Digital Juries: A Civics-Oriented Approach to Platform Governance

Jenny Fan
Harvard University
Cambridge, MA
fan.jenny@gmail.com

Amy X. Zhang
University of Washington
Seattle, WA
axz@cs.uw.edu

ABSTRACT

As concerns have grown regarding harmful content spread on social media, platform mechanisms for content moderation have become increasingly significant. However, many existing platform governance structures lack formal processes for democratic participation by users of the platform. Drawing inspiration from constitutional jury trials in many legal systems, this paper proposes *digital juries* as a civics-oriented approach for adjudicating content moderation cases. Building on existing theoretical models of jury decision-making, we outline a 5-stage model characterizing the space of design considerations in a digital jury process. We implement two examples of jury designs involving blind-voting and deliberation. From users who participate in our jury implementations, we gather informed judgments of the *democratic legitimacy* of a jury process for content moderation. We find that digital juries are perceived as more procedurally just than existing common platform moderation practices, but also find disagreement over whether jury decisions should be enforced or used as recommendations.

Author Keywords

content moderation; platforms; social media; online speech; democracy; civics; juries; governance; institutional design

CCS Concepts

•Human-centered computing → Collaborative and social computing;

INTRODUCTION

Toxic content on social media, such as hate speech [29], misinformation [92], and conspiratorial "dog-whistling", or coded messages [16], has been well studied in terms of its harm to individuals and society in addition to its challenge for content moderation. The scale and ease by which content spreads on platforms have raised new alarm about online speech that incites violence [4, 88], radicalizes public discourse [95, 12], and even impacts elections [9].

Many platforms have struggled with how to adjudicate this content, much of which is borderline [97] and requires knowledge of local sociocultural norms and other context [42, 60]. Part of the challenge is that platforms are tasked with making difficult decisions about speech standards that profoundly affect public discourse [19]. Meanwhile, the processes that many large commercial platforms employ for content moderation—namely paying human content moderators [42, 78] and training oftentimes biased or brittle algorithms [5, 52] to spot violations—do not draw upon the perspective of users beyond superficial tasks such as flagging [25]. In their adopted role of the "new governors" of speech [59], social media platforms risk losing democratic legitimacy [87, 31].

A major corollary for how citizens can be democratically involved in governance decisions is the jury process in many legal systems, such as the American civil jury. In this paper, we consider how this process could translate online and propose *digital juries* as a civics-oriented approach for adjudicating online content moderation questions. Building on existing theoretical models of jury decision-making [49, 32], we present a 5-stage model outlining the space of considerations when designing a digital jury process: jury selection, onboarding, case trial, consensus formation, and outcome enforcement.

We then gather empirical evidence to explore whether digital juries are perceived as more democratically legitimate than the status quo of paid and automated moderation, as well as how aspects of jury design relate to perceptions of democratic legitimacy. We implement two prototype jury workflows that vary the consensus formation stage of our model, with one emphasizing blind voting and the other emphasizing group deliberation. We recruit 82 "jurors" to make decisions online about difficult content moderation cases using our workflows in groups of around 6. These experiences allow our participants to gain an impression of how a digital jury process could potentially work, as no well-known examples currently exist. We then survey jurors to capture whether they found the different processes for making a decision to be democratically legitimate, following a framework of procedural justice.

From our study, we find that digital juries improved user perceptions of justice in the *process* of content moderation on five different attributes, *legitimacy, trust, equality, fairness,* and *care*, though not in *efficacy*, compared to the status quo of content moderation decisions arrived at through automation and paid moderators. We also find evidence that users preferred a deliberative jury over a blind-voting jury. Finally, jurors had conflicting opinions about whether enforcing jury outcomes as-is or using them as recommendations would be more democratically legitimate. We conclude with a discus-

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376293

sion on the feasibility and design criteria for such a system in moderating speech at scale. This is a challenge made trickier with competing views on trust and fairness, as we found in qualitative feedback. Altogether, our results suggest that a change towards a civics-oriented digital jury process could help re-frame users' relationship to platform speech regulators.

RELATED WORK

A "citizen-sovereign" view of platform governance

Across platforms, regulation in online spaces has been proposed as consisting of four modalities—law, markets, norms, and architecture—that can interact and intersect [70]. Within online social platforms, regulation is operationalized by moderation practices [42, 65] that form a wide taxonomy of tactics [44], from norm-based administrator feedback to technical actions such as banning or blocking [13]. There are multiple theoretical frameworks for how platforms gain governing legitimacy to regulate behavior, ranging from contractual [2] to constitutional [86]. Unlike a contractual, merchant-sovereign lens [76], a constitutional, citizen-sovereign lens of platform governance [70] re-frames the role of the user from customer to citizen. The regulatory role of platforms is also re-framed from market vendor to governing sovereignty. Legal scholars acknowledge that the law is designed for a merchant-sovereign relationship [2, 86] while a citizen-sovereign view may be desirable to reflect platform realities.

Most industrial-scale platforms [17] today rely on a combination of proactive algorithmic filtering, user-flagged reports of offensive content [25], and *ex post* human moderation for oversight. Although CEOs like Mark Zuckerberg have suggested governance structures that are quasi-constitutional in nature [97], ranging from a "Supreme Court" to an appeals process and independent oversight committee, these policies lack civic participation from users as stakeholders and decision-makers [50]. As platforms increasingly intervene in speech regulation [41, 57], scholars and regulators alike are raising concerns about their lack of procedural fairness and accountability to the public [24, 86], calling into question the democratic legitimacy of their moderation practices [31, 59].

Self-governance in online communities

Meanwhile, a number of online communities have a rich history of establishing legitimacy through self-governance [75]. Researchers have studied platforms that take a community-driven approach to moderation, observing the governance structures that develop bottom-up over time and finding the factors that drive success [65]. Much research has focused on Wikipedia's decentralized structure, including the many locally governed, self-contained WikiProjects [37] and emphasis on open deliberation to bestow responsibilities [14] and resolve disputes [53]. Research has shown that successful online communities have greater structure [93] and more diverse rules [39]. However, Wikipedia's highly flexible governance has resulted in a system that is considered bureaucratic [15] and criticized as difficult to navigate or hostile for newcomers [45]. Another platform with a more localized governance structure is Reddit [21], where many rules are governed at the "subreddit" or group level. However, while anyone may start a subreddit and become an administrator, individual subreddits have formal mechanisms for rule creation and enforcement only by moderator users [35], as opposed to by the subreddit members as a whole. The same can be said for Minecraft servers, Mastodon instances, or many other federated systems. It is unclear whether the concentration of power in the hands of administrators is due to the fact that it promotes success [39] or that the existing software more easily supports it by design.

While the above platforms have hierarchical governance systems with roles for administrators and moderators, platforms can also take a more crowdsourced approach to governance. Systems such as League of Legends (LoL) tribunals [63, 67], Weibo committees [62], Slashdot moderators [68], and Facebook's short-lived policy voting system [91] allow a large proportion of members to participate in governance decisions. In addition, many social platforms today have some element of crowdsourced social moderation involving voting on pieces of content that then alter their visibility. While more inclusive and arguably more democratic by design, one drawback to their current design is that the scale of participation means that each individual contributes at a granular level, acting as "human processors" [67]. In addition, they rely on the platform as a centralized clearinghouse for receiving and assigning cases or aggregating votes using an oftentimes opaque algorithm.

