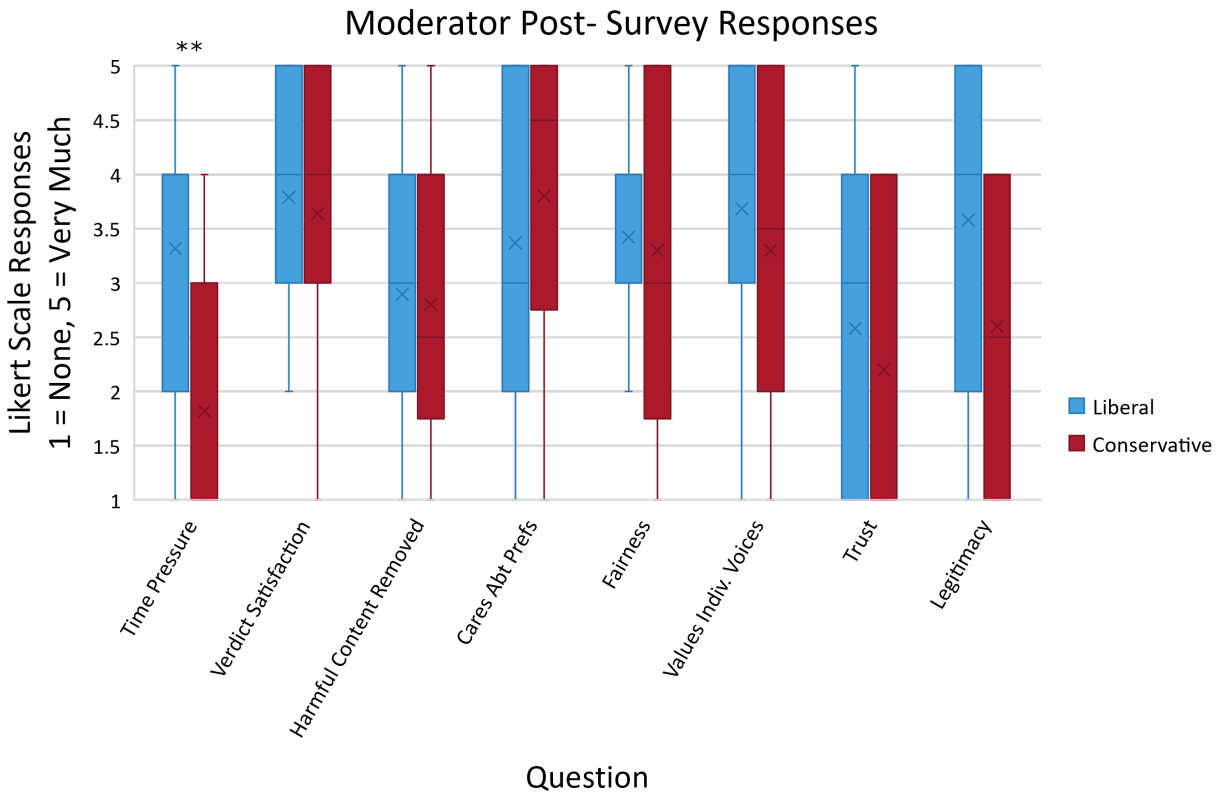

Figure 15: Moderator post-survey responses assessing the legitimacy of the moderation process, as well as satisfaction in verdict outcomes and subjective time pressure.

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.

We also categorized the subjective responses provided by nine users and observed five common themes.

Users disliked content Two users viewed the content they were shown 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 also 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.*"

None of the responses indicated that users were aware of the moderation system they were assigned.

4.3 RQ3: Drawbacks and Improvements

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

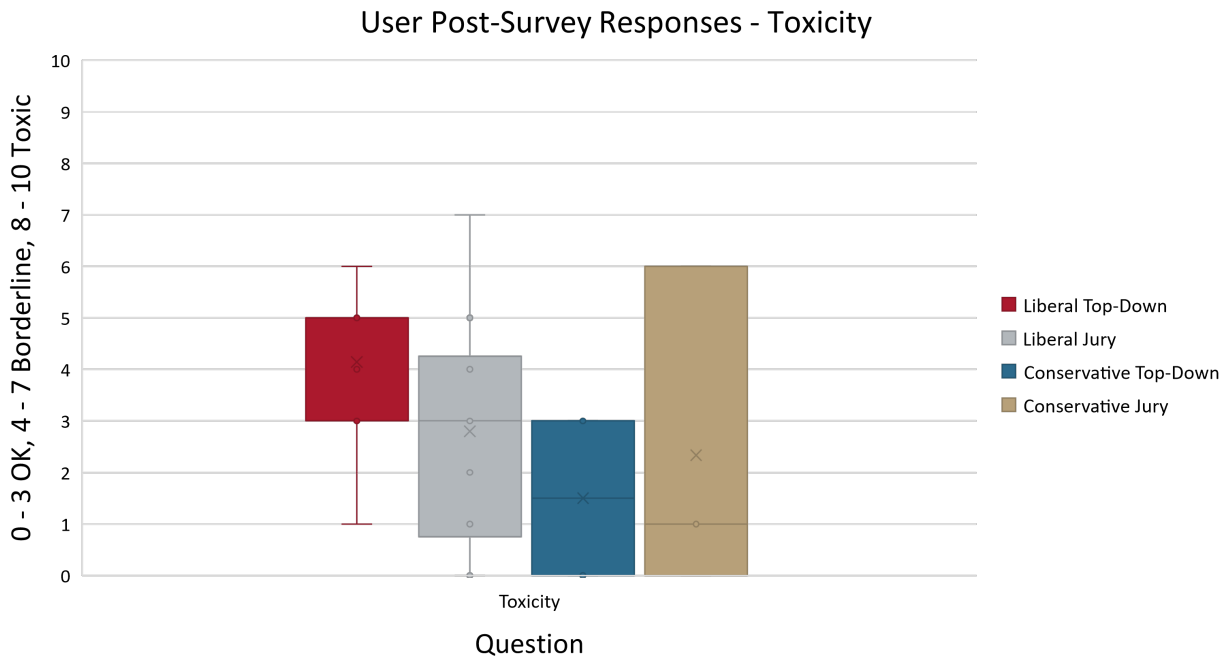

Figure 16: User post-survey responses assessing how toxic they believed the content they saw to be.

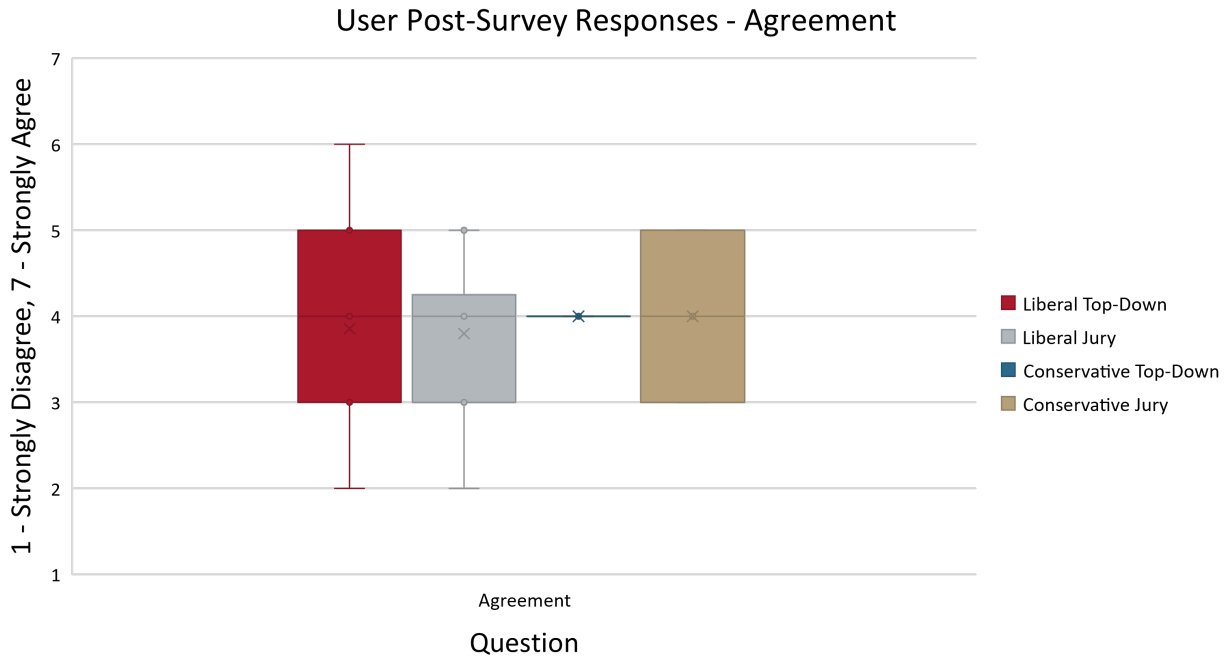

Figure 17: User post-survey responses assessing the extent users agreed with the content they saw.

