
CHRISTINA A. PAN∗, Stanford University, USA
SAHIL YAKHMI∗, Stanford University, USA
TARA P. IYER∗, Stanford University, USA
EVAN STRASNICK, Stanford University, USA
AMY X. ZHANG, University of Washington, USA
MICHAEL S. BERNSTEIN, Stanford University, USA

While research continues to investigate and improve the accuracy, fairness, and normative appropriateness of content moderation processes on large social media platforms, even the best process cannot be effective if users reject its authority as illegitimate. We present a survey experiment comparing the perceived institutional legitimacy of four popular content moderation processes. We conducted a within-subjects experiment in which we showed US Facebook users moderation decisions and randomized the description of whether those decisions were made by paid contractors, algorithms, expert panels, or juries of users. Prior work suggests that juries will have the highest perceived legitimacy due to the benefits of judicial independence and democratic representation. However, expert panels had greater perceived legitimacy than algorithms or juries. Moreover, outcome alignment—agreement with the decision—played a larger role than process in determining perceived legitimacy. These results suggest benefits to incorporating expert oversight in content moderation and underscore that any process will face legitimacy challenges derived from disagreement about outcomes.

CCS Concepts:
• Human-centered computing → Empirical studies in collaborative and social computing.

Additional Key Words and Phrases: content moderation; platform governance; legitimacy; social media

ACM Reference Format:

1 INTRODUCTION

Efforts to improve platform design are ineffective if users do not trust platforms and their processes. Large social media platforms—like Facebook, YouTube, and Twitter—have become the new “town
squares” for public discourse [88], but the legitimacy of the rules and processes governing these platforms have increasingly been called into question. Social media platforms face widespread criticism for regulating speech in an opaque [116] and unrepresentative [77] manner without meaningful oversight [77].

Since online platforms have become the “new governors” of speech [83], they have been analyzed through the lens of political theory and legitimacy [16, 48, 83, 100, 131, 149, 151]. Moreover, a large body of sociological work points to the practical importance of perceived legitimacy—the acceptance of authority by those subject to it—for the functioning of institutions [21, 51, 69, 104, 139]. Empirical studies show that when institutions are perceived as highly legitimate by the public, this results in greater acceptance of unpopular decisions along with more cooperation and compliance with authorities in the long term. For instance, when the US Supreme Court decided a contentious election in Bush v. Gore, compliance was swift and the standing of the court was not measurably diminished [53]. As institutions that regularly must make contentious decisions, online platforms similarly depend upon perceived legitimacy.

In this paper, we compare the perceived legitimacy of several content moderation processes that are in wide use or are specifically designed to increase legitimacy of moderation decisions. Understanding the impact of different content moderation processes on perceived legitimacy is critical—crafting even a “perfect” moderation process will not help a platform if that process is viewed by the population as illegitimate. In recent years, scholars have applied the lens of legitimacy to online platforms, including surveying the governance mechanisms they use [34, 59], proposing frameworks with which to evaluate platform legitimacy [129], and proposing more legitimate methods of platform governance [2, 36, 44, 78, 135]. However, most prior work lacks robust, empirical methods of evaluating legitimacy, and existing empirical work does not establish a common basis for comparing disparate processes. Stakeholders seeking to design more legitimate content moderation processes, whether platform owners, academics, or policymakers, currently lack data on how specific processes and proposals affect perceived legitimacy and the extent to which process design matters at all when making decisions about highly disagreed-upon content.

We conducted an online, within-subjects survey experiment in which US Facebook users evaluated moderation decisions presented as made by one of four processes: paid contractors, algorithms, expert panels, and juries of users. Paid contractors and algorithms are the two common types of content moderation used at scale [56], while expert panels like the Facebook Oversight Board [150] and digital juries [44] are both recent moderation processes gathering substantial support and debate that are designed to enhance legitimacy.

In our within-subjects survey experiment, for each moderation process, participants were given a randomly selected Facebook post along with a randomly assigned decision outcome. For each post, participants were asked to answer questions about their attitudes towards the post and decision outcome, which measure components of perceived institutional legitimacy. At the end, participants were asked to compare and discuss the four processes. From participants’ responses to the individual posts, we constructed a model that estimates the effect on perceived legitimacy of each moderation process, user alignment with the decision, and demographic variables. In addition to the quantitative analysis, we coded the comparative responses to identify and analyze all meaningfully distinguishable attitudes.

We find that expert panels have greater perceived legitimacy than both algorithms and digital juries. These results suggest that users value expertise, even when the nature of that expertise is not well understood. Additionally, we find qualitative evidence of a user preference for group decision making over decisions made by individuals and of acceptance of algorithmic decisions being conditional on factors like human oversight, despite being perceived as impartial. However, we also find that the alignment of user preferences with decision outcomes dominates all tested
process factors in determining perceptions of legitimacy. In other words, whether users agree with the decisions of the content moderation process has a greater impact on the legitimacy users attach to that process than the process itself. While these results suggest platforms may struggle to create processes that can be perceived as legitimate by all users when dealing with highly disagreed-upon content, they also suggest incorporating expert oversight and multiple perspectives into moderation processes can help.