Deliberative democracy and the citizens' jury

Alternatively, a citizen jury system could also allow for democratic participation but with the opportunity to have a deeper and more deliberative process for consensus-building. Theories of deliberative democracy posit that democratic legitimacy comes from *authentic deliberation* on the part of those affected by a collective decision [18]. The role of juries as group decision-making systems is two-fold, as a mechanism for determining a group's preference structure and for determining social norms and their relationship to law [3]. Many crowdsourced governance systems are primarily designed for voting at scale to assess the former rather than the latter as a norm-enforcing institution. As a result, they are often considered a transitional step in automating human judgment [61], rather than maintained as a source of participatory governance.

While critics of a jury system debate the drawbacks of human judgment and group deliberation, such as perceived irrational bias or insufficient expertise, proponents cite the benefits of juries as an error-checking system, mechanism for incorporating diverse views [48], and means of facilitating cooperation and shaping one's social identity within groups [90]. The primary advantage is *procedural*, offering all jury members a chance to deeply engage in the decision-making process and endorse a resulting consensus [47]. Existing literature of public participation in deliberative forums, including juries, report a number of benefits when the process is perceived as fair [18]. In addition to increased perceptions of procedural fairness [51], jurors report a higher sense of legitimacy of a governing institution [40] and increased civic engagement following jury service [46]. The civic labor of moderation [72] also shapes a participant's identity as "citizen" or "juror" [73].

In LoL tribunals—a rare instance of a governance system resembling digital juries in practice—participants reported the

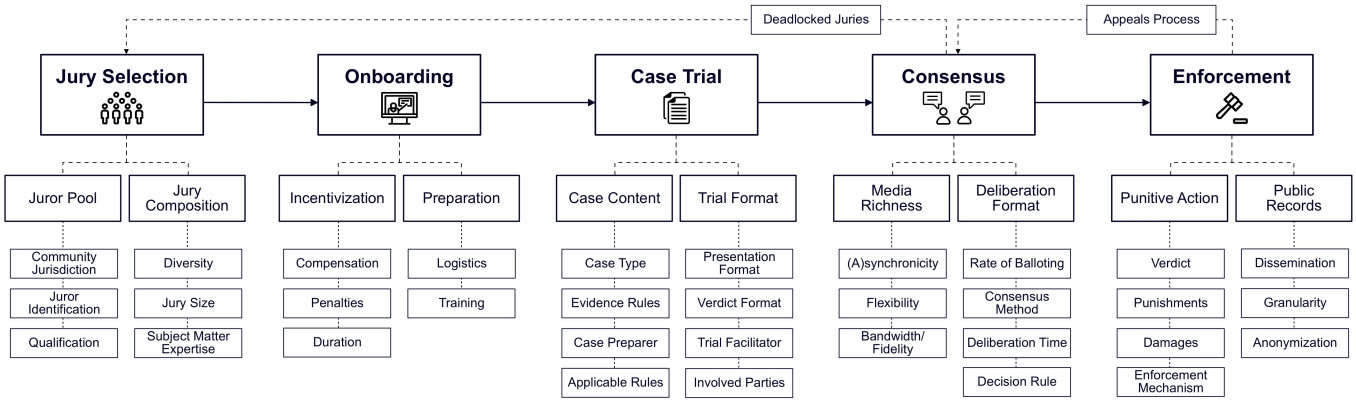

Figure 1. The Digital Jury Model shows the stages and process flow in a digital jury, along with design parameters within each stage.

experience as helpful for learning norms and building community [67]. However, the game developer Riot Games prioritized efficiency over legitimacy in design, leading to decisions such as a lack of transparency about cases or votes, random assignment of judges to cases, and no interaction between judges (despite user-initiated efforts to communicate) [61]. The lack of deliberation resulted in a system offering a trial by many *judges*, but was not truly an implementation of trial by *jury*. After three years, the tribunals were replaced by a more opaque but also more efficient automated system.

Design of deliberation and adjudication systems

The design of our digital jury implementation draws from research in deliberation and crowdsourcing interfaces. Systems for improving online decision-making are well-explored in social computing [63, 33, 96, 66] and grapple with the difficult tension of presenting quality argumentation within time and attention constraints. The *Virtual Agora* project compares face-to-face deliberation with online deliberation, measuring its impact on decision quality and perceived legitimacy of choices [73], though all sessions were relatively time-consuming. Other studies examine how to improve the time efficiency of online deliberation. The *Microtalk* system structures comments into discrete, persuasive arguments for one-round debate [30]. Other systems expand on Microtalk's structure with synchronous workflows for multi-turn and contextual argumentation [80], resulting in improved accuracy [23]. Finally, the *Stanford Online Deliberation Platform* conducts group video sessions for civic deliberation using an automated facilitator [36]. These systems inform our design of experimental digital jury processes. However, while many of these workflows have been evaluated on the quality of the outcome, such as accuracy in finding a pre-determined answer, in our case, we are primarily interested in how the design of the jury process affects perceived democratic legitimacy.

MODEL OF DIGITAL JURY DESIGN

In this section, we present the *Digital Jury Model (DJM)*, a 5-stage model of a digital jury system as a sociotechnical implementation inspired by a constitutional jury system. We outline the design considerations at each stage (Figure 1). In order to develop design parameters for a distributed, online-only adjudication system, we look to existing frameworks of digital constitutionalism and jury decision-making for validity.

Digital constitutionalism based on the rule of law has been proposed to evaluate the legitimacy of governing institutions, including private, online platforms [86]. To incorporate constitutional values of consent, predictability, and procedural fairness, Frey et al. posit that institutions must have formal, direct mechanisms (inspired by the Ostrom workshop [75]) for "low-level agents" to participate in rule-making and enforcement [38]. Theoretical foundations for how such agents could serve as jurors is most informed by Hastie et al.'s *Inside the Jury*, an empirical study spanning 828 mock jury participants [49]. Hastie's psychocognitive model has been thought to more accurately model juror behavior in contrast to mathematically-based alternatives, which depict an idealized juror that rationally listens to evidence with a prejudice-free thought process [94]. This model was further explored by Epstein's *Agent_Zero* model of a three-phased jury process, which builds upon Hastie et al. to incorporate social dynamics and consider the broader system around a trial, such as juror selection [32]. To translate these models to the digital context, where platforms and communities can be far more specialized and diverse in their norms, we expand upon prior models to explicitly incorporate an onboarding and enforcement stage.

Stage 1. Jury selection

Juror pool *(jurisdiction, juror identification, qualification)*: The jurisdictional boundaries from which a "jury of one's peers" is formed could be defined in many ways with the help of technology, such as within geographic locality (e.g., country designation or distance from a point), community boundaries (e.g., members of a specific thread, group, or network), affinity groups (e.g., users with particular tags, interests, or "likes"), or even audiences of the content in question, such as is the case with social moderation. Once selected, jurors could either be identified to each other in the jury, kept anonymous, or use pseudonyms while communicating to each other. For instance, the LoL tribunals had anonymous judges [62]. Jurors may also go through a qualification process, such as the *voir dire* examination of a potential juror done by a counsel to make

disqualifications in civil juries. More broadly, jurors could be vetted, such as Slashdot's meta-moderation [68], where moderators evaluate other moderators.