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.*" Another stated that "*a clear guideline with examples of toxicity would be helpful. For example, where is the line on political statements? Also, does toxicity just need to represent an attitude, or does there have to be implicit or explicit chance of harm (physical or mental)?*"

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. They all either lean left or they all either lean right. 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 criticisms as they relate to our experiment. While our juries were ideologically homogeneous to better reflect the nature of partisan communities online, defendants accused of crimes in the United States are entitled to a trial by an impartial jury. Although the popular notion of a "jury of one's peers" may evoke the notion of a jury composed of members who share similar similar views and experiences as the defendant, in practice this has come to mean that jurors need only be "those who have enough in common with the accused, or who have enough sympathy for the accused, to be able to give a realistic evaluation of his story" [62]. The U.S. Supreme Court has ruled that for jurors to fulfill this requirement they must be drawn from a representative cross-section of the district population, though the degree to which this occurs in actuality is disputed [31, 106]. Diverse, representative juries can draw from their various life experiences and predispositions to better evaluate the evidence at a trial, and are perceived as more legitimate than juries that are unrepresentative [31].

UI confusion Six participants also noted some confusion with the moderation site interface (*"It was a little clunky at first but eventually we got the hang of it."*), and suggestions for improving it: "*I find the behavior of the bottom panel to be a bit counterintuitive. When you "lock in your decision," it minimizes into the bottom of the screen. While you can click the white bar to bring it back up and get back into the chat, there isn't any symbol that shows users they can do that; it's just a white bar. They need to experiment by clicking it to see that it opens back up...I believe users still want to stay in the conversation after voting, and taking them out of it every time they press that button isn't helpful.*"

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*": "*If it's 6 minutes per thread and you're dealing with threads with many comments, then the time pressure is too much to be able to make a meaningful group decision, or results in a hung vote.*" Two also commented on the presence of social pressure to change their vote when deliberating with other jurors: "*I definitely felt some social pressure to change my votes to be more in line with what others said.*" Expressing similar sentiments, another juror responded:"*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 about how to decide what was toxic, and maybe pressured by the other jurors to vote zero, 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... The amount of cultural and political power a moderator team could have is far beyond appropriate.*") though others didn't 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.*") These phenomena are also observed in real-world juries. Other work has shown that a variety of social factors can influence juror decisions, such as peer pressure, desire for conformity, or attitudes toward characteristics of defendants or aspects of the case, that can lead jurors to be swayed by positions of those who speak first, the views of a small minority group, or the majority [11]. Jury size also impacts the deliberation process and its outcomes: larger juries are more likely to have diverse compositions, but also tend to require more time to deliberate, and are more likely to hang [85, 88].

Cases were easy Despite the prevalence of time pressure, five jurors also commented on the relative ease of the moderation cases they encountered. One said, "*I was really thinking you were going to get way rawer with us. These were just mild political opinions. The only times I get upset at social media...is when I get upset at the content.*" Another likewise remarked "*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 also had suggestions for novel improvements to the moderation 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.*" This stands in contrast to a real-world jury, where "locking in"

the decision involves all jurors reaching a consensus and submitting the final verdict to a judge. One person suggested adding a real-time 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.*" Two indicated that time to practice with the user interface before voting began would have been a welcome edition: "*Start with a few trial rounds that don't count. Let users test out the system without feeling too much pressure before the actual experiment begins...It could be helpful to have someone who the group can ask technical questions during these trial rounds, merely to clear up any potential confusion as quickly as possible among participants about the system itself and how it's supposed to work, before the actual experiment begins.*" 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.")

Enjoyed participating Finally, six moderators said they enjoyed the experience and appreciated the opportunity to participate. ("*The study was really interesting and fun to be a part of and I hope the data has been useful."; "Overall it was interesting and I am glad to have been a part of it.*")

4.4 Limitations

There were several limitations to our study that future work could address. A key limitation that limits the statistical power and generalizability of our results is our small 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.

Additionally, several participants noted that the content we had selected for moderation

challenged them much less than they had expected. While the intention of our content selection process was to avoid showing participants posts that could be overly harmful, this may have also undermined somewhat the ecological validity of the study, as the content we omitted might have been well at home on real social media platforms. It is possible that a slightly less restrictive process for content selection might have struck a better balance between curating a realistic and meaningful set of content to be moderated while also preventing undue emotional harm to moderators.

Several of our moderators also noted that they experienced some confusion at first about how to navigate the user interface of our moderation website. While we included a video overview of the sites features and the moderation workflow to supplement text-based instructions, many participants who experienced difficulties reported that they had not watched the video, and therefore were unaware of some of the sites key features, such as selecting different components from the dropdown menu or the presence of chat functionality for communicating with other jurors. Several steps could be taken to reduce the risk for such confusion in the future, such as making the instructional video mandatory before proceeding, including practice rounds for moderators to familiarize themselves with the site interface, and creating a tutorial overlay the first time moderators interact with the moderation system.

5 Discussion

Our use of digital jury moderation did not appear to have a significant impact on the political polarization of moderators or users, though there were some effects on measures related to social polarization—social distance for conservative moderators decreased after the experiment, and the combined social distance score for liberal users in the top-down condition also decreased, indicating 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 the combined social score may relate to this as well, though pairwise comparisons of the changes in polarization between user groups from the same affiliation but different experiment conditions did not yield meaningful differences. It is difficult for us to make strong claims, due to our small sample sizes, and the fact there were not many effects mediated by the choice in experiment group. We can, however, say that digital jury moderation did not *increase* the polarization of users that interacted with its output, even if it 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. It is worth noting, however, that our study only assessed the impact of moderating content that was non-targeted and political in nature. The impacts of moderating different types of content, such as targeted online harassment, could differ substantially, but are out of scope for this work.

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 & Zhang's findings. Implementing such a system could therefore empower platform users to moderate their own communities, facilitating a "logic of care" where moderator contributions toward building and maintaining communities are valued [87]. The inter-group dialog a deliberative jury can foster may also serve as a less intrusive means of moderation that can align moderation with the directives of the UN International Covenant on Civil and Political Rights [9] and mitigate conflicts between the necessity of moderation and the profit motives of platforms.

Differences in how liberals and conservative moderators used and perceived the system were also interesting to observe. Conservatives tended to have a much narrower view of what qualified as toxic content, and reported in their subjective evaluations that they valued free speech, and were wary of the power moderators could hold to censor dissenting voices. 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. One person reported that their group had decided that the standard for any amount of toxicity was the likelihood of causing physical, bodily harm, and thus rated their case components as having zero toxicity. 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 their existing biases.

Despite this, the system is nonetheless a good jumping off point for future iterations. In their feedback in the post-experiment survey, moderators discussed several potential differences and improvements that could be implemented. We present further discussion of these in Section 5.2.

5.1 Broader Impacts

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. Introducing a peer-based moderation system to existing and future social media platforms could advance several of the National Science Foundation's societally relevant outcomes [4].

Increased Public Engagement with Technology Increased public engagement with technology (specifically, with the proposed digital jury moderation system that could be

implemented by social media platforms) is perhaps the most likely outcome of our work. Currently, users of several prominent social media platforms (such as Facebook, Twitter, or Instagram) have little direct input on community standards or content moderation, and instead report content that violates community standards for moderation by site employees or contractors [26]. However, users are the largest group of stakeholders for social media platforms: they create and consume the majority of content, and are social media platforms' most valuable asset, since their actions and engagement with site content drive lucrative advertising purchases by companies hoping to target specific demographics. In the first financial quarter of 2022, of Facebook's roughly $28 billion in revenue, approximately 96.7% was driven by advertising [7].