2 BACKGROUND AND RESEARCH QUESTIONS

In this section, we draw from prior work on platform governance, content moderation, and political legitimacy to motivate our study and methods. In contrast to the body of normative and qualitative work on content moderation, this paper contributes an empirical study that allows for user perceptions of multiple moderation processes to be compared with common methods of evaluating perceived legitimacy.

2.1 Content Moderation Processes

Each day, users post billions of pieces of content on online platforms [56]. This content must be reviewed so that illegal and harmful posts can be removed in a timely manner, while nevertheless respecting the users’ right to self-expression [56, 92]. In this work, we consider the case of post hoc content moderation takedown decisions, excluding processes involved in crafting content policy, to narrow the focus of the study to an intervention that is easily understood by participants and can be undertaken by several processes.

Online platforms employ varied methods to carry out this task. Consequently, researchers have sought to identify patterns in these strategies, for example, distinguishing between artisanal, community-reliant, and industrial moderation processes [22, 56]. Artisanal and community-reliant processes have been used by small platforms and niche communities within larger platforms like Reddit [56, 121]. However, the largest platforms heavily rely upon industrial moderation processes—defined as processes that enable platforms to 1) operate at large scale, 2) enforce well-defined rules, and 3) maintain separation between a) policy creation and b) interpretation and enforcement [22]. Industrial moderation processes heavily overlap with commercial content moderation processes [60, 116].

In this work, we limit our scope to investigating the legitimacy of industrial moderation processes because they impact the most people, being employed by the largest platforms (Facebook, Youtube, Twitter, etc.), and are the processes most central to ongoing public debate over content moderation [22]. As a result, we do not investigate community or artisanal moderation. Although community moderation is also employed by some large platforms, it is more closely linked to subcommunity norms rather than platform-wide rules, and generally does not strictly separate policy creation and enforcement. Community moderation, therefore, should be studied in context of specific community norms and not only from the perspective of post hoc decisions.

We select the four following industrial moderation processes for our study.

1. Paid Individual Contractors, who are hired and trained on a company’s moderation policy
2. Automated Systems, commonly powered by databases of known infringing content and machine learning algorithms, trained with the help of human contractors, that detect certain types of banned content (e.g., explicit language, hate speech, and pornographic images)
3. Digital Juries, or ad-hoc deliberative bodies drawn from the user population
4. Expert Panels, composed of experts in content moderation and related fields like law, human and digital rights, media and journalism, and political science
Paid Individual Contractors (1) and Automated Systems (2) are selected because they are the industrial processes in widespread use by large platforms, and have been extensively described by many authors (e.g., [22, 56, 60, 116]).

Digital Juries (3) and Expert Panels (4) are selected because they are emerging processes that also fall under the definition of industrial moderation. Both processes are well examined in the literature [44, 115, 128], are used in some form in industry [7, 43, 85], and were proposed specifically to help address the legitimacy issues plaguing earlier methods [25, 115, 129, 141]. Digital Juries, as described by Fan and Zhang [44], draw legitimacy from democratic norms [25] and use of authentic deliberation [44]. In industry, Digital Juries resemble juries as used on platforms such as League of Legends, Weibo, and even Parler [7, 85]. Expert Panels are representative of bodies of experts like Facebook’s fact-checking program using 3rd party fact-checkers [43] or the Facebook Oversight Board [150], which are intended to be transparent and independent. Because they are an emerging process, platforms are still developing the design of expert panels such as the Facebook Oversight Board, including how they can potentially scale and whether they should conduct policy creation separate from interpretation. However, we chose to include expert panels as a counterpoint to digital juries and limit it to be a process for policy interpretation in line with our other three processes.

2.2 Legitimacy

2.2.1 What is Legitimacy? Legitimacy can be understood on either a normative or descriptive basis [72, 113]. In its normative sense, legitimacy “refers to some benchmark of acceptability or justification of political power or authority and—possibly—obligation” [113]. Influential examples of normative legitimacy frameworks include constitutional legitimacy [117] and democratic legitimacy [112], discussed later in the context of emerging content moderation processes. By contrast, in its descriptive sense, legitimacy refers to the acceptance of authority [113, 145].

This study examines legitimacy in its descriptive sense, i.e., as a measurable, subjective, sociological phenomenon, referred to using the more common phrase perceived legitimacy [137]. However, while we emphasize perceived legitimacy for its established practical benefits, we recognize the role that normative principles like fairness play in shaping attitudes [72]. As such, we will discuss prior work examining both conceptions of legitimacy, using legitimacy to refer to the expansive concept in both its normative and descriptive senses. We use the term democratic legitimacy to refer to the normative concept that political systems derive legitimacy through adherence to democratic norms, procedures, and values [25].