Jury composition *(diversity, jury size, expertise)*: The selected jury should be from a representative cross-section of impartial peers, with no cognizable class or group of users excluded from the selection process [49]. This could mean a proportional representation of the entire community or representative of the affected parties. As a digital system, large platforms have the ability to incorporate a wider or more targeted juror selection pool or enforce greater juror diversity based on collected data about users. This pool could also span varying lay or expert users, depending on the level of subject matter expertise required for the case. In terms of jury size, research has found that in juries of more than 6, participants are more reluctant to speak out or disagree with the majority [49]. While most offline juries are statically sized, in the online case, juries can be flexibly sized and rapidly sourced, such as in LoL tribunals [61]. The same case could also be run across multiple juries of varying make-up or size to gather inter-jury reliability.

Stage 2. Onboarding

Incentivization *(compensation, penalties, duration)*: After selecting jurors, a jury system should consider how to motivate jurors in proportion to the amount of time they are expected to participate. Micro-tasking sites like Amazon Mechanical Turk have set a familiar pattern of micro-payments for short human intelligence tasks, a pattern that has also been used by platforms to pay moderators [78]. In municipal courts, a minimal compensation for jury duty is common, although the strongest motivator for jury duty is the penalties for ignoring summons. In contrast, many crowdsourced governance systems rely on volunteer judges and moderators who may wish to give back to the community, develop social capital [77], or learn from cases [64]. However, few online systems have explored making broad participation in governance mandatory or tied to explicit rewards. One example is the now-defunct Civil Comments platform that required users to vote on other comments before they could comment [71].

Preparation *(logistics, training)*: Selected jurors might need to prepare before becoming a juror. This preparation may be logistical, such as scheduling a time (if synchronous), installing necessary tools, or arriving at the right online destination and authenticating themselves. Preparation may also involve training jurors about how cases should be evaluated. Within some settings such as Wikipedia, decision criteria are determined by users themselves in the form of collaboratively-edited community guidelines [15]. In other cases, platform operators release guidelines. Finally, jurors could be trained in how to be a good juror. For instance, trial cases could be taught in an interactive tutorial environment using a case study method. Instructions could be delivered via a human facilitator or a chat bot.

Stage 3. Case Trial

Case content *(case type, evidence rules, case preparer, applicable rules)*: Juries would see different types of cases depending on whether they were conducting rule-making, adjudication, or applying human intelligence where algorithms struggle (e.g., detecting misinformation or "newsworthiness"). The standards of what constitutes evidence and how it is evaluated in each case could be derived from national laws, platform rules, community norms, or precedents set by previous cases. Some evidence, such as behavior of an individual off-platform, may be considered in or out of scope. For example, the rules of evidence in LoL allow judges to see all reports and their comments as well as raw in-game chat to gain context [64]. Cases should be prepared in a way that reduces biases and incorporates different viewpoints. This may require additional human labor or technical tools to gather evidence and supporting documents and to write a report prior to presenting the case to a jury. Finally, it should be clear which set of rules and standards a case should be evaluated against, particularly in cases of conflicting community standards or national laws.

Trial format *(presentation format, verdict format, trial facilitator, involved parties)*: The core of the model involves design decisions regarding the mechanics of how the trial is conducted, including who is involved in the process and the format of their decisions. The standard format of how case content is presented to jurors could prioritize certain readings or behavior. Depending on the type of case, jurors may have to deliver a verdict that not only attributes guilt, but also evaluates the level of damage (e.g., toxicity of a harmful content) or even suggests appropriate punishments and consequences. In order for the trial to proceed in a particular way, there may be one or more trial facilitators who play the role of judge, clerk, or foreperson (a spokesperson of the jury). A trial facilitator could be a juror, a separate user, or a platform employee. Aggrieved parties could also be incorporated in different ways, ranging from completely abstracting away associated users to involving them directly in presenting the case. In systems such as LoL and Weibo, the reported user and reporter can make arguments directly to judges [62].

Stage 4. Consensus

Media richness *(synchronicity, bandwidth, fidelity)*: Media Richness Theory suggests that richer communication media (such as in-person, video conferencing, or synchronous text-based conversations) are more effective for communicating [26] and facilitating trust [10] than leaner media, such as asynchronous email. However, richness comes at a direct cost to scaleability. Because online juries are limited to computer-mediated interaction, the impacts of media format must be considered in evaluating the two axes of *scaleability* (ease of implementation at scale) and *immersiveness* (richness of communication media). Synchronous communication such as chatrooms or video conferencing are more immersive, while asynchronous comments and voting are easier to scale.

Deliberation format *(consensus method, deliberation time, decision rule, rate of balloting)*: Online consensus building must necessarily be constrained by time limits, consensus method (e.g., voting, Delphi method [27], blocking), and decision rules (e.g., simple majority, super majority, unanimity) as well when and how frequently votes are taken (rate of balloting). Time limits imposed on deliberation may trade off with deliberation quality [58], but even semi-synchronous ju-

Figure 2. After completing onboarding, participants see information about one of three possible cases. Below information about the case, they see instructions and space for providing input and interaction, depending on the condition. After all inputs are in and a minimum time has passed, the final decision and actions are posted on the right pane. In the jury conditions, the right pane updates in real-time if any juror chooses to change their vote.

ries would likely require an enforced time limit to prevent juror dropout and cases becoming irrelevant. For instance, Wikipedia deliberations, which happen asynchronously, are closed by default after 30 days [53]. The rate of balloting determines when and how frequently a vote is taken to check for quorum, affecting faction formation in the deliberation process [49]. Various decision rules from simple majority (51%) to unanimity affect not only the verdict and deliberation time, but also juror satisfaction. For unanimous decision rules, juries take longer to deliberate and are more likely to hang than majority-rule juries [49], but jurors report higher satisfaction and a greater sense of civic responsibility [8, 74, 46]. LoL tribunals require only a majority to reach a verdict [64].

Stage 5. Enforcement

Punitive action *(verdict, punishments, damages, enforcement mechanism)*: While crowdsourced governance systems mostly focus on binary verdicts, jurors could also specify sanctions and scale them with the severity of the violation. In content moderation, sanctions could be made for content as well as users. Content could be de-amplified, deleted, or affixed with a label, and users could be sent warnings, placed in a time-out, or banned from the platform. Platforms should have mechanisms to enforce these sanctions, such as formalized processes for recommending jury decisions to the platform or enforcing the outcome directly, or else they risk undermining legitimacy by lacking credible commitment [50].

Public records *(dissemination, granularity, anonymity)*: The results of each jury's findings could be publicly disseminated for transparency and learning. Records could be released at different levels of granularity, such as by case, by jury assembly (i.e., groups of users), or by juror. In platforms where users are likely to participate multiple times, juror-specific records could help audit for abuse of the system as well as identify exemplary jurors for additional roles, such as case preparation or trial facilitation. However, in sensitive cases, juror identities should be anonymized for their protection.

STUDY

We conduct an empirical study to investigate what design qualities would allow content moderation processes to be perceived as more democratically legitimate. Using participants recruited from a crowdsourcing platform, we gather feedback on three workflows (Figure 2). For two of the workflows, we implement a "minimum viable" digital jury process, hosted on the website digitaljuries.com. By implementing a functioning jury system, we are able to weigh design along with feasibility considerations, as well as provide users with first-hand experience of being in a digital jury.

Guided by the model parameters identified earlier, we design and implement three randomized conditions: two are jury workflows, with one emphasizing blind voting (IMMERSIVE) and one emphasizing deliberation (SCALEABLE). For these two conditions, we place participants into juries of around 6 people to decide one case per workflow and survey them about their perception of the process after each case. We also show users a third condition (STATUS QUO) with a scripted case outcome, presented as having been decided by the platform's algorithmic and paid human moderation. As in the jury conditions, users are surveyed about their perception of this process as described to them. This allows users to contrast their jury experiences with a process akin to the status quo of many commercial social media platforms today.