By introducing peer-based moderation to these platforms, users would get a "seat at the table" when making decisions about platform governance, gain a greater role in nurturing a civil and welcoming platform ecosystem, and engage in civic participation with fellow users. We believe that social media ecosystems that are diverse, inclusive, and which foster civil discussion are necessary for a healthy Internet, and perhaps social discourse at large; platforms driven solely by engagement likely require checks (e.g., via moderation) to meet these criteria [84].

Improved National Security Several politicians, scientists, and citizens have expressed alarm in recent years that disinformation online poses a threat to democracy [101]. Online disinformation may threaten democratic self-determination by undermining citizens' abilities to participate in government and democratic decision-making processes (e.g., circulating false statements misattributed to political officials); accountable representation (e.g., promoting misleading information about voting locations during an election); and democratic deliberation (e.g., disseminating false claims and promoting distrust of experts or institutions). While Facebook is largely successful at enforcing its own community standards [59], there are several occasions where disinformation, misinformation, or hateful speech have slipped through the cracks, such as in the months preceding the 2016 US presidential election, when US intelligence and law enforcement agencies announced that Russian hackers were executing a sophisticated influence campaign on social media to manipulate voters and sow distrust in the election and US political institutions [80], or more recently, false claims of voter fraud in the wake of the 2020 US presidential election made by former President Donald Trump and his supporters [78].

Sometimes content may require context-specific knowledge and information about local sociocultural norms to be moderated effectively [59]. However, on large platforms where top-down moderation occurs, such moderation decisions are often made with little or no input from platform users themselves, who may have the specialized knowledge that is pertinent to making them. It is entirely possible, therefore, that introducing a moderation system where platform users play a more central and direct role in moderating content could prevent some

of the instances where content is not removed by platform-employed moderators even when it is harmful, especially if the moderation system operates in addition to the top-down moderation infrastructure already in place (as is the case with Twitter's Birdwatch community fact-checking initiative [22], for example). As a result, disinformation or misinformative content could be caught and removed more quickly, limiting its spread and thus reducing the potential risks posed to public peace and national security.

5.1.1 Implications

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, would it still be worth implementing in its own right? If the impact of the jury system is net neutral, it may still be worthwhile to implement because of the values it upholds. As outlined by Fan & 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 a digital jury is worth considering due to the way they positively transform the relationship between social media platforms and their users. (For more discussion of values, see Section 5.1.2 below.)

Presuming that digital jury moderation is viewed as a just, legitimate, and effective choice by users, the degree to which this is true and the specific implications thereof would likely be consequences of the design decisions and trade-offs made in its implementation on platforms. Our work and Fan & Zhang's work raises several considerations:

- Which users should be jurors? Would it be a mandatory requirement for all users, or would users volunteer to be jurors? Should jurors be restricted by geography, or be members of communities with particular interests? Should jurors be anonymous? What should the size of a jury be?

- Should jurors be paid for their time, or should juries operate on a volunteer basis? Should jurors be trained, and how? Are written instructions sufficient, or are videos or interactive tutorials more beneficial?

- Should systems exist to mitigate juror bias? How might jurors be prevented from gaming the system?

- Which content should be moderated by the jury? Would all content, or only a subset of content (e.g., content other users report)? Would the jury operate in addition to other forms of moderation (algorithms, top-down), or would it operate alone?
- Should moderating cases require synchronous deliberation among jurors, or should jurors vote on actions to take asynchronously? If juries deliberate, should deliberation have a time limit? Should decisions by the jury be unanimous, or majority vote? If voting is unanimous, what should happen to content if the jury is hung?
- Should the verdicts the jury issues be requirements that social media platforms must enforce, or merely recommendations? Should jury decisions be publicly accessible?

Each of the questions and possibilities above may not have a single correct choice; 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. 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.

A key remaining question we believe will be central to the effectiveness of a digital jury is whether moderation cases should be decided via synchronous deliberation or asynchronous voting. The participants in Fan & Zhang's experiment were divided on this issue: on one hand, users believed that exposure to a diverse array of perspectives when moderating content was valuable; on the other, some jurors were distrustful of the opinions of other participants, and believed that deliberating could lead to "groupthink," where the decisions and perspectives of a majority are upheld at the expense of minorities. Work by Hu and colleagues [47] has shown that decisions of online juries that deliberate are more consistent (the same jurors would make similar decisions on similar cases), but that these decisions arose in part due to hardening of majority opinions. Therefore, another key related question is whether verdicts reached by the jury should require unanimity. Prior work has shown that jurors that are required to reach decisions unanimously are more satisfied with case outcomes than those that required a simple majority [71], and unanimity serves to limit the influence of a majority opinion (since compromise with any minorities is required to reach a consensus).

However, this comes at the cost of longer turnaround times for moderation decisions, since juries that fail to reach a consensus would be "hung," and the case would need to be retried by different jurors. The juries in our study are required to reach a unanimous verdict in order for their decisions to stand; it remains to be seen how jurors felt about this, or whether a large proportion of cases resulted in a hung jury where a consensus was not reached. In practice, it may be the case that requiring unanimous verdicts is not feasible, in which case a different decision threshold (such as a 75% super-majority) may be an effective compromise. Future work should explore the different possible configurations of a jury moderation system to determine which arrangements are most effective.

While a jury system comparable to that proposed by Fan & Zhang and our own work has not yet been implemented on a major platform (to our knowledge), a start-up social media company currently seeking venture capital funding, Podium [79], purports to implement such a system, along with a way to ensure unbiased juries by mapping each user's individual bias, and staking their site reputation against their actions on the platform. While the site is not operational yet, we think it represents a promising potential implementation that could answer some of the outstanding questions that have been posed.

Twitter, similarly, is slowly deploying its Birdwatch system [22], which implements a community-based approached to labeling misinformative content on Twitter, allowing users to provide context for tweets and reach a consensus as a community on which context is helpful. It is currently available for community members to contribute to, and (as of March 2022) is in the process of being deployed on Twitter in the US. So far the results of the rollout have been promising, with users largely finding the user-added context notes helpful, and less likely to agree with misinformative content that had these notes [23]. Such support for a community-driven approach to tackling harmful content online by a major social media platform represents an encouraging step forward.

5.1.2 Ethically Aligned Design

IEEE's Ethically Aligned Design (EAD) [3] presents a series of principles, questions, and guidelines outlining ethical considerations and promoting human well-being in a society where autonomous and intelligent systems (AI/S) have become, and will continue to be, pervasive. While digital jury moderation does not strictly fall under the AI/S umbrella, the principles in EAD are a useful lens through which we can examine any technology we might develop.

We maintain that any real-world implementation of a digital jury moderation system should uphold the principles presented in EAD. A digital jury moderation system, first and foremost, should be a peer-based system that promotes and protects the human rights of social media users. A digital jury system could empower users and enhance their self-governance, agency, and civil rights on these platforms, enabling them to more easily deter-

mine for themselves which content is acceptable, ensure their voices are heard, and exert their own control over the platform ecosystem, instead of merely being at the mercy of decisions made by the platforms themselves. 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 likely be necessary.

A transparent decision-making process is essential for awareness of misuse—if jurors deliberately attempt to sabotage cases by being purposefully combative or non-cooperative, the verdicts and any transcripts for these cases should be available so such malicious actors can be held to account. A system of checks and balances, such as a process for appeals, or automated or manual detection of juror bias, would also be a valuable safeguard. 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.

5.2 Future Work

Our implementation of a digital jury was only one potential implementation, and there are several potential configurations that could be created in the future. As indicated in Section 5.1.1 above, and as noted by our participants, there are several aspects of the system that could be changed and investigated in future work. 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.*"

This also ties into another point raised by participants, which 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. It is possible jurors might have chosen to remove more content if, for example, we had removed the unanimity requirement (allowing jurors to remove content if consensus with a majority, rather than the whole deliberating body, was reached); if the nature of the content were different (e.g., bullying another user instead of non-targeted political content); or if users could view the entire selection of cases before making decisions, allowing them to rate toxicity in the context of all borderline content rather than on a per-case basis.

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). 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.