2.2.2 Measuring Perceived Legitimacy. Modern social scientists have contributed a wealth of work on measuring the perceived legitimacy of governance. We draw primarily from work studying the perceived legitimacy of the courts. Among the best established of this work is that of law and psychology professor Tom Tyler and political scientist James Gibson. In Tyler’s framework, fair procedure, quality of decision making, quality of treatment, and motive-based trust contribute to greater perceived legitimacy, while perceived legitimacy in turn fosters compliance, cooperation, and empowerment [136]. Tyler also highlights that, to a plurality, perceived legitimacy is analogous to obtaining the person’s desired outcome [138], implying that there is a limited extent to which process design can create perceived legitimacy at all. While Tyler examines individuals’ interactions with the state, Gibson instead frames perceived legitimacy around institutions. Gibson measures the legitimacy of institutions like courts through procedural values like trustworthiness and neutrality. In addition, he measures institutional commitment, the extent to which people support an institution’s existence, and decisional jurisdiction, the support for the institution’s power over a particular application [55]. Gibson places special emphasis on “diffuse support,” a “reservoir of
favorable attitudes or good will that helps members to accept or tolerate outputs to which they are opposed or the effects of which they see as damaging to their wants,” as opposed to “specific support,” or support for a particular action or policy [21, 39].

In the domain of content moderation, however, little work exists that measures perceived legitimacy. Instead, most prior work investigates questions of normative legitimacy, for example outlining fundamental rights and procedural values known to correspond to legitimate governance [129, 130]. Of the studies that take a more descriptive and empirical approach, none have tackled the question of perceived legitimacy head on, focusing instead on adjacent questions [44, 119, 140].

In the absence of an established measure of perceived legitimacy of content moderation, we select Gibson’s formulation of institutional legitimacy as our overarching framework for measuring perceived legitimacy. In addition, we follow Gibson in using population-wide measures of attitudes to capture “diffuse support.”

2.3 Studying the Legitimacy of Content Moderation Processes

From prior work, we can conclude that while legitimacy is broadly accepted as an important and desirable quality in content moderation systems and a variety of perspectives exist regarding how it can be accrued, the impact of specific processes on perceived legitimacy remains largely unknown.

This missing data in the literature motivates the primary research question of our study:

RQ1: How do content moderation processes impact levels of perceived institutional legitimacy?

For those seeking to design legitimate content moderation systems, a major open question is the extent to which process design can create perceived legitimacy at all. Because a legitimate process is most valuable when it can mitigate the negative effects of an unfavorable decision, it is important to contextualize the magnitude of process effects by comparing them with the strength of outcome effects. Thus we also ask:

RQ2: To what extent does the alignment of outcome of decisions made by a process with individual preferences influence the perceived institutional legitimacy of that process?

2.4 Known Determiners of Legitimacy

Academics have proposed a plethora of principles and frameworks that contribute to legitimate governance, including transparency and public participation [49, 62, 68, 101, 130], adherence to established legal principles [12, 83], and upholding individual rights [26, 76]. However, to develop hypotheses for RQ1, we focus our discussion on prior work examining the factors that differ between the four processes.

2.4.1 Independence. Among a large space of dispute resolution processes, prior work shows a consistent preference among litigants for greater decision control by an impartial third-party [70, 123]. National high courts enjoy special legitimacy [51], and it is often taken for granted that their greater independence contributes to public trust [93]. Multiple empirical studies also find support for independence conferring greater perceived legitimacy to political institutions and courts [20, 54]. Consequently, in the domain of content moderation, decision-making by independent bodies commonly features in high level frameworks designed to enhance legitimacy, such as FAITE [135],
national social media councils (SMCs) [78], and policy proposals by the Cato Institute and the Bookings Institute [102, 118].

Among the four moderation processes in this study, we consider the digital jury to have high independence by analogy with criminal juries, which serve as an independent check on government power [126]. Additionally, we expect jury members’ loose and impersonal relationship with the platform would limit the platform’s influence over their decisions. We accept the judgment of prior work that expert panels can benefit from independence [122], but note that potential platform influence over the body and its composition may limit practical independence. Conversely, we deem human contractors and algorithmic moderation to have no practical independence from the platform.

Due to the greater independence of the deliberative bodies, we hypothesize:

H1.1: The expert panel and digital jury will be perceived as more legitimate than the paid contractor and algorithm.

2.4.2 Automated Decision Making. As the use of Algorithmic Decision Making (ADS) has grown, researchers have studied its characteristics relative to human decision making from multiple perspectives [23]. ADS is often evaluated according to to specific normative criteria, including fairness [14, 71], accountability [147], explainability [8, 103], and contestability [67, 140]. Such inquiry is motivated in part by evidence that ADS can be biased and can cause various types of harms [13, 15, 31, 109], and in part by application of theories of justice [18, 75, 98]. Despite major theoretical and practical issues commonly known in academia, prior work shows that perception of the trustworthiness of algorithms relative to humans can be favorable, though it is highly dependent on context, subjectivity of the domain, and performance [11, 89, 95]. Moreover, public perceptions of algorithms are subject to cognitive biases, including overconfidence in their capabilities [143], outcome favorability bias [41, 144], excessive aversion to mistakes [35], and folk theories [40]. In general, algorithms tend to benefit from being perceived as impartial, objective, and authoritative by the public [57, 127].

In this study, algorithmic moderation embodies automated judgment, standing in contrast to deliberative bodies like the expert panel and digital jury. While paid contractors can exercise human judgment, we assess that because paid contractors are given extremely limited time and detailed guidelines to make decisions [56], their decisions involve significantly less discretion.