Following the three conditions, we provide free-response prompts to each user to survey which of the three processes—the status quo of no user input, a blind voting jury, and a deliberating jury—they preferred as a way for platforms to moderate content and why. We also ask users whether they would prefer the outcome of a jury process like the ones they participated in to be directly enforced or taken as recommendation by a platform. From these free-response answers as well as post-round surveys, we gain insight into the following research questions:

- **RQ1:** How do users perceive the democratic legitimacy of a content moderation process involving user juries vs. the status quo of no user input?
- **RQ2:** How do users perceive the democratic legitimacy of a jury process with consensus by voting vs. by deliberation?
- **RQ3:** Would users prefer jury decisions to be directly enforced or passed on to platforms as recommendations?

Procedure and Conditions

We design our study to be within-subjects so that users can compare all three content moderation processes. The order of conditions and cases are counterbalanced. The first 3 stages of the 5-stage model are held constant for all users:

Stage 1 (juror selection) has participants recruited from Amazon Mechanical Turk (AMT), limited to US-only and ages 18+. Tasks are released on AMT in batches during peak times to maximize the likelihood that users are synchronous. Upon arriving at the site, users are informed about the potentially harmful content they will judge and must consent to continue. As users join the study, they get added to a group. Each group is set to 6 people. **Stage 2 (onboarding)** instructions are standardized at the beginning of the study, with no supervisory roles defined. Jurors receive a short explanation about content moderation, the web interface, and the different assessments they can make. In **Stage 3 (trial)**, jurors see one of three standardized cases that contain a screenshot of the violating post, background context, and list of potential rule violations. In **Stage 4 (consensus)**, the conditions diverge:

STATUS QUO *(no user input)*: In this condition, there is no user input from the participant, and users are simply shown the platform's decision and rationale. Consistent with the status quo of commercial content moderation on many platforms today, the outcome is described as a mix of algorithmic and paid moderator effort.

SCALEABLE *(semi-synchronous, blind-voting jury)*: Optimizing for scaleability, jurors independently decide on a toxicity score and recommended action in a one-round blind vote, similarly to systems like LoL tribunals. Jury consensus is based on an average of the toxicity scores and majority rule for the recommended action. After voting, participants wait for two minutes before the rest of the jury's results are shown to them, allowing time for other members of the jury to vote.

IMMERSIVE *(synchronous, deliberating jury)*: Optimizing for immersive communication, participants are able to freely deliberate with other jurors in a chatroom for four minutes before submitting their vote. After the minimum deliberation time elapses, jurors are allowed to continue discussion for as long as desired, as well as are able to freely change their vote.

Participants go through each condition with a different case and are surveyed after each condition, as well as after all three conditions. The order of cases are selected randomly. Finally, to investigate **Stage 5 (enforcement)**, we conclude with a survey that asks for participants' desired enforcement level for the jury decisions: either as *recommendations* (taken into advisement but with no action required) or *enforced* decisions (binding actions directly implemented).

Cases and Assessments

Our study uses three cases as there are three within-subjects conditions. The cases are written by the first author and designed to be contextually nuanced and borderline with respect to violating standards.[1] Each case contains a screenshot of the violating post, background and context about the issue, and a list of potential standard violations drawn from Facebook's Community Standards. Chosen topics are meant to have disjoint coverage of topics that have received some public comment from major platforms, though in practice, some cases had overlapping themes. This study features:

- *Hate speech:* An antisemitic comic of Pepe the Frog.
- *Graphic violence:* A mixed-reaction post linking to the Christchurch shooter manifesto.
- *Child safety:* A link to a children's cartoon, altered to depict disturbing content of the urban legend Momo.

In the two jury rounds, jurors provide three assessments:

- Toxicity score of content, defined as likelihood to cause harm (0-3 OK, 4-7 Borderline, 8-10 Toxic), mimicking [55].

[1] We provide details about each case within Supplementary Materials.

- Punishment for the content, if any (`unlist` from users' feeds, `delete` from the site, and `report` to authorities).
- Punishment for the user, if any (`warn`, `ban` for 1 week, and `permanently ban`).

Measurement and Data Analysis

The inherently subjective nature of values-based judgments is challenging to measure. To measure the democratic legitimacy of a digital jury, procedural justice frameworks provide a way to evaluate success. Procedural justice asserts that regardless of *outcome*, an individual can be satisfied with the system if the individual considers the underlying *process* to be just [89]. Procedural justice frameworks evaluate the fairness and procedural regularity of decision-making processes, such as the opportunity for participation, transparency, neutrality, and objectivity of decision-makers, belief in fair motivations, and dignity in standing that the system offers to them [90].

We structure survey questions of each condition's democratic legitimacy from participants using variants of Tyler's procedural justice framework [89, 69] (e.g., perceived fairness of the *process*, satisfaction with the *outcome*) and Haidt's Moral Foundations theory [43], which describes a common set of universal moral values persistent across cultures (e.g., care, fairness, and authority). From these sources, we survey participants on *outcome* satisfaction and six core criteria that we then ask participants to rate with regards to the *process* of each condition: legitimate exercise of power, trust, equal valuing of individual voices, fairness, care of personal preferences, and efficacy in moderating content.

To measure differences in the mean perceptions of procedural justice between the three conditions, we conduct Univariate Type III Repeated-Measures (within-subjects) ANOVA tests on the six ratings related to procedural justice, followed by post hoc Tukey HSD tests when differences are significant. This is calculated in `R` using the `car` package. We also run tests on users' self-reported sense of time constraints and difficulty of decision-making in the jury conditions, along with satisfaction with the outcome of each case.

In the final survey following all three conditions, jurors provide open-ended responses regarding whether they would prefer enforcement or recommendation of a jury's outcomes. They also provide open-ended responses as to which content moderation process they preferred and why. To analyze all sets of qualitative responses, the first author used a standard open-coding approach [22] to code the *values* expressed in each response. They then grouped the resulting 35 codes into major categories of values. Codes and categories were iteratively discussed during the process with all authors.

Participants

We chose to recruit from AMT because the IMMERSIVE condition requires synchronous activity, and AMT is a large enough platform where it would be likely for 6 workers to join a task near in time to each other. Due to the logistical constraints of recruiting for synchronous trials, we did not sample to achieve a certain demographic proportion. However, this would be a reasonable expectation for a platform since major platforms have even larger active user bases.

Jurors were paid $6 for the task and spent an average of 22.21 minutes. In total, we record 82 active participants across 15 juries. Each jury is set at 6 people each, though due to availability and drop-outs, 8 of the juries have 5 people. Of the participants, 35 (43%) are age 25–34, while 19 (23%) are 35–44. Of the remaining, 12 (15%) are age 45–54, 9 (11%) are age 18–24, and 3 are 55–64. 67% of participants are male, while 27% are female. All participants have some high school education, with 48 people (59%) completing college or higher degrees. When it comes to political affiliation, 46% identify as Democrat, 22% as Republican, and 22% as Independent. It is notable that the demographics of the participants skew young, male, and Democratic, perhaps reflecting the demographics of the AMT population at the time of data collection [28].