5.2.1 Evaluation

Once a digital jury moderation system is implemented in a large-scale social media environment (whether it be an existing platform such as Facebook, or one yet to exist, such as Podium described above), assessing the degrees to which users engage with other users and the moderation structure; how positive these interactions are, and how diverse the overall user base and specific sub-communities are; user mental health is impacted; and the amount of harmful content has changed, would be difficult and time-consuming, and would most likely need to be an ongoing effort as the platform evolves and grows (or shrinks) over time. Confounding factors (the effects of specific content, which could be seen by different users in different geographic areas, changes in user interface or site design, and perhaps many others), would be difficult, if not impossible, to eliminate. The most rigorous way to test for specific impacts would likely be to have identical content streams passed through two pipelines on the same website (one implementing a digital jury for content moderation, the other either a top-down or some other approach) for an extended duration, perhaps for several years—similar to our study. However, this is not a realistic approach, as it would require a large social media platform to divide its user base and manage two distinct ecosystems simultaneously for the sake of what would amount to a long-running experiment, requiring potentially much

more personnel and resources (to say nothing about the ethical considerations of doing so in the wild).

A more realistic approach would likely be to employ either covert or overt measures of user activity, or a combination of the two, and compare users and content on either a single platform before and after the implementation of a digital jury moderation system, or a platform implementing a digital jury versus another implementing a different method of content moderation. The former would introduce fewer confounds, since the platform user base, interaction paradigms, content type, and style would be equivalent, but the latter could also be compelling in the absence of the former. Covert measures could be performed to assess user engagement (e.g., tracking the number of users who are jurors, the average number of posts/comments/other actions made by users), the sentiment of interactions among and the diversity of the user base (sentiment analysis of posts/comments, user demographics), and change in the amount of harmful content (likely via a third-party fact checking analysis), while overt measures asking users about their mental health and their perceptions of platform toxicity and moderation practices as in Cook et al. [26] could assess user mental health and change in the amount of harmful content, respectively.

6 Conclusion

In this work, we investigated whether democratizing content moderation on social media platforms would impact the polarization of social media 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 & 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 peer-based digital jury moderation system for content moderation on social media platforms would have the benefit of providing users agency in platform governance without adversely impacting user experience. While our study had several limitations (such as a small sample size and content that was seen as easy to moderate), there are several potential improvements and alterations to our implementation of a digital jury that would be promising 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.

References

[1] 2017. *The partisan divide on political values grows even wider.* Report. Pew Research Center. https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2017/10/10-05-2017-Political-landscape-release-updt..pdf

[2] 2019. Combatting Misinformation on Instagram. https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

[3] 2019. Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (2019), 1–294. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[4] 2020. Broader Impacts Review Document for National Science Foundation Proposals. https://researchinsociety.org/wp-content/uploads/2021/02/GuidingPrinciplesDoc2020.pdf

[5] 2021. How Facebook's third-party fact-checking program works. https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works

[6] 2021. Reddiquette. https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette

[7] 2022. *Meta Earnings Presentation Q1 2022.* Report. Meta. https://s21.q4cdn.com/399680738/files/doc_financials/2022/q1/Q1-2022_Earnings-Presentation_Final.pdf

[8] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review* 110, 3 (2020), 629–76. https://doi.org/10.1257/aer.20190658

[9] Evelyn Mary Aswad. 2018. The future of freedom of expression online. *Duke L. & Tech. Rev.* 17 (2018), 26. https://scholarship.law.duke.edu/dltr/vol17/iss1/2/

[10] Brooke Auxier and Monica Anderson. 2021. *Social Media Use in 2021.* Report. Pew Research Center. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/04/PI_2021.04.07_Social-Media-Use_FINAL.pdf

[11] Michelle Baddeley and Sophia Parkinson. 2012. Group decision-making: An economic analysis of social influence and individual difference in experimental juries. *The Journal of Socio-Economics* 41, 5 (2012), 558–573. https://doi.org/10.1016/j.socec.2012.04.023

[12] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander

Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221. https://doi.org/10.1073/pnas.1804840115

[13] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542. https://doi.org/10.1177/0956797615594620

[14] Pablo Barberá. 2020. *Social media, echo chambers, and political polarization.* Book section 3, 34–55.

[15] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 830–839. https://ojs.aaai.org/index.php/ICWSM/article/view/7347

[16] Tetyana Bohdanova. 2014. Unexpected Revolution: The Role of Social Media in Ukraine's Euromaidan Uprising. *European View* 13, 1 (2014), 133–142. https://doi.org/10.1007/s12290-014-0296-4

[17] Adam Bonica, Nolan McCarty, Keith T Poole, and Howard Rosenthal. 2013. Why hasn't democracy slowed rising inequality? *Journal of Economic Perspectives* 27, 3 (2013), 103–24. https://doi.org/10.1257/jep.27.3.103

[18] Samantha Bradshaw and Philip N Howard. 2018. The global organization of social media disinformation campaigns. *Journal of International Affairs* 71, 1.5 (2018), 23–32. https://www.jstor.org/stable/26508115

[19] Julia Cambre, Scott R. Klemmer, and Chinmay Kulkarni. 2017. Escaping the Echo Chamber: Ideologically and Geographically Diverse Discussions about Politics. , 2423–2428 pages. https://doi.org/10.1145/3027063.3053265

[20] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), Article 174. https://doi.org/10.1145/3359276

[21] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Trans. Comput.-Hum. Interact.* 29, 4 (2022), Article 29. https://doi.org/10.1145/3490499

[22] Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

[23] Keith Coleman. 2022. Building a better Birdwatch. https://blog.twitter.com/en_us/topics/company/2022/building-a-better-birdwatch

[24] MacKenzie F. Common. 2020. Fear the Reaper: how content moderation rules are enforced on social media. *International Review of Law, Computers & Technology* 34, 2 (2020), 126–152. https://doi.org/10.1080/13600869.2020.1733762 doi: 10.1080/13600869.2020.1733762.

[25] Meredith Conroy, Jessica T Feezell, and Mario Guerrero. 2012. Facebook and political engagement: A study of online political group membership and offline political engagement. *Computers in Human behavior* 28, 5 (2012), 1535–1546. https://doi.org/10.1016/j.chb.2012.03.012

[26] Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. *Frontiers in Human Dynamics* 3, 3 (2021). https://doi.org/10.3389/fhumd.2021.626409

[27] Larry Diamond. 2010. Liberation technology. *Journal of Democracy* 21, 3 (2010), 69–83. https://www.journalofdemocracy.org/articles/liberation-technology/

[28] Thiago Dias Oliva. 2020. Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review* 20, 4 (2020), 607–640. https://doi.org/10.1093/hrlr/ngaa032

[29] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173785

[30] James N Druckman, Erik Peterson, and Rune Slothuus. 2013. How elite partisan polarization affects public opinion formation. *American Political Science Review* 107, 1 (2013), 57–79. https://doi.org/10.1017/S0003055412000500

[31] Leslie Ellis and Shari Siedman Diamond. 2003. Race, diversity, and jury composition: Battering and bolstering legitimacy. *Chi.-Kent L. Rev.* 78 (2003), 1033. https://scholarship.kentlaw.iit.edu/cklawreview/vol78/iss3/6/

[32] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376293

[33] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday* 25, 11 (2020). https://doi.org/10.5210/fm.v25i11.11431

[34] Casey Fiesler, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*. https://ojs.aaai.org/index.php/ICWSM/article/view/15033

[35] Kathy Frankovic. 2016. Belief in conspiracies largely depends on political identity. *YouGov, December* 27 (2016), 17–20. https://today.yougov.com/topics/politics/articles-reports/2016/12/27/belief-conspiracies-largely-depends-political-iden

[36] Ryan J Gallagher, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. 2018. Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PloS one* 13, 4 (2018), e0195644. https://doi.org/10.1371/journal.pone.0195644

[37] Kiran Garimella. 2018. Polarization on social media. (2018). https://users.ics.aalto.fi/kiran/content/thesis.pdf

[38] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611

[39] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977), 67–82.