For highly disagreed-upon posts, we anticipate that perceptions of algorithms’ impartiality will outweigh concerns about their lack of ability in a subjective domain and the larger penalties they receive for poor performance. Thus, making a direct comparison between the automated and human processes with low independence, we hypothesize:

H1.2: The algorithm will be perceived as more legitimate than the paid contractor.

2.4.3 Democratic Legitimacy vs. Expertise. The debate over judge vs. jury trials in the judicial system can be understood as a debate over democratic vs. expert authority. Juries, despite well known drawbacks [82], have long been justified on the grounds that they bind the legal system to community norms and provide legitimacy through democratic representation and the exercise of popular sovereignty [45, 120, 126]. Surveys in the US have consistently found broad public support for juries as an institution (“diffuse support”) [65]. However, little rigorous empirical work exists
that compares the perceived legitimacy of juries to judges \[82, 126\]. Nevertheless, surveys tend to show a preference for juries over judges \[126\] and some empirical evidence of perceived legitimacy benefits of citizen participation have been noted in multiple countries \[17, 99\].

Prior work on the perceived legitimacy of expert authority is mixed. On the one hand, public trust in experts appears pervasive \[50, 133\], and expertise has traditionally been seen as a way to establish legitimate authority \[47, 146\]. However, critiques of expert authority are common, with scholars pointing to unequal relationships between experts and the public and other issues \[46, 64, 134\]. Steven Turner resolves this tension by noting that claims to cognitive authority must be legitimated through acceptance by the public, observing that across fields experts achieve varying levels of success \[134\].

In this study, the expert panel embodies expert knowledge and judgment, while the digital jury represents democratic participation. Contractors and algorithms, though they may act in accordance with expert-designed guidelines, do not exercise sufficient individual discretion to represent either type of knowledge.

Prior work generally supports the idea that juries are perceived as more legitimate decision makers, and suggests that the benefits of democratic legitimacy extend to the domain of content moderation \[44\]. Moreover, because the domain of content moderation is relatively novel, it is reasonable to expect that the legitimation process of experts among the public—which can lag legitimation among professionals by decades \[134\]—is still in its infancy. Consequently, we hypothesize:

**H1.3:** The digital jury will be perceived as more legitimate than the expert panel.

In the following subsection, we review prior work relating to RQ2.

### 2.4.4 Role of Pre-Existing Views

While much work on perceived legitimacy is concerned with its ability to promote the acceptance of adverse or unpopular outcomes, prior work suggests that perceived legitimacy is itself shaped by alignment with individual preferences and beliefs. A large body of work finds evidence for various confirmation or congeniality biases \[19, 81\] whereby pre-existing views affect how information is collected \[80\], interpreted \[96\], and evaluated \[38\]. Similarly, motivated reasoning theory describes mechanisms by which directional goals bias cognitive processes \[87\]. These biases can be strongly mediated by partisan identification and cues \[58, 91\]. There is some indication that these cognitive effects may extend to perception of legitimacy. While Gibson finds that controversial decisions do not necessarily impair the legitimacy of an institution like the Supreme Court \[53\], other work finds that strongly held moral convictions do magnify the effect of outcomes on perceived legitimacy of the court \[125\].

Based on this prior work, we hypothesize:

**H2:** Users will report higher perceived legitimacy when content moderation systems make decisions that align with their individual normative preferences about whether a post should stay up or be taken down, and this association will be at least as strong as that of process factors.
3 METHODS
This study records and analyzes how US Facebook users perceive the institutional legitimacy of various content moderation processes in an online survey setting. In contrast to prior work [73, 119], we examine the attitudes of users who are not directly involved in content takedown decisions (i.e., bystanders). The literature on legitimacy indicates that public attitudes (i.e., “diffuse support”) are what determine the legitimacy of institutions [21]. Moreover, in online communities, the vast majority of users are never involved in content moderation disputes [108]. We designed the survey around Facebook due to the platform’s scale [28], broad adoption in the USA [27], and representative user base [111], and we recruited participants from Amazon Mechanical Turk (AMT), following a common practice in political science studies [30].

3.1 Materials
We followed the example set by prior work in collecting real social media posts rather than synthesizing controlled examples [44]. This approach mitigates potential biases in post creation and improves the ecological validity of the study, as prior work suggests that hypothetical choices can differ from choices made in concrete situations [86]. To reduce the impact of biases in post selection, we employed a two-stage strategy, described below.

We first compiled a list of Facebook posts representing a wide array of topics common in takedown decisions (e.g., racism, protest, vaccination, electoral fraud, government conspiracy) and viewpoints (e.g., both liberal and conservative), taking care to avoid specific posts that participants were likely to have already encountered in the media. We collected posts that might be viewed as violating Facebook’s Community Standards [3] in three of its categories: inciting violence, hate speech, and misinformation. These categories were chosen for their high frequency and prominence in public disagreements about content moderation. We collected 58 candidate posts from three sources: public Facebook groups, low-traffic news articles, and the Plain View Project (PVP) [1]. The PVP is a journalistic database of Facebook posts authored by police officers expressing themes of violence, racism, and bigotry. To find Facebook groups and news articles, we identified common topics within each category that elicited public disagreement (e.g., anti-vaccination in misinformation). We then used these topics as search terms on Facebook to find groups and on news search engines to find articles containing posts. The resulting posts may or may not have actually been removed by Facebook.