RESULTS

RQ1: User Juries versus Status Quo of No User Input

We explored users' perceptions of democratic legitimacy when it came to content moderation by jury (both IMMERSIVE and SCALEABLE) versus by the STATUS QUO. Table 1 shows results for the survey questions. Content analysis of the open-ended responses supports the presence of a procedural justice framework among participants, with most users discussing process over outcome when comparing conditions. Only a few (8) of the reflections remarked on disagreement in the outcome itself, with one person after the SCALEABLE condition saying: "*I didn't like the outcome. It was somewhat frustrating. I want to know why other users voted differently from me.*" There was also no statistically significant difference in ratings between conditions when it came to satisfaction with the decision.

From the ratings after each condition, participants expressed a greater sense of procedural justice in the jury conditions versus the status quo on all aspects of procedural justice except for "efficacy". From the open-ended responses for why they favored either of the jury conditions, participants described **democratic** values of popular sovereignty, equality, and justice, as well as **humanistic** values of trust in humans. For instance, 30 respondents mentioned how user input in platform governance would lend greater legitimacy. One person said: "*It would mean that actual people's voices matter...Since they are the ones using the site, what they say should matter more.*" Participants also described a feeling of empowerment from not needing to rely on moderators: "*...my voice was actually being heard. Instead of...hoping for moderation, I would feel connected to the process and like I actually had a part and some control in making decisions.*" In terms of humanistic values, 12 respondents expressed distrust in automation and preference for human insight: "*I like the human element... it involves the use of compassion and that cannot be programmed into AI.*"

Overall, the STATUS QUO condition was the least preferred out of the three conditions by 55% of participants. Twelve respondents described a lack of fairness and due process in the STATUS QUO. However, values related to **efficacy**, such as time efficiency or quality of outcome, was one dimension respondents cited when they discussed the downsides of juries. For instance, one person placed more trust in the expertise that would develop in a paid moderator and algorithm process: "*It*

Question *(1=Strongly Disagree, 5=Strongly Agree)*	Average Rating Status Q.	Scale.	Immers.	Type III SS	df	F	p	post hoc Tukey HSD Test
How satisfied are you with the decision?	3.46	3.44	3.92	7.08	2	2.286	$p = 0.105$	
This process feels like a legitimate exercise of the social media platform's power.	3.25	3.84	4.01	21.50	2	8.971	$p < 0.001$***	immersive-status quo*** scaleable-status quo**
This process improves my trust in how content moderation decisions are made.	2.91	3.43	3.86	32.95	2	12.335	$p < 0.001$***	immersive-status quo*** scaleable-status quo*
This process values individual voices equally.	2.64	3.69	4.06	78.93	2	31.281	$p < 0.001$***	immersive-status quo*** scaleable-status quo***
This process is fair.	3.23	3.74	3.94	18.07	2	6.557	$p = 0.002$**	immersive-status quo*** scaleable-status quo*
This process cares about my preferences.	2.41	3.43	3.87	80.41	2	34.600	$p < 0.001$***	immersive-status quo*** scaleable-status quo*** scaleable-immersive*
This process is an effective way to protect users from unwanted or toxic content.	3.35	3.64	3.84	7.10	2	2.833	$p = 0.062$	

Table 1. After each condition of STATUS QUO, SCALEABLE, and IMMERSIVE, questions were asked about the procedural justice of the process (rows 4-9), as well as time constraints, satisfaction, and difficulty. Average ratings across the conditions and results of statistical significance tests are reported. Significance codes: < 0.05*, < 0.01**, < 0.001***

combined an algorithm...with a paid person who presumably has training and guidelines and a decision making process. I don't know if I trust a random collection of individuals...[it] could lead to a lot of bad outcomes." This suggests more could be done to not only train jurors on the rules, but also consider qualities that would facilitate trust in the process.

RQ2: Jury Consensus by Deliberation versus Voting

The differences between the two jury conditions was less pronounced, with only "care" being perceived as significantly greater in the IMMERSIVE (deliberating) condition over the SCALEABLE (voting) condition. However, 57.5% of participants most preferred the IMMERSIVE condition, while only 35% most preferred the SCALEABLE condition.

In the IMMERSIVE condition, jurors each published on average 4.01 chat messages (38.9 words). As jury groups, juries exchanged an average of 21.93 messages (212.4 words). In open-ended responses, 20 participants mentioned the **democratic** value of exposure to diversity of viewpoints in a deliberation, with one person saying: *"because I got to see other people's rationale...it might help me temper my own biases... It made me question whether I was overreacting...in my moderation."* Fourteen jurors also described how the deliberative aspect gave them the chance to work together towards consensus: "*...we all collaborated and came to the same conclusion. It was good to receive and give insight to my peers.*"

However, participants also described aspects they did not like about deliberation. Eight people mentioned lower **efficacy**. One user identified a trade-off between efficacy and richer user input, with SCALEABLE achieving the best balance between all three conditions: *"...it was efficient, allowed each individual to provide their personal opinion/preference."* Twelve participants emphasized **individualism** and felt that deliberation would bias jurors or cause "groupthink" [54]: "*It felt the most honest way, because you are going by your own opinions rather than be convinced of someone else's.*" A few participants also described disliking hearing others' perspectives or conversing with others in this setting: "*It lets me vote without hearing some other people's disregard of real issues...Round 2 [deliberation] made me more angry than any other.*" This suggests that facilitation or more structured workflows [36] could perhaps help guide deliberation.

Echoing participants' comments, we saw that the standard deviation of toxicity votes was greater in the SCALEABLE than in the IMMERSIVE condition (2.44 vs. 2.39, respectively), echoing prior work that deliberation can reduce disagreement [80]. Interestingly, we also saw that the SCALEABLE juries were overall more harsh. Users gave toxicity scores of 7.42 on average in the SCALEABLE condition compared to 6.59 in the IMMERSIVE condition, though this difference was not significant. In punitive actions for the user, juries chose to warn the user with 45 votes in IMMERSIVE and 30 in SCALEABLE. However, in the SCALEABLE condition, users voted to ban for 1 week or permanently ban users at nearly twice the rate of the IMMERSIVE condition (13.3% and 15.6% for IMMERSIVE, respectively, versus 24.4% and 25.6% for SCALEABLE).

RQ3. Recommendation versus Enforcement

We saw no statistical difference in preference for either enforcement or recommendation of jury decisions with both options at a mean of 3.34 and mode of 3 on the 5-point Likert scale. Despite a lack of a clear preference, participants expressed strong views both in favor and against the legitimacy of direct enforcement in open-ended responses. Proponents of direct enforcement were concerned about the lack of **accountability** to the public. Remarking on the status quo, one juror said, "*Frankly, [platforms] already receive recommendations, albeit not from a formal jury, right now. It has done little to nothing to stem toxic content—even when highly influential voices with absolutely enormous followings weigh in.*" Twelve jurors expressed cynicism with regards to the motivations and interests of large platforms such as Facebook and Twitter towards listening to users. One person said "*If it's just a recommendation, it would feel like a waste of time... It's just something they will ignore when they want to and cite as evidence of caring about users opinions when it's convenient.*"

Supporters of recommendation preferred for platforms to have the final say, with 12 jurors expressing a lack of confidence in

the quality or neutrality of juror input. Citing concerns about juror bias, one user said, "*You could just get one-sided votes or opinions...The companies should be managing what is right or wrong, because the public doesn't seem to be able to do it on our own anymore.*" This distrust of public discernment is directly at odds with the competing view of distrust of platform discernment, presenting a paradox in platform governance. As one pro-enforcement user stated, "*Social media has proven that they can't police themselves unless it hurts their bottom line. An enforced decision at this point might improve things for everybody. Is worth a shot.*" Four jurors additionally advocated for **transparency** in publishing jury decisions, including as an accountability mechanism when doing recommendation: "*I'd like for there to be a public record of what the jury decided so that the platform would have to explain themselves as to why they went against the decision.*"

DISCUSSION

Conflicting views on democratic legitimacy

Competing views on who to trust and what is fair emerged within the context of this study, exposing two possible mental models for how participants may assess the democratic legitimacy of a digital jury system.