[40] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

[41] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3502004

[42] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* (2020). https://doi.org/10.1177/2053951719897945

[43] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380. https://www.jstor.org/stable/2776392

[44] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. (2021). https://doi.org/10.1145/3479610

[45] Danula Hettiachchi and Jorge Goncalves. 2019. Towards Effective Crowd-Powered Online Content Moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Fremantle, WA, Australia) *(OZCHI'19)*. Association for Computing Machinery, New York, NY, USA, 342–346. https://doi.org/10.1145/3369457.3369491

[46] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. 2011. Opening closed regimes: what was the role of social media during the Arab Spring? *Available at SSRN 2595096* (2011). https://doi.org/10.2139/ssrn.2595096

[47] Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. 2012. Breaking News on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2751–2754. https://doi.org/10.1145/2207676.2208672

[48] Xinlan Emily Hu, Mark E Whiting, and Michael S Bernstein. 2021. *Can Online Juries Make Consistent, Repeatable Decisions?* Association for Computing Machinery, Article 142. https://doi.org/10.1145/3411764.3445433

[49] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 74 (nov 2018), 24 pages. https://doi.org/10.1145/3274343

[50] Shanto Iyengar and Kyu S Hahn. 2009. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication* 59, 1 (2009), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

[51] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), Article 192. https://doi.org/10.1145/3359294

[52] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (2019), Article 31. https://doi.org/10.1145/3338243

[53] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), Article 150. https://doi.org/10.1145/3359252

[54] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE* 16, 8 (2021), e0256762. https://doi.org/10.1371/journal.pone.0256762

[55] Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 278–289. https://ojs.aaai.org/index.php/ICWSM/article/view/3229

[56] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proc. ACM Hum.-Comput. Interact.* 4, GROUP (2020), Article 17. https://doi.org/10.1145/3375197

[57] Scott Keeter and Ruth Igielnik. 2016. *Can likely voter models be improved?* Report. Pew Research Center. https://www.pewresearch.org/methods/2016/01/07/can-likely-voter-models-be-improved/

[58] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the Other Way:Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information. *Communication Research* 36, 3 (2009), 426–448. https://doi.org/10.1177/0093650209333030

[59] Jason Koebler and Joseph Cox. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. *Vice Motherboard* (2018). https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

[60] Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S. Bernstein, and Scott R. Klemmer. 2015. Talkabout: Making distance matter with small groups in massive classes. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1116–1128. https://doi.org/10.1145/2675133.2675166

[61] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28, 2 (2012), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

[62] Lewis H LaRue. 1976. A Jury of One's Peers. *Wash. & Lee L. Rev.* 33 (1976), 841. https://scholarlycommons.law.wlu.edu/wlulr/vol33/iss4/3

[63] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174160

[64] Megan Nicole Mancini. 2019. *Development and Validation of the Secondary Traumatic Stress Scale in a Sample of Social Media Users*. Thesis. https://engagedscholarship.csuohio.edu/etdarchive/1132/

[65] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute* (2017). https://www.datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

[66] J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. , Paper 9 pages. https://doi.org/10.1145/3173574.3173583

[67] Maria Milosh, Marcus Painter, David Van Dijcke, and Austin L Wright. 2020. Unmasking Partisanship: How Polarization Influences Public Responses to Collective Risk. *University of Chicago, Becker Friedman Institute for Economics Working Paper* 2020-102 (2020). https://bfi.uchicago.edu/wp-content/uploads/BFI_WP_2020102.pdf

[68] Patricia Moravec, Randall Minas, and Alan R Dennis. 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* 18-87 (2018). https://doi.org/10.2139/ssrn.3269541

[69] Laura Murphy and Megan Cacace. 2020. *Facebook's Civil Rights Audit – Final Report*. Report. https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

[70] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383. https://doi.org/10.1177/1461444818773059

[71] Charlan Nemeth. 1977. Interactions between jurors as a function of majority vs. unanimity decision rules. *Journal of Applied Social Psychology* 7, 1 (1977), 38–56. https://doi.org/doi/10.1111/j.1559-1816.1977.tb02416.x

[72] Xia-mu Niu and Yu-hua Jiao. 2008. An overview of perceptual hashing. *ACTA ELECTONICA SINICA* 36, 7 (2008), 1405. https://www.ejournal.org.cn/EN/Y2008/V36/I7/1405

[73] Shannon M. Oltmann, Troy B. Cooper, and Nicholas Proferes. 2020. How Twitter's affordances empower dissent and information dissemination: An exploratory study of the rogue and alt government agency Twitter accounts. *Government Information Quarterly* 37, 3 (2020), 101475. https://doi.org/10.1016/j.giq.2020.101475

[74] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31. https://doi.org/10.1145/3512929

[75] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science* 66, 11 (2020), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

[76] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* 147, 12 (2018), 1865. https://doi.org/doi/10.1037/xge0000465

[77] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526. https://doi.org/10.1073/pnas.1806781116

[78] Gordon Pennycook and David G Rand. 2021. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review* (2021).

[79] Podium. 2019. Introducing Podium. https://medium.com/@PodiumNetwork/introducing-podium-796111865f18

[80] Dana Priest, Ellen Nakashima, and Tom Hamburger. 2016. U.S. investigating potential covert Russian plan to disrupt November elections. https://www.washingtonpost.com/world/national-security/intelligence-community-investigating-covert-russian-influence-operations-in-the-united-states/2016/09/04/aec27fa0-7156-11e6-8533-6b0b0ded0253_story.html

[81] Steve Rathje, Jay J Van Bavel, and Sander van der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118, 26 (2021). https://doi.org/10.1073/pnas.2024292118

[82] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (2021), Article 316. https://doi.org/10.1145/3476057

[83] Joshua Robison and Kevin J Mullinix. 2016. Elite polarization and public opinion: How polarization is communicated and its effects. *Political Communication* 33, 2 (2016), 261–282. https://doi.org/10.1080/10584609.2015.1055526

[84] Kaleigh Rogers. 2021. Facebook's Algorithm Is Broken. We Collected Some Suggestions On How To Fix It. https://fivethirtyeight.com/features/facebooks-algorithm-is-broken-we-collected-some-spicy-suggestions-on-how-to-fix-it/

[85] Robert T Roper. 1980. Jury Size and Verdict Consistency:" A Line Has to Be Drawn Somewhere"? *Law and Society Review* (1980), 977–995. https://www.jstor.org/stable/3053217

[86] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information

[87] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society* 22, 6 (2020), 1026–1042. https://doi.org/10.1177/1461444819875990

[88] Michael J Saks and Mollie Weighner Marti. 1997. A meta-analysis of the effects of jury size. *Law and Human Behavior* 21, 5 (1997), 451–467. https://doi.org/10.1023/A:1024819605652

[89] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020), Article 107. https://doi.org/10.1145/3415178

[90] Christina Shane-Simpson, Adriana Manago, Naomi Gaggi, and Kristen Gillespie-Lynch. 2018. Why do college students prefer Facebook, Twitter, or Instagram? Site affordances, tensions between privacy and self-expression, and implications for social capital. *Computers in Human Behavior* 86 (2018), 276–288. https://doi.org/10.1016/j.chb.2018.04.041

[91] Grace Shao. 2019. Social media has become a battleground in Hong Kong's protests. https://www.cnbc.com/2019/08/16/social-media-has-become-a-battleground-in-hong-kongs-protests.html

[92] Qinlan Shen, Michael Miller Yoder, Yohan Jo, and Carolyn P Rosé. 2019. Perceptions of censorship and moderation bias in political debate forums. In *Twelfth International AAAI Conference on Web and Social Media*. https://ojs.aaai.org/index.php/ICWSM/article/view/15002

[93] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

[94] Elizabeth N. Simas, Scott Clifford, and Justin H. Kirkland. 2020. How Empathic Concern Fuels Political Polarization. *American Political Science Review* 114, 1 (2020), 258–269. https://doi.org/10.1017/S0003055419000534

[95] Stuart Soroka, Patrick Fournier, and Lilach Nir. 2019. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences* 116, 38 (2019), 18888–18892. https://doi.org/10.1073/pnas.1908369116

[96] Melissa Spinner. 2012. *The Effects of Social Media on Democratization*. Thesis. https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1109&context=cc_etds_theses

[97] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160. https://doi.org/10.1177/0266382117722446

[98] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927. https://doi.org/10.1177/1461444819834317

[99] Alexander J. Stewart, Mohsen Mosleh, Marina Diakonova, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. 2019. Information gerrymandering and undemocratic decisions. *Nature* 573, 7772 (2019), 117–121. https://doi.org/10.1038/s41586-019-1507-6

[100] Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. 2018. The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. *The International Journal of Press/Politics* 23, 1 (2018), 95–115. https://doi.org/10.1177/1940161217740697