To further mitigate bias, we narrowed this broad pool of candidates to nine posts (i.e., 3 in each category of potential infringement) by selecting the posts that were the most disagreed-upon by Facebook user participants on AMT. We ran a pre-survey that asked 56 participants (not eligible for the main survey) their opinion about whether a given post ought to be removed (i.e., normative preference). Responses were recorded on a five-point Likert scale, ranging from strongly disagree (1) to strongly agree (5). For each candidate post, we calculated its disagreement score as a combination of the standard deviation of the responses and the absolute deviation of the median from the neutral response value of 3: disagreement\_post = σ\_post − |3 − μ\_post|. This formulation was chosen to ensure that posts would not only elicit a wide spread of opinions, but these opinions would be well balanced between favoring taking down and leaving up.\(^1\) The median standard deviation of

\(^1\)After the study was conducted, we discovered that we mistakenly included a post (Post 5 in the Supplementary Materials) that was not among the top 3 posts by disagreement score in its category—inciting violence. Although the study was designed around highly disagreed-upon content to enhance our ability to measure effects of process on perceived legitimacy, content moderation processes also deal with content for which opinions are more homogeneous. After performing additional analysis on a dataset that excluded the post in question, we found no meaningful change in the magnitude or direction of effects but observed higher p-values due to the loss of about 11% of data.
responses for the final nine posts was 1.5, and the median of the posts’ median response was 3.0. These posts are available in the Supplementary Materials.

3.2 Experimental Design

We constructed a within-subjects survey experiment to assess the perceived institutional legitimacy of content moderation decisions made by a paid contractor, an algorithm, an expert panel, and a digital jury. Participants were given 4 randomly constructed content moderation decisions—this randomization exposed participants to many many combinations of posts and processes to help mitigate biases introduced by individual posts. For each decision, participants were asked to answer several questions regarding their attitudes toward the post and the decision outcome. At the end, participants were asked to discuss the four processes on a comparative basis. The study design was reviewed and approved by our institution’s institutional review board (IRB) under protocol #57848. Selected screenshots of the survey are available in supplementary materials.

3.2.1 Participants. The survey was sent to US Facebook users on AMT. AMT allows only workers 18 years or older, and gives workers the option to self-report being Facebook users. Participants were required to go through an IRB-approved consent process with appropriate content warnings and resources. Participants were informed that neither the moderation decisions nor processes were real only in the survey debrief, to improve the survey realism. Participants were compensated $1.82 for the 15 minutes spent completing the survey, based on the 2020 federal minimum wage of $7.25/hr [110], a rate above the mean and median hourly wages for AMT workers ($3.13/hr and $1.77/hr, respectively) [66].

We set a target sample size of 100 participants based on a small pilot study in which we already observed significant outcome-preference alignment effects, and power analysis aiming to detect an effect size of 0.5 points (out of 20) for process effect. After the data validation described below, 93 responses remained. No additional stratified (sub)sampling was performed.

Participants were 57% female and 43% male. Participants were also well balanced between political affiliations, with 35% identifying as liberal, 31% as conservative, 31% as independent, and a remaining 2% refraining from reporting affiliation. Roughly 60% of participants were between the ages of 25 and 44, with 35% 45 or older. This age distribution mirrors that of US Facebook users [27], although it underrepresents the 18-24 age group. The survey population reported as 80% White, 11% Asian, 4% Mixed Race, 3% Black, and 1% Native American, with a further 1% declining to report. Additionally, 11% of participants reported as Hispanic or Latino, across all race categories. Compared to the US population, our survey population was more educated, with only 23% reporting highest attainment as high school, 38% with a Bachelor’s degree, and 18% with a Master’s or higher.

3.2.2 Experimental Manipulation. Each participant was shown four moderation decisions consisting of 1) a post randomly selected from the nine, 2) one of the four moderation processes, 3) a random decision outcome—taken down or left up, and 4) a brief indication of the violation category if the post was taken down. Each moderation process was shown exactly once, in random order, and posts were sampled such that each of the three categories of content violation would be seen at least once in the four decisions. The moderation process descriptions shown to participants were intentionally kept short to allow pre-existing attitudes and assumptions to be captured in responses, and to approximate the opaque nature of content moderation as practiced today [116]. These descriptions are provided in Appendix A. From our pilot studies, we found that the descriptions were adequate for users to be able to understand and differentiate between the moderation processes, aligning with prior work [11, 89, 95].
Table 1. Questions used to measure each aspect of institutional legitimacy in the survey experiment. Users responded to these questions on a 5 point Likert scale. [Moderation process] should be replaced with the moderation process in question (e.g., digital juries of Facebook users).