12 participants shared some level of distrust towards automation, favoring human input. The lack of user—not just human—input also frustrated participants, who connected it to a perceived decline in platform quality: *"It was frustrating to have no input whatsoever into the process. It is literally why the nether regions of Twitter have become breeding grounds for neo-Nazis."* Distrust in opaque processes appears in other systems as well, such as among the LoL tribunal judges who circulated rumors in the absence of information [62].

On the other hand, critics of a user-led jury mentioned distrust in the opinions and motivations of other jurors. One juror said, *"I don't trust other people's opinions...I would rather have one person be a consistent judge and jury than a group of people with varied personal opinions that could be inconsistent and sway back and forth depending on issues."* In flat governance structures where jurors lack differentiated roles, prosocial behavior may be more difficult to cultivate or motivate [84]. Other jurors questioned the motives of those who would participate, with one suspicious of volunteers: *"I don't know if I trust a random collection of individuals, especially if they seek out the moderation duty..."*, while another was suspicious of paid jurors: "*If you have paid shills hanging out, waiting for jury duty, they can manipulate the results and bend results to suit their political agenda.*" One juror even doubted whether they were truly deliberating with other humans, potentially reflecting the greater proliferation of bots that further erode trust online. Researchers have proposed ways to increase trust in online communities, such as persistent identity, quality control, and coordination mechanisms [65] which could be incorporated as judgment records, juror accountability processes, and support communities for jurors. Support structures are explicitly mentioned as missing in LoL tribunals, and cited as factors that eroded player trust in the tribunal system [62].

Similarly, perceptions of what was "fair" seemed divided in two camps, depending on whether they considered receiving other people's perspectives as unwanted bias or as inclusion of multiple perspectives. In the bias camp, users described the IMMERSIVE round as unfair for allowing people to influence each other: "*I like [scaleable] because no one is influenced by others. People need to form their own opinion.*" Other users held the opposing view that more communication would be a fairer process. One user expressed that *"It would help to see other view points as some things are more offensive to others than to me."* Future work could explore designs that seek to reduce groupthink [83] and promote authentic consensus [80], partly alleviating the concerns of the first camp.

Feasibility

This "minimum viable jury" study raises several implementation considerations related to aspects of the jury model:

Juror selection bias (stage 1): In lieu of a mandatory system, many jurors expressed concerns about biased or bad actors exploiting the selection process. One juror wrote, *"Social media is entirely too biased and needs a democratic, unbiased panel to litigate these matters. My only concern is how biased the panelists are."* In this study, jurors were only allowed to participate once on the crowdsourcing platform, though future versions could track participation and voting records across cases. This process also did not enforce any qualifications beyond age and national locality, though the opportunity to prefilter users could address perceived expertise on the platform. Without a clearly identified juror community, users' distrust towards the jury selection process may also reflect the absence of a civics-oriented framework among platform users.

Jury incentivization (stage 2): Jury service could re-frame moderation work as an empowering act of civic duty rather than micro-tasks outsourced to contractors. We asked users in the survey under what conditions they would participate in a jury for a social media platform. With the ability to select multiple options, 23.3% said they would actively volunteer, while 36.7% of users would participate in their idle time. Research has shown in the offline case, however, that many jurors bear financial hardships due to jury duty [81]. Though there will likely be fewer hardships in the online case, financial compensation should still be provided, and time should be reduced without dropping below a minimum level of quality engagement and interaction. We saw that 82.2% would participate if paid for their time. A final question is whether to compel users via sanctions, such as by locking their account. However, only 10.0% said they would participate if it was mandatory, with the alternative presumably being that they would leave the platform. Only 3.3% said they would not participate at all.

Juror exposure to harm (stage 3): Much like the challenges facing commercial content moderators, exposure to traumatic material in a jury system process could potentially cause participants harm [56, 79]. The cases in this study were displayed in a way to intentionally minimize exposure to graphic or potentially traumatizing material. To ameliorate this in practice, a digital jury system could follow best practices from criminal court cases with sensitive material and rely on the existing ecosystem of platform policy teams to synthesize and prepare

cases. Exposure could also be rate-limited. Ethical research practices such as informed consent should be consulted and implemented. To protect against identification or retaliation, our implementation is pseudonymous, showing only a self-selected username but no other demographic information.

Synchronicity (stage 4): A challenge for this study was the difficulty in assembling an entire jury of distributed jurors synchronously during the consensus-forming stage. In the first few trials, the AMT task did not attract participants at the same time, causing some jurors to have to wait. Future systems could support automated scheduling and jury assignment.

LIMITATIONS AND FUTURE WORK

The use of AMT has downsides due to biases in its population demographics, though it was helpful for synchronicity. Future studies could collect larger and more representative samples and also broaden to non-U.S. and non-Western groups. In addition, all study participants were also jurors, and thus, our results may not reflect how a non-juror would perceive a case decided by jury. However, as digital juries for content moderation are rare in today's online platforms, we opted to have participants actually *experience* the workflows, as opposed to simply presenting scenarios, so that they could better understand and compare them. Future work could partner with real communities to deploy juries and survey both jurors and non-jurors. Studies could explore the side benefits of media literacy education for jurors and track user sentiment towards the challenges of content moderation and platform governance. Case transcripts could also be published to solicit the opinions of the broader community. In addition, as communities vary greatly, different groups could be involved in co-designing customized jury processes to fit their community's needs. Finally, this work focused on adherence to community standards but other tasks for juries could be tackling issues like identifying misinformation [82] or even going beyond simply interpreting policy to surfacing or guiding the creation of new policies.

Jury bias and diversity

The impacts of individual juror bias and collective group polarization on democratic deliberation have been explored at length [85]. However, sampling from a heterogeneous, diverse population has been shown to mitigate these negative effects, such as in deliberative polling [34, 85]. Future work on digital juries could gather randomized samples of multiple jury profiles to evaluate the same case, in order to study the polarization or alignment of online groups. Jury outcomes could also be compared against expert opinions as an accuracy check to measure alignment between experts and the "majoritarian view" [11]. While jury diversity is effective in reducing polarization, studies have shown that majority voices can obscure minority racial and gender perspectives during deliberation [3]. Due to the limited sample size and pseudonymous jury design used in this study, we could not analyze this impact on minorities. As digital juries have the benefit of targeted recruiting and pre-screening participants, future studies with larger populations should consider varying juror recruiting strategies as well as analyzing the effects of different forms of identity presentation on deliberation.

Local juries and global impacts

Social norms and rules vary dramatically across not only macro-level platforms, but also within meso-level communities [20, 6, 1, 7]. Depending on the platform, a "local" jury could be measured by geographic distance, nationality, servers, group membership, social network distance, or more. Future studies could vary jury sizes, either as fixed sizes (e.g., 6-person vs. 12-person juries) or as algorithmically-determined sizes depending on the case, such as in LoL. Future studies could also experiment with sampling multiple rounds of juries and different decision rules. Our current study uses simple majority rules, but prior research in deliberation has shown the positive civic impacts of unanimous decision rules, despite the longer required time investment [46].