[101] Chris Tenove. 2020. Protecting democracy from disinformation: Normative threats and policy responses. *The International Journal of Press/Politics* 25, 3 (2020), 517–537. https://doi.org/10.1177/1940161220918740

[102] Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Social media, political polarization, and political disinformation: A review of the scientific literature* (2018). https://hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf

[103] Aleksandra Urman. 2020. Context matters: political polarization on Twitter from a comparative perspective. *Media, Culture & Society* 42, 6 (2020), 857–879. https://doi.org/10.1177/0163443719876541

[104] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 417–426. https://doi.org/10.1145/2702123.2702191

[105] Andreas Veglis. 2014. Moderation techniques for social media content. In *International conference on social computing and social media*. Springer, 137–148. https://doi.org/10.1007/978-3-319-07632-4_13

[106] Robert C Walters, Michael D Marin, and Mark Curriden. 2005. Jury of Our Peers: An Unfulfilled Constitutional Promise. *SMUL Rev.* 58 (2005), 319. https://scholar.smu.edu/smulr/vol58/iss2/5/

[107] Galen Weld, Amy X. Zhang, and Tim Althoff. 2021. What Makes Online Communities' Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. *arXiv preprint arXiv:2111.05835* (2021). https://doi.org/10.48550/arXiv.2111.05835

[108] Anne E. Wilson, Victoria A. Parker, and Matthew Feinberg. 2020. Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences* 34 (2020), 223–228. https://doi.org/10.1016/j.cobeha.2020.07.005

[109] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378. https://doi.org/10.1145/3379337.3415858

[110] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators

in News Articles. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, 603–612. https://doi.org/10.1145/3184558.3188731

A Survey Instruments

Our study used the following survey instruments:

A.1 Pre-Experiment Screening Questionnaire

1. Are you 18 years of age or older?
2. Are you a permanent or temporary resident of the USA?
3. What is your time zone?
4. Do you visit a social media site at least three times a week in order to read messages/posts?

Question to assess engagement with politics, adapted from Keeter & Igielnik [57]:

5. How engaged do you consider yourself with US politics? (Very Disinterested / Disinterested / Neutral / Interested / Very Interested)

Question 6 is the variant of the ideological consistency scale adopted by Bail et al. [12], used to assess ideological polarization. Half of these statements are worded in a manner that is designed to appeal to liberals (5b, 5d, 5g, 5i, 5j), and the other half are intended to appeal to conservatives (5a, 5c, 5e, 5f, 5h). The order of the statements is randomized to account for ordering effects. We use this in our screening questionnaire to group people by their partisan opinions (liberal or conservative), to facilitate assigning participants for the moderator portion to jury groups of the same political ideology, so the experiment can begin as soon as possible.

6. Please rate the degree to which you agree or disagree with the following statements from 1 (Strongly disagree) to 7 (Strongly agree):
   (a) Stricter environmental laws and regulations cost too many jobs and hurt the economy.
   (b) Government regulation of business is necessary to protect the public interest.
   (c) Poor people today have it easy because they can get government benefits without doing anything in return.
   (d) Immigrants today strengthen our country because of their hard work and talents.
   (e) Government is almost always wasteful and inefficient.
   (f) The best way to ensure peace is through military strength.
   (g) Racial discrimination is the main reason why many black people can't get ahead these days.
   (h) The government today can't afford to do much more to help the needy.
   (i) Business corporations make too much profit.
   (j) Homosexuality should be accepted by society.

7. What made you decide to participate in this study?

8. 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?

For the moderation portion of the study (Phase I), we also asked participants to indicate their availability for jury deliberation time slots for the two weeks after the date of the screening questionnaire, to facilitate the formation of juries with overlapping availability:

9. The moderators for our experiment will meet in groups online for one-hour periods on two different days.

   Please indicate whether you will be available for one hour at any of the dates/times below. All times are in Eastern Standard Time (UTC-05:00).

   Please be as generous as possible!

A.2 Pre-Experiment Survey

1. Tell us about yourself:

   (a) Age Range: (18-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65+)

   (b) Gender: (Female / Male / Non-binary / Other)

   (c) Highest completed level of education: (Some High School / High School / College / Graduate/Professional)

2. What do you consider your political affiliation? (Republican / Democrat / Independent / Libertarian / Other / Not Sure)

3. Would you call yourself a strong [political affiliation] or a not very strong [political affiliation]?

4. Would you say you follow what's going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all? (Most of the time / Some of the time / Only now and then / Hardly at all)

Question 5 is the variant of the ideological consistency scale adopted by Bail et al. [12], used to assess ideological polarization. Half of these statements are worded in a manner that is designed to appeal to liberals (5b, 5d, 5g, 5i, 5j), and the other half are intended to appeal to conservatives (5a, 5c, 5e, 5f, 5h). The order of the statements is randomized to account for ordering effects. We ask this question again in the pre-experiment survey to account for any changes that may have occurred since the screening questionnaire, as some time may have passed.

5. Please rate the degree to which you agree or disagree with the following statements from 1 (Strongly disagree) to 7 (Strongly agree):

(a) Stricter environmental laws and regulations cost too many jobs and hurt the economy.

(b) Government regulation of business is necessary to protect the public interest.

(c) Poor people today have it easy because they can get government benefits without doing anything in return.

(d) Immigrants today strengthen our country because of their hard work and talents.

(e) Government is almost always wasteful and inefficient.

(f) The best way to ensure peace is through military strength.

(g) Racial discrimination is the main reason why many black people can't get ahead these days.

(h) The government today can't afford to do much more to help the needy.

(i) Business corporations make too much profit.

(j) Homosexuality should be accepted by society.

Question 6 is the feeling thermometer Suhay et al. [100] used to assess affective polarization:

6. Please rate each group below using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group or that you don't care too much for that group. You would rate the group at the 50-degree mark if you don't feel particularly warm or cold toward the group. You may choose any number between (and including) 0 and 100. Use the slider to indicate your rating.

   (a) Democrats

   (b) Republicans

Questions 7 and 8 from Suhay et al. [100] assess marriage preference, which we consider as a component of social polarization:

7. How do you think you would react if a member of your family told you they were going to marry a Republican? Would you be... (1 Very unhappy / Unhappy / Somewhat unhappy / Neither happy nor unhappy / Somewhat happy / Happy / 7 Very happy)

8. How do you think you would react if a member of your family told you they were going to marry a Democrat? Would you be... (1 Very unhappy / Unhappy / Somewhat unhappy / Neither happy nor unhappy / Somewhat happy / Happy / 7 Very happy)

Questions 9 and 10 from Simas et al. [94] assess social distance, which we consider as a component of social polarization:

9. How upset would you feel if your neighbor placed a "Joe Biden for President" sign in their yard? (1 Not upset at all / Not too upset / Somewhat upset / Very upset / 5 Extremely upset)

10. How upset would you feel if your neighbor placed a "Donald Trump for President" sign in their yard? (1 Not upset at all / Not too upset / Somewhat upset / Very upset / 5 Extremely upset)

Question 11 from Suhay et al. [100] assesses likeminded community preference, which we consider as a component of social polarization:

11. Imagine for a moment that you are moving to another community... In deciding where to live, how important would it be to live in a place where most people held political views similar to your own? (1 Not important, Somewhat important, Moderately important, 4 Very important)

A.3 Post-Experiment Survey

A.3.1 Post-Experiment Polarization Survey

Identical to questions 2 – 11 of the pre-experiment survey.

A.3.2 Post-Experiment User Experience Survey

For moderators:

1. Did you feel any time constraints doing this process? (Not at all / A little / Moderately / Somewhat / Very much)
2. How satisfied were you on average with the jury's decisions? (Not at all / A little / Moderately / Somewhat / Very much)

Please reflect upon the process for reaching the jury's outcomes.

3. This process is an effective way to protect users from unwanted or toxic content. (Not at all / A little / Moderately / Somewhat / Very much)
4. This process cares about my preferences. (Not at all / A little / Moderately / Somewhat / Very much)
5. This process is fair. (Not at all / A little / Moderately / Somewhat / Very much)
6. This process values individual voices equally. (Not at all / A little / Moderately / Somewhat / Very much)
7. This process improves my trust in content moderation decisions. (Not at all / A little / Moderately / Somewhat / Very much)
8. This process feels like a legitimate exercise of the social media platform's power. (Not at all / A little / Moderately / Somewhat / Very much)
9. Are there ways you think the moderation system could be improved?
10. Other thoughts/comments?