3.2.3 Measures. The perceived institutional legitimacy of moderation decisions served as the primary quantitative measure of the survey. Five survey questions, given in Table 1, were posed to participants for each moderation decision, corresponding to five component measures of perceived institutional legitimacy—outcome satisfaction, users’ trust in the process, perceived fairness and impartiality, institutional commitment, and decisional jurisdiction. The questions assessing trustworthiness, institutional commitment, and decisional jurisdiction were adapted from Gibson’s work surveying the institutional legitimacy of national high courts [51, 55], while the question assessing fairness and impartiality was adapted from a study measuring perceived legitimacy of state Supreme Courts [54]. These questions were modified to fit the domain, and institutional commitment and decisional jurisdiction were flipped from negative to affirmative to better suit our hypothetical setting. We also included a question assessing outcome satisfaction (found to be positively correlated with institutional legitimacy [55]) using similar language to prior work evaluating content moderation [44]. As in prior empirical work [44, 51, 55], terms like fairness and impartiality were not rigorously defined to avoid unduly influencing participants with prescriptive normative criteria. The responses to these questions were captured on a five point Likert scale, and we calculated Cronbach’s alpha (a common measure of internal consistency) between these five measures in our data as 0.92. These component measures were summed to create a composite measure.

In addition to this quantitative measure, the survey also collected qualitative data through free response questions. Participants were randomly asked to elaborate on their responses to quantitative questions 50% of the time. Additionally, after answering questions about the four moderation decisions, all participants were asked 1) to select the process they saw as the most trusted, least trusted, most fair and impartial, and least fair and impartial, and 2) to provide a brief rationale(s) behind their choices. These comparative questions were included not only to corroborate quantitative results, but also because prior research shows that people can be more effective in making comparative judgments [148].

Demographic information, including age, gender, race, ethnicity, education level, work experience, political affiliation, income, and Facebook usage, was also collected primarily to assess the representativeness of the participant group, and in limited cases to test for association with perceived legitimacy (detailed below).

3.2.4 Data Validation. In order to validate responses, users were asked to answer attention check questions (repeating back details about the moderation process and outcome). Any participants that
failed these attention check questions were removed from the dataset. In addition, any spam-like submissions were removed.

3.2.5 Quantitative Modeling. Quantitative responses were analyzed using a linear mixed effects (LME) model in which the degree of alignment of individual normative preference with outcome (Alignment), content moderation process (Process), Gender, and Political Affiliation serve as explanatory variables, and measures of perceived legitimacy, as the response variable (as described in Section 3.2.3). The inclusion of Alignment and Process relate to RQ2 and RQ1 respectively, while Gender and Political Affiliation are included because they have been shown to relate to perceived legitimacy in prior work [52, 136]. We intentionally do not control for participants’ prior exposure to content moderation, as perceived legitimacy measures population-wide attitudes—it is a sociological phenomenon that must be assessed within a representative population sample. In this model, each participant is given a random intercept and slope for decision outcome, allowing for the possibility that each participant may have a different inclination to take down or leave up posts. Additionally, each post is given a random intercept and slope for decision outcome and political affiliation, as specific posts may be more or less objectionable across the population, and many posts have a significant political dimension.

The composite measure of perceived legitimacy serves as the dependent variable for the primary model, which is used for all hypothesis tests. To further understand how the explanatory variables relate to individual measures of perceived legitimacy, parallel submodels were also fit with each of the five perceived legitimacy measures as dependent variables.

3.2.6 Qualitative Coding. To analyze the four final comparative free response questions, two co-authors identified the moderation process named by each participant. If no process could be identified, the entire response was excluded from analysis. If multiple processes were indicated, only the process identified as a first choice was coded if a relative ordering was given, otherwise all processes were coded. The same two co-authors then performed an open coding procedure to identify all meaningfully distinguishable attitudes expressed by participants in their answers. Noting that participants frequently hedged their answers and expressed multiple attitudes at a time, one co-author then developed a framework of axial codes in which attitudes were coded as a series of (role, subject, predicate) triples, each component of which is defined as follows:

1. role: Indicates whether the attitude served as a rationale for the answer, qualification of the answer, or condition for the answer.
2. subject: Indicates to which of the four moderation process(es) the attitude pertains.
3. predicate: Indicates the idea being expressed about the subject.

For example, the attitude expressed in, “I think a panel of experts can be most trusted because they have the training needed to make good decisions, but the platform can select experts in a biased way,” might be coded as: [(rationale, Expert, Code 8 : "has necessary training"), (qualification, Expert, Code 13 : "controlled by platform")]]. The two co-authors independently rated all responses, and Cohen’s kappa, a metric of inter-rater reliability, was calculated separately for each possible code. Across processes we calculated a mean kappa of 0.99, and across attitude triples we calculated a frequency-weighted average kappa of 0.61. Subsequently, the two co-authors discussed inconsistencies and reached unanimous agreement on the final coding of each response. The full attitude coding scheme contains approximately 50 distinct predicates, which are given in Appendix B.