This flexibility also lends the possibility of *localized* interpretations of policies, as opposed to having to a single, consistent interpretation across a platform. Such is the case for platforms like Reddit with many subreddit-specific rules [1, 35]. Particularly on platforms where users do not share the Western legal tradition of juries, juries could be adapted to local contexts, such as using juries only to supplement expert opinion. Another consideration is local *enforcement* of policies, such as removing content only in specific parts of the platform. The implementation of this depends greatly on the platform's architecture and content distribution method.

CONCLUSION

In this work, we propose *digital juries* as a civics-oriented, decision-making approach for adjudicating online content moderation questions at scale. Building on existing models of jury decision-making, we present the Digital Jury Model, characterizing the space of design considerations when developing a digital jury process. In our empirical analysis of prototype jury workflows, we find evidence that digital juries improve user perceptions of procedural justice in the content moderation process on all measured attributes, with the exception of efficiency. Specifically, the jury processes are perceived as a more legitimate exercise of platform power, improving trust in how content moderation processes are made, valuing individual voices, and caring about user preferences.

While digital juries may have potential drawbacks in efficiency and trustworthiness, they can be a valuable participatory mechanism that improves perceptions of the democratic legitimacy of platform governance and encourages a civics-oriented social identity. The design dimensions we outline in our 5-stage model as well as our empirical results comparing two instances from the model point to a host of potential future dimensions to explore. Beyond design, there is a rich space to investigate the use of digital jury systems in real communities as part of the growing field of internet governance.

ACKNOWLEDGMENTS

We are grateful to the Harvard Design Engineering program, particularly the directors, advisors Krzysztof Gajos and Robert Pietrusko, and MDE '19 cohort. Thank you to Claire Wardle and Alexios Mantzarlis for their support, our friends at MIT CSAIL and Berkman Klein Center for their valuable insight, and our reviewers, many proofreaders, and study participants.

REFERENCES

[1] Kimberley R. Allison, Kay Bussey, and Naomi Sweller. 2019. 'I'm Going to Hell for Laughing at This': Norms, Humour, and the Neutralisation of Aggression in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 152 (Nov. 2019), 25 pages. DOI: http://dx.doi.org/10.1145/3359254

[2] Jack M. Balkin. 2004. Virtual liberty: Freedom to design and freedom to play in virtual worlds. *Virginia Law Review* 90, 8 (2004), 2043–2098.

[3] Robert V. Bartlett and Walter F. Baber. 2015. The Citizen Jury as a Deliberative Forum: Juries as Instruments of Democracy. In *Consensus and Global Environmental Governance: Deliberative Democracy in Nature's Regime*. The MIT Press, Cambridge, Massachusetts; London, England.

[4] Susan Benesch. 2013. Dangerous Speech: A Proposal to Prevent Group Violence. (Feb 2013). https://perma.cc/K97E-9U78

[5] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*. Springer, Springer, New York, 405–415.

[6] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages. DOI: http://dx.doi.org/10.1145/3134659

[7] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 100 (Nov. 2019), 25 pages. DOI: http://dx.doi.org/10.1145/3359202

[8] Andrew J. Bloeser, Carl Mccurley, and Jeffery J. Mondak. 2012. Jury Service as Civic Engagement: Determinants of Jury Summons Compliance. *American Politics Research* 40, 2 (2012), 179–204.

[9] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012).

[10] Nathan Bos, Judy Olson, Darren Gergle, Gary Olson, and Zach Wright. 2002. Effects of Four Computer-mediated Communications Channels on Trust Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 135–140. DOI: http://dx.doi.org/10.1145/503376.503401

[11] Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. *Report of the Facebook Data Transparency Advisory Group*. The Justice Collaboratory, New Haven. https://perma.cc/4W9X-WJWA

[12] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7313–7318. DOI: http://dx.doi.org/10.1073/pnas.1618923114

[13] Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. 1994. Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion*. ACM, New York, 183–184.

[14] Moira Burke and Robert Kraut. 2008. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 27–36.

[15] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1101–1110.

[16] Justin Caffier. 2017. Get to Know the Memes of the Alt-Right and Never Miss a Dog-Whistle Again. (Jan. 2017). https://perma.cc/EM8H-AYVL

[17] Robyn Caplan. 2018. Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches. (Nov. 2018). https://perma.cc/W6D3-HMPE

[18] Michael X. Carpini, Fay L. Cook, and Lawrence R. Jacobs. 2004. Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature. *Annual Review Of Political Science* 7, 1 (2004), 315–344.

[19] Anupam Chander and Vivek Krishnamurthy. 2018. The Myth of Platform Neutrality. (Symposium Issue: Information Platforms and the Law). *The Georgetown Law Technology Review* 2, 2 (2018).

[20] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[21] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[22] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[23] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. 2018. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. (2018).

[24] Jennifer Cobbe. 2019. Algorithmic Censorship on Social Platforms: Power, Legitimacy, and Resistance. *Legitimacy, and Resistance* (Aug. 2019).

[25] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[26] Richard L. Daft and Robert H. Lengel. 1986. Organizational Information Requirements, Media Richness and Structural Design. *Management Science* 32, 5 (1986), 554–571.

[27] Norman Dalkey and Olaf Helmer. 1963. An experimental application of the Delphi method to the use of experts. *Management science* 9, 3 (1963), 458–467.

[28] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on web search and data mining (WSDM '18)*, Vol. 2018-. ACM, 135–143.

[29] Joan Donovan. 2019. How Hate Groups' Secret Sound System Works. (Mar 2019). https://perma.cc/9L5J-ZLAC

[30] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. In *HCOMP*.

[31] Kristen Eichensehr. 2019. Digital Switzerlands. *University of Pennsylvania Law Review* 167 (2019), 665–732. Issue 3.

[32] Joshua M. Epstein. 2013. *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*. Princeton University Press, Princeton, New Jersey ; Oxfordshire, England.

[33] Cynthia Farina, Hoi Kong, Cheryl Blake, Mary Newhart, and Nik Luka. 2014. Democratic Deliberation in the Wild: the McGill Online Design Studio and the RegulationRoom Project. *Fordham Urban Law Journal* 41 (2014), 1527–1759.

[34] Cynthia Farrar, James S. Fishkin, Donald P. Green, Christian List, Robert C. Luskin, and Elizabeth Levy Paluck. 2010. Disaggregating Deliberation's Effects: An Experiment within a Deliberative Poll. *British Journal of Political Science* 40, 2 (2010), 333–347.

[35] 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*.

[36] James Fishkin, Nikhil Gargl, Lodewijk Gelauffl, Ashish Goell, Kamesh Munagala, Sukolsak Sakshuwong, Alice Siu, and Sravya Yandamuri. 2019. Deliberative Democracy with the Online Deliberation Platform. (2019). https://perma.cc/C53W-PA5L

[37] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.

[38] Seth Frey, P. M. Krafft, and Brian Keegan. 2019. "This Place Does What It Was Built For": Designing Digital Institutions for Participatory Change. *Proceedings of the ACM on Human-Computer Interaction* 3 (Nov. 2019), 1–31. DOI:http://dx.doi.org/10.1145/3359134

[39] Seth Frey and Robert W. Sumner. 2019. Emergence of integrated institutions in a large population of self-governing communities. *PloS One* 14, 7 (2019), e0216335.