For users:

1. On average, how toxic was the content you read during the course of the experiment? Toxicity is defined as the likelihood to cause harm. (0 – 3 OK, 4 – 7 Borderline, 8 – 10 Toxic)

2. On average, to what extent did you agree or disagree with the content you read during the course of the experiment? (1 (Strongly disagree) to 7 (Strongly agree))

3. Other thoughts/comments?

A.4 Secondary Traumatic Stress Scale for Social Media Users (STSS-SM)

The STSS-SM, adopted from the STSS by Mancini [64]. The test was scored by coding Never as 0 and Very Often as 5, then adding the responses for all items.

The following is a list of statements made by persons who have been impacted by their experiences on social media with traumatized individuals or traumatic experiences. Read each statement, then indicate how frequently the statement was true for you in the past seven days (Never/Rarely/Occasionally/Often/Very Often):

1. I felt emotionally numb after using social media.

2. My heart started pounding when I thought about things I've seen on social media.

3. It seemed as if I was reliving the trauma(s) experienced by people I've seen on social media.

4. I had trouble sleeping.

5. I felt discouraged about the future.

6. Reminders of things I've seen on social media upset me.

7. I had little interest in being around others.

8. I felt jumpy.

9. I was less active than usual.

10. I thought about upsetting things I've seen on social media when I didn't intend to.

11. I had trouble concentrating.

12. I avoided people, places, or things that reminded me of upsetting things I've seen on social media.

13. I had disturbing dreams about things I've seen on social media.

14. I wanted to avoid social media.
15. I was easily annoyed.
16. I expected something bad to happen.
17. I noticed gaps in my memory about social media.

B Outcome Measures

From our survey responses, we computed several outcome measures to assess ideological, affective, and social polarization. All calculations refer to question numbers from the pre-experiment survey (see A). These outcome measures (and the components of the social polarization score) are all scaled to ± 1 so they are comparable. For each, -1 is the most "liberal" extremum and $+1$ is the most "conservative" extremum.

B.1 Ideological Score

We use Question 5 to create an ideological score, which we used to measure ideological polarization. Half of its items are worded in a manner that is designed to appeal to liberals (5b, 5d, 5g, 5i, 5j), and the other half are intended to appeal to conservatives (5a, 5c, 5e, 5f, 5h). The order of the statements is randomized to account for ordering effects. Responses for each item could range from Strongly Disagree (coded as 1) to Strongly Agree (coded as 7). The procedure for calculating the ideological score is as follows:

1. Conservative-aligned questions are centered about 0 and normalized to ± 1.
2. Liberal-aligned questions are reverse-coded, centered about 0, and normalized to ± 1.
3. The ideological score is the mean of all normalized scores from each item.

$$\text{Ideological\_Score} = \frac{1}{10} \times \left(\frac{\sum_{i=1}^{5} (\text{Conservative\_Ques}_i - 4)}{3} + \frac{\sum_{i=1}^{5} (4 - \text{Liberal\_Ques}_i)}{3} \right) \tag{1}$$

The end result is a score between ± 1, with -1 as the most liberal and $+1$ as the most conservative. A score of 0 indicates no ideological leaning toward either end of the spectrum.

B.2 Affective Score

We use Question 6 to create an affective score for measuring affective polarization. The feeling thermometers used for Question 6 range from 0 (least favorable) to 100 (most favorable). Question 6a assessed feelings toward Democrats, and Question 6b assessed feelings toward Republicans. We define the affective score as:

$$\text{Affective\_Score} = \frac{\text{Q6b} - \text{Q6a}}{100} \tag{2}$$

The end result is a score between ± 1, with -1 is the most favorable to liberals and $+1$is the most favorable to conservatives.

B.3 Social Score

We use Questions 7 – 11 to develop a score to assess social polarization. The score is composed of three components:

- **Marriage Preference**: Calculated from Questions 7 and 8. (How do you think you would react if a member of your family told you they were going to marry a [Republican/Democrat]?) Responses could range from Very Unhappy (coded as 1) to Very Happy (coded as 7).

$$\text{Marriage\_Pref} = \frac{\text{Q7} - \text{Q8}}{6} \tag{3}$$

  This yields a measure of marriage preference from ± 1, with $+1$ indicating a strong preference for Republicans and -1 indicating a strong preference for Democrats.

- **Social Distance**: Calculated from Questions 9 and 10. (How upset would you feel if your neighbor placed a "[Joe Biden/Donald Trump] for President" sign in their yard?) Responses could range from Not Upset At All (coded as 1) to Extremely Upset (coded as 5).

$$\text{Social\_Distance} = \frac{\text{Q9} - \text{Q10}}{4} \tag{4}$$

  This yields a measure of social distance from ± 1, with $+1$ indicating a strong desire for distance from Democrats (favorable to conservatives) and -1 indicating a strong desire for distance from Republicans (favorable to liberals).

- **Likeminded Community**: Calculated from Question 11. (In deciding where to live, how important would it be to live in a place where most people held political views similar to your own?) Responses could range from Not Important (coded as 1) to Very Important (coded as 4).

$$\text{Likeminded\_Community} = \frac{\text{Q11} - 1}{3} \tag{5}$$

  This yields a measure of preference for a likeminded community ranging from 0 to $+1$, with 0 indicating no preference $+1$ indicating a strong preference.

We then combine these components into a **combined social score** for social polarization:

$$\begin{aligned}\text{Combined\_Social\_Score} = \text{mean}(&\text{Marriage\_Preference}, \text{Social\_Distance}, \\ &\text{sign}(\text{Ideological\_Score}) \times \text{Likeminded\_Community})\end{aligned} \tag{6}$$

This yields a measure of social polarization from ± 1, with $+1$ indicating a strong conservative preference and -1 indicating a strong liberal preference. Note that the sign of the Ideological_Score is used for the Likeminded_Community component, since the measure for Likeminded_Community refers to the political views of the participant.

Dep. Variable:	Q5a	**Log-Likelihood:**	-25.373
Model:	OrderedModel	**AIC:**	66.75
Method:	Maximum Likelihood	**BIC:**	76.17
No. Observations:	24	**Pseudo R-squared**	0.151
Df Residuals:	16		
Df Model:	8		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
C(Affiliation)[T.liberal]	1.9414	1.447	1.342	0.180	-0.894	4.777
C(Condition)[T.topdown]	4.6327	1.842	2.515	0.012	1.022	8.244
C(Affiliation)[T.liberal]:C(Condition)[T.topdown]	-5.5537	2.239	-2.481	0.013	-9.942	-1.166
-3.0/-2.0	-1.7913	1.435	-1.248	0.212	-4.604	1.021
-2.0/-1.0	-0.2098	0.979	-0.214	0.830	-2.130	1.710
-1.0/0.0	-0.1707	0.685	-0.249	0.803	-1.513	1.172
0.0/1.0	1.2217	0.246	4.966	0.000	0.740	1.704
1.0/2.0	1.0069	0.452	2.230	0.026	0.122	1.892

Table 5: Ordinal logistic regression showing the effects of partisan affiliation and experiment condition on answers to pre-experiment survey Q5a ("Please rate the degree to which you agree or disagree with the following statements from 1 (Strongly disagree) to 7 (Strongly agree): Stricter environmental laws and regulations cost too many jobs and hurt the economy.") There was a significant interaction between affiliation and condition ($p < 0.05$) as well as a significant main effect from experiment condition ($p < 0.05$).