4 RESULTS

The primary quantitative model of survey responses estimates the effect on perceived institutional legitimacy of Alignment, Process, Gender, and Political Affiliation. Regression coefficients
Variable | Alternative: No Interactions | Primary Model
--- | --- | ---
Alignment | 1.86*** (0.15) | 1.87*** (0.15)
Algorithm | -0.66 (0.34) | -0.68* (0.34)
Expert Panel | 1.18*** (0.34) | 1.14*** (0.34)
Digital Jury | -0.56 (0.34) | -0.55 (0.34)
Male | -0.52 (0.67) | -0.45 (0.67)
Conservative | -1.04 (0.84) | -1.10 (0.84)
Independent | -2.14* (0.83) | -2.18** (0.82)
Unreported Affiliation | -3.80 (2.32) | -3.78 (2.32)
Alignment * Algorithm | 0.06 (0.26) | 
Alignment * Expert | 0.15 (0.25) | 
Alignment * Jury | 0.17 (0.24) | 
Constant | 17.68*** (0.64) | 17.71*** (0.65)

Table 2. Determinants of perceived legitimacy as modeled by primary and alternative LME models. Process is modeled using deviation contrasts such that the Constant reflects the mean across processes, and coefficients are comparable between models. Significance levels are indicated for readability purposes only and are calculated with t-tests using Satterthwaite’s method [97]. These levels do not constitute formal hypothesis tests. \( p < 0.001^{***}, p < 0.01^{**}, p < 0.05^* \)

4.1 Perceived Legitimacy of Moderation Processes

To evaluate H1, we conduct Tukey’s HSD test as a *post hoc* analysis of the pairwise differences in perceived institutional legitimacy across moderation processes. This test allows for significance testing across more than two groups and makes fewer assumptions than t-tests, which are not universally accepted for LME model parameters [97]. Results are presented in Table 4. From this test,
### Table 3. Proportion of participants indicating each moderation process as ranking the highest and lowest with respect to *trustworthiness* and *impartiality*. Raw participant counts are given in parentheses.

<table>
<thead>
<tr>
<th>Process</th>
<th>Trustworthiness</th>
<th></th>
<th>Impartiality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest (%)</td>
<td>Lowest (%)</td>
<td>Highest (%)</td>
<td>Lowest (%)</td>
</tr>
<tr>
<td>Contractor</td>
<td>14% (13)</td>
<td>35% (31)</td>
<td>9% (8)</td>
<td>46% (41)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>30% (28)</td>
<td>34% (30)</td>
<td>51% (46)</td>
<td>13% (12)</td>
</tr>
<tr>
<td>Expert</td>
<td>41% (38)</td>
<td>9% (8)</td>
<td>28% (25)</td>
<td>8% (7)</td>
</tr>
<tr>
<td>Jury</td>
<td>28% (26)</td>
<td>27% (24)</td>
<td>18% (16)</td>
<td>36% (32)</td>
</tr>
</tbody>
</table>

Table 4. Tukey’s HSD test results of significance in the difference in mean perceived legitimacy across moderation processes. $p < 0.001$ *, $p < 0.01$ **, $p < 0.05$ *.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm - Contractor</td>
<td>-0.78</td>
<td>0.60</td>
</tr>
<tr>
<td>Expert - Contractor</td>
<td>1.02</td>
<td>0.60</td>
</tr>
<tr>
<td>Expert - Algorithm</td>
<td>1.81*</td>
<td>0.60</td>
</tr>
<tr>
<td>Jury - Contractor</td>
<td>-0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>Jury - Algorithm</td>
<td>0.13</td>
<td>0.62</td>
</tr>
<tr>
<td>Jury - Expert</td>
<td>-1.68*</td>
<td>0.61</td>
</tr>
</tbody>
</table>

we can conclude that decisions made by the expert panel are perceived as more legitimate, according to our definition, than decisions made by both the digital jury and algorithm; however there is not sufficient evidence to draw conclusions about the perceived legitimacy of other moderation processes. Consequently, we find partial support for H1.1 and are able to disprove H1.3, but do not find evidence to support or disprove H1.2. To better visualize the varying perceived legitimacy of the four moderation processes, a marginal effects plot is presented in Figure 1. Additionally, a summary of coded free responses to the comparative questions is presented in Table 5 and 6, and corresponding quantitative results are presented in Figure 2.

4.1.1 H1.1. Quantitative results show that the expert panel has higher perceived legitimacy than the algorithm, supporting one component of H1.1. In free response, an important factor for participants appeared to be whether decisions were made by groups or individuals. 24% of respondents suggested that contractors would make more biased decisions as single individuals, and 24% suggested that contractors would apply their own beliefs and agenda. Moreover, a full 42% of participants expressed support in some form for the idea that groups of moderators can be more trustworthy and/or impartial that single moderators. In contrast to our expectations, qualitative results cast doubt on independence as a major factor driving perceived legitimacy. Many participants acknowledged the greater independence of the two deliberative bodies—25% of participants expressed a belief that paid contractors would carry out the agenda and biases of the platform and 16% expressed concern that the algorithm could be programmed with platform biases, while much smaller proportions expressed similar ideas about expert panels or digital juries (2% for both). However, only 8% of participants selecting expert panels and 6% of those selecting the digital jury as the most impartial process provided independence as a rationale, with similar or smaller proportions among those selecting these processes as the most trustworthy.
Fig. 1. The mean perceived legitimacy of each process is plotted, shown with a 95% confidence interval calculated from variance of the fixed effect estimates.

(a) The mean trustworthiness of each process. (b) The mean impartiality of each process.

Fig. 2. Trustworthiness and impartiality of each moderation process. Each mean is shown with a 95% confidence interval calculated from variance of the fixed effect estimates. The Bonferroni correction used for hypothesis tests is not applied here.
Comparing The Perceived Legitimacy of Content Moderation Processes 82:15

<table>
<thead>
<tr>
<th>Pr.</th>
<th>Rationale</th>
<th>Highest Trustworthiness</th>
<th>% (n)</th>
<th>Rationale</th>
<th>Lowest Trustworthiness</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contractor</td>
<td>It’s their job</td>
<td>38% (5)</td>
<td>Single person bias</td>
<td>61% (19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Has necessary training and knowledge</td>
<td>31% (4)</td>
<td>Implements platform agenda and biases</td>
<td>39% (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>Decides based on logic, data, rules</td>
<td>32% (9)</td>
<td>Lacks human factors of cognition</td>
<td>50% (15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doesn’t apply own beliefs and agenda</td>
<td>18% (5)</td>
<td>Generally performs poorly</td>
<td>40% (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td>Has necessary training and knowledge</td>
<td>50% (19)</td>
<td>Performed worse (in survey)</td>
<td>25% (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple people helps mitigate bias</td>
<td>26% (10)</td>
<td>Applies own beliefs and agenda</td>
<td>13% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jury</td>
<td>Multiple people helps mitigate bias</td>
<td>31% (8)</td>
<td>Applies own beliefs and agenda</td>
<td>54% (13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doesn’t apply own beliefs and agenda</td>
<td>12% (3)</td>
<td>Random selection process not sufficient</td>
<td>38% (9)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Most frequent rationales given for answer among participants selecting each process as having the highest and the lowest trustworthiness. Proportions are given as percentages, with raw participant counts given in parentheses.

<table>
<thead>
<tr>
<th>Pr.</th>
<th>Rationale</th>
<th>Highest Impartiality</th>
<th>% (n)</th>
<th>Rationale</th>
<th>Lowest Impartiality</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contractor</td>
<td>Faithfully adheres to guidelines</td>
<td>38% (3)</td>
<td>Single person bias</td>
<td>41% (17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is accountable for decisions</td>
<td>25% (2)</td>
<td>Applies own beliefs and agenda</td>
<td>34% (14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>Doesn’t apply own beliefs and agenda</td>
<td>52% (24)</td>
<td>Generally performs poorly</td>
<td>50% (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decides based on logic, data, rules</td>
<td>37% (17)</td>
<td>Lacks human factors of cognition</td>
<td>42% (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td>Has necessary training and knowledge</td>
<td>24% (6)</td>
<td>Applies own beliefs and agenda</td>
<td>43% (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple people helps mitigate bias</td>
<td>20% (5)</td>
<td>Unaccountable (e.g., lacks oversight)</td>
<td>43% (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jury</td>
<td>Multiple people helps mitigate bias</td>
<td>44% (7)</td>
<td>Applies own beliefs and agenda</td>
<td>56% (18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is independent of platform</td>
<td>6% (1)</td>
<td>Lack necessary training and knowledge</td>
<td>25% (8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Most frequent rationales given for answer among participants selecting each process as having the highest and the lowest impartiality. Proportions are given as percentages, with raw participant counts given in parentheses.

4.1.2 H1.2. Quantitative estimates of the perceived legitimacy of algorithms and paid contractors were not statistically distinguishable. However, qualitative analysis of free response provides more clues. Pluralities of respondents designated the paid contractor as the least trustworthy and least impartial decision maker, while a majority (51%) chose the algorithm as the most impartial, suggesting some support for H1.2. The discrepancy between the quantitative estimates and free response answers for paid contractors is notable. These discrepancies might be due to estimation error and lack of statistical significance, or alternatively by substantive differences in the framing of the quantitative and free response questions. In quantitative questions, participants were asked to provide ratings in isolation, while in free response, they were asked to consider all four processes simultaneously. Moreover, in free response questions, participants were only asked to discuss the most and least trustworthy and impartial processes.

Participants were concerned with paid contractors implementing their own agenda and biases (24%), despite the limited role of personal interpretation in contractor moderation in most platforms [10, 107], or implementing the biases of the platform (25%). Interestingly, some participants
viewed a paid relationship as a corrupting influence, while others viewed it as source of account-ability. While a large proportion of respondents who labeled the contractor as untrustworthy (39%) and partial (34%) also expressed that the contractor would be subject to platform control, 10% of participants in each case expressed concern that to contractors, moderation would be “just a job.” We anticipated that paid contractors would be perceived as less legitimate due to lack of clarity about their background and lack of faith in their expertise and ability to make nuanced judgments. In free response, however, these types of concerns were expressed by <5% of respondents.

By contrast, 25% of respondents made comments like “The least trustworthy would likely be the algorithm due to the complex nature, nuance, and context