[40] John Gastil. 2010. *The Jury and Democracy: How Jury Deliberation Promotes Civic Engagement and Political Participation*. Oxford University Press, Oxford ; New York.

[41] Tarleton Gillespie. 2015. Platforms intervene. *Social Media + Society* 1, 1 (2015), 2056305115580479.

[42] Tarleton Gillespie. 2018. *Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven.

[43] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean Wojcik, and Peter Ditto. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances In Experimental Social Psychology, Vol 47* 47 (2013), 55–130.

[44] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17 (2015), 42–368.

[45] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia@MISCs reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[46] Valerie P. Hans, John Gastil, and Traci Feller. 2014. Deliberative Democracy and the American Civil Jury. 11 (2014), 697–894.

[47] Stephan Hartmann and Soroush Rafiee Rad. 2018. Voting, deliberation and truth. *Synthese* 195, 3 (2018), 1273–1293.

[48] Nigel Harvey. 2008. *Blackwell Handbook of Judgment and Decision Making*. Wiley, Hoboken.

[49] Reid Hastie. 1983. *Inside the Jury*. Harvard University Press, Cambridge, Mass.

[50] Margaret Levi Henry Farrell and Tim O'Reilly. 2018. Mark Zuckerberg runs a nation-state, and he's the king. (April 2018). https://perma.cc/WE8M-2SDT

[51] Mitchel N. Herian, Joseph A. Hamm, Alan J. Tomkins, and Lisa M. Pytlik Zillig. 2012. Public Participation, Procedural Fairness, and Evaluations of Local Governance: The Moderating Role of Uncertainty. *Journal of Public Administration Research and Theory* 22, 4 (2012), 815–840.

[52] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *CoRR* abs/1702.08138 (2017). http://arxiv.org/abs/1702.08138

[53] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 74.

[54] Irving L Janis. 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. (1972).

[55] Jigsaw. 2019. Perspective API. (2019). https://perspectiveapi.com/

[56] Stanley M. Kaplan and Carolyn Winget. 1992. The occupational hazards of jury duty. *The Bulletin of the American Academy of Psychiatry and the Law* 20, 3 (1992).

[57] David Kaye. 2019. Speech Police: The Global Struggle to Govern the Internet. (2019).

[58] Sara Kiesler and Lee Sproull. 1992. Group decision making and communication technology. *Organizational Behavior and Human Decision Processes* 52, 1 (1992), 96–123.

[59] Kate Klonick. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, 6 (2018), 598–670.

[60] Jason Koebler and Joseph Cox. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. (aug 2018). https://perma.cc/PZ7W-PDYY

[61] Yubo Kou and Xinning Gui. 2017. When Code Governs Community. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*. 1–9.

[62] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior Through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 62 (Dec. 2017), 17 pages. DOI: http://dx.doi.org/10.1145/3134697

[63] Yubo Kou and Bonnie Nardi. 2013. Regulating Anti-Social Behavior on the Internet: The Example of League of Legends. (Feb. 12-15 2013), 616–622.

[64] Yubo Kou and Bonnie Nardi. 2014. Governance in League of Legends: A Hybrid System. *Foundations of Digital Games* (April 3-7 2014).

[65] Robert E. Kraut. 2011. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, Cambridge, Mass.

[66] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with Considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 265–274. DOI: http://dx.doi.org/10.1145/2145204.2145249

[67] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3739–3748.

[68] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, 543–550.

[69] Melvin J. Lerner, E. Allan Lind, and Tom R. Tyler. 1988. *Social Psychology of Procedural Justice*. Springer, Boston.

[70] Lawrence Lessig. 2006. *Code version 2.0* ([2nd ed.]. ed.). Basic Books, New York.

[71] Natasha Lomas. 2015. Can Civil Comments Kill the Internet Troll? (Oct. 2015). https://perma.cc/Z5XH-C4KP

[72] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (2019). https://doi.org/10.1177/2056305119836778

[73] Peter Muhlberger. 2005. The Virtual Agora Project: A Research Design for Studying Democratic Deliberation. *Journal of Public Deliberation* 1, 1 (2005).

[74] 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.

[75] Elinor Ostrom. 1990. *Governing the commons : the evolution of institutions for collective action*. Cambridge University Press, Cambridge ; New York.

[76] David G. Post. 1995. Anarchy, State, and the Internet: An Essay on Law-Making in Cyberspace. *Journal of Online Law* 1995 (1995), 3–5.

[77] Paul Resnick. 2001. Beyond bowling together: Sociotechnical capital. *HCI in the New Millennium* 77 (2001), 247–272.

[78] Sarah T. Roberts. 2016. Commercial content moderation: digital laborers' dirty work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*, Safiya Umoja Noble and Brendesha M. Tynes (Eds.). Peter Lang Publishing, New York, Chapter 8.

[79] Sarah T. Roberts. 2019. *Behind the screen : content moderation in the shadows of social media*. Yale University Press, New Haven.

[80] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[81] Richard Seltzer. 1999. The Vanishing Juror: Why Are There Not Enough Available Jurors? *The Justice System Journal* 20, 3 (1999), 203–218.

[82] Henry Silverman. 2019. Helping fact-checkers identify false claims faster. (Dec. 2019). https://perma.cc/S7F6-WUV4

[83] Miriam Solomon. 2006. Groupthink versus the wisdom of crowds: The social epistemology of deliberation and dissent. *The Southern Journal of Philosophy* 44, S1 (2006), 28–42.

[84] Lee Sproull. 2011. Prosocial Behavior on the Net. *Daedalus* 140, 4 (2011), 140–153.

[85] Cass R. Sunstein. 2002. The Law of Group Polarization. *Journal of Political Philosophy* 10, 2 (2002), 175–195.

[86] Nicolas Suzor. 2018. Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society* 4, 3 (2018), 2056305118787812.

[87] Nabiha Syed and Ben Smith. 2016. A First Amendment For Social Platforms. (Jun 2016). https://perma.cc/THC3-UNQ4

[88] Lynne Tirrell. 2018. Toxic Speech: Inoculations and Antidotes. *Southern Journal of Philosophy* 56, S1 (2018), 116–144.

[89] Tom Tyler. 1988. What Is Procedural Justice? Criteria Used by Citizens to Assess the Fairness of Legal Procedures. *Law and Society Review* 22, 1 (1988), 103–103.

[90] Tom R. Tyler. 1989. The Psychology of Procedural Justice: A Test of the Group-Value Model. *Journal of Personality and Social Psychology* 57, 5 (1989), 830–838.

[91] Ted Ullyot. 2009. Results of the Inaugural Facebook Site Governance Vote. (Apr 2009). https://web.archive.org/web/20090430215524/http://blog.facebook.com/blog.php?post=79146552130

[92] Claire Wardle and Hossein Derakhshan. 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report* 27 (2017).

[93] Morten Warncke-Wang, Vladislav R. Ayukaev, Brent Hecht, and Loren G. Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 2015 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 743–756.

[94] Ryan Winter and Edie Greene. 2008. *Juror Decision-Making*. Wiley Online Library, 739 – 761. DOI:http://dx.doi.org/10.1002/9780470713181.ch28

[95] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. *CoRR* abs/1805.12512 (2018). http://arxiv.org/abs/1805.12512

[96] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, 2082–2096.

[97] Mark Zuckerberg. 2018. A Blueprint for Content Governance and Enforcement. (Nov 2018). https://perma.cc/TC7X-YUXF