Dep. Variable:	Q11	**Log-Likelihood:**	-19.785
Model:	OrderedModel	**AIC:**	51.57
Method:	Maximum Likelihood	**BIC:**	58.64
No. Observations:	24	**Pseudo R-squared**	0.120
Df Residuals:	18		
Df Model:	6		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
C(Affiliation)[T.liberal]	2.0199	1.567	1.289	0.197	-1.051	5.090
C(Condition)[T.topdown]	3.9793	1.862	2.137	0.033	0.330	7.628
C(Affiliation)[T.liberal]:C(Condition)[T.topdown]	-3.2274	2.056	-1.570	0.116	-7.256	0.802
-1.0/0.0	-0.6170	1.171	-0.527	0.598	-2.911	1.677
0.0/1.0	1.4475	0.263	5.514	0.000	0.933	1.962
1.0/2.0	0.7996	0.455	1.755	0.079	-0.093	1.692

Table 6: Ordinal logistic regression showing the effects of partisan affiliation and experiment condition on answers to pre-experiment survey Q11 ("Imagine for a moment that you are moving to another community... In deciding where to live, how important would it be to live in a place where most people held political views similar to your own? (1 Not important, Somewhat important, Moderately important, 4 Very important).") There was a significant main effect from experiment condition ($p < 0.05$).

Dep. Variable:	Post_Ideological_Score	**R-squared:**	0.973
Model:	OLS	**Adj. R-squared:**	0.932
Method:	Least Squares	**F-statistic:**	23.50
No. Observations:	24	**Prob (F-statistic):**	2.27e-05
Df Residuals:	9	**Log-Likelihood:**	28.929
Df Model:	14	**AIC:**	-27.86
Covariance Type:	nonrobust	**BIC:**	-10.19

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.0383	0.326	-0.118	0.909	-0.776	0.699
C(Gender)[T.Male]	0.0940	0.106	0.891	0.396	-0.145	0.333
C(Gender)[T.Other]	0.0443	0.181	0.245	0.812	-0.364	0.453
C(Partisan Self-ID)[T.Democrat]	-0.1318	0.181	-0.727	0.485	-0.541	0.278
C(Partisan Self-ID)[T.Independent]	-0.0480	0.183	-0.263	0.799	-0.461	0.365
C(Partisan Self-ID)[T.Libertarian]	0.0286	0.127	0.225	0.827	-0.259	0.316
C(Partisan Self-ID)[T.Other]	-0.1071	0.225	-0.476	0.645	-0.616	0.401
C(Affiliation)[T.liberal]	0.0152	0.142	0.107	0.917	-0.306	0.336
C(Condition)[T.topdown]	-0.0073	0.066	-0.110	0.914	-0.157	0.142
Age Range	0.0348	0.051	0.684	0.511	-0.080	0.150
Education Level	0.0148	0.050	0.292	0.777	-0.099	0.129
Partisanship Strength	-0.0104	0.092	-0.114	0.912	-0.218	0.197
Political Engagement	-0.0532	0.101	-0.530	0.609	-0.281	0.174
Ideological_Score	1.0031	0.243	4.121	0.003	0.452	1.554
C(Condition)[T.topdown]:Ideological_Score	0.0076	0.145	0.052	0.959	-0.321	0.337

Omnibus:	2.666	**Durbin-Watson:**	2.528
Prob(Omnibus):	0.264	**Jarque-Bera (JB):**	1.757
Skew:	-0.663	**Prob(JB):**	0.415
Kurtosis:	3.024	**Cond. No.**	81.2

Table 7: Ordinary least squares regression to predict the post-experiment ideological score. Age, gender, education level, self-identified political affiliation, strength of political affiliation, political engagement, experiment political affiliation, experiment condition, and pre-experiment ideological score, as well as the interaction between the latter two, were all predictor variables. The pre-experiment ideological score was the only significant predictor of the post-experiment score ($p < 0.01$).

Dep. Variable:	Post_Affective_Score	**R-squared:**	0.957
Model:	OLS	**Adj. R-squared:**	0.891
Method:	Least Squares	**F-statistic:**	14.45
No. Observations:	24	**Prob (F-statistic):**	0.000174
Df Residuals:	9	**Log-Likelihood:**	28.393
Df Model:	14	**AIC:**	-26.79
Covariance Type:	nonrobust	**BIC:**	-9.115

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.2078	0.316	0.658	0.527	-0.506	0.922
C(Gender)[T.Male]	-0.0045	0.112	-0.040	0.969	-0.258	0.249
C(Gender)[T.Other]	0.1681	0.185	0.910	0.386	-0.250	0.586
C(Partisan Self-ID)[T.Democrat]	-0.1368	0.203	-0.674	0.517	-0.596	0.322
C(Partisan Self-ID)[T.Independent]	-0.0803	0.183	-0.439	0.671	-0.494	0.334
C(Partisan Self-ID)[T.Libertarian]	-0.1546	0.132	-1.173	0.271	-0.453	0.144
C(Partisan Self-ID)[T.Other]	-0.1355	0.199	-0.680	0.514	-0.586	0.315
C(Affiliation)[T.liberal]	-0.2008	0.122	-1.641	0.135	-0.478	0.076
C(Condition)[T.topdown]	-0.0201	0.061	-0.331	0.748	-0.157	0.117
Age Range	0.0084	0.050	0.168	0.870	-0.104	0.121
Education Level	-0.0409	0.050	-0.821	0.433	-0.154	0.072
Partisanship Strength	0.0158	0.084	0.188	0.855	-0.174	0.205
Political Engagement	0.0713	0.098	0.728	0.485	-0.150	0.293
Affective_Score	0.6243	0.162	3.856	0.004	0.258	0.990
C(Condition)[T.topdown]:Affective_Score	-0.0561	0.149	-0.376	0.716	-0.393	0.281

Omnibus:	2.733	**Durbin-Watson:**	1.407
Prob(Omnibus):	0.255	**Jarque-Bera (JB):**	1.482
Skew:	-0.296	**Prob(JB):**	0.477
Kurtosis:	1.936	**Cond. No.**	69.2

Table 8: Ordinary least squares regression to predict the post-experiment affective score. Age, gender, education level, self-identified political affiliation, strength of political affiliation, political engagement, experiment political affiliation, experiment condition, and pre-experiment affective score, as well as the interaction between the latter two, were all predictor variables. The pre-experiment affective score was the only significant predictor of the post-experiment score ($p < 0.01$).

Dep. Variable:	Post_Combined_Social_Score	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.757
Method:	Least Squares	**F-statistic:**	6.126
No. Observations:	24	**Prob (F-statistic):**	0.00489
Df Residuals:	9	**Log-Likelihood:**	27.308
Df Model:	14	**AIC:**	-24.62
Covariance Type:	nonrobust	**BIC:**	-6.946

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.3671	0.322	-1.140	0.284	-1.096	0.361
C(Gender)[T.Male]	0.0387	0.113	0.342	0.740	-0.217	0.294
C(Gender)[T.Other]	0.1892	0.190	0.995	0.346	-0.241	0.619
C(Partisanship Self-ID)[T.Democrat]	-0.0288	0.202	-0.143	0.890	-0.485	0.428
C(Partisanship Self-ID)[T.Independent]	0.2069	0.203	1.018	0.335	-0.253	0.667
C(Partisanship Self-ID)[T.Libertarian]	0.0939	0.133	0.709	0.497	-0.206	0.394
C(Partisanship Self-ID)[T.Other]	0.1629	0.215	0.758	0.468	-0.323	0.649
C(Affiliation)[T.liberal]	-0.1015	0.128	-0.795	0.447	-0.390	0.187
C(Condition)[T.topdown]	0.0881	0.066	1.329	0.217	-0.062	0.238
Age Range	0.0029	0.053	0.054	0.958	-0.117	0.123
Education Level	0.0283	0.052	0.543	0.600	-0.090	0.146
Partisanship Strength	0.0829	0.095	0.870	0.407	-0.133	0.299
Political Engagement	0.0373	0.104	0.360	0.727	-0.197	0.272
Combined_Social_Score	0.4822	0.265	1.822	0.102	-0.116	1.081
C(Condition)[T.topdown]:Combined_Social_Score	0.5158	0.244	2.110	0.064	-0.037	1.069

Omnibus:	0.585	**Durbin-Watson:**	2.351
Prob(Omnibus):	0.747	**Jarque-Bera (JB):**	0.669
Skew:	-0.217	**Prob(JB):**	0.716
Kurtosis:	2.307	**Cond. No.**	74.9

Table 9: Ordinary least squares regression to predict the post-experiment combined social score. Age, gender, education level, self-identified political affiliation, strength of political affiliation, political engagement, experiment political affiliation, experiment condition, and pre-experiment combined social score, as well as the interaction between the latter two, were all predictor variables. There were no significant predictors in the model, but the interaction between experiment condition and pre-experiment combined social score approached significance ($p = 0.064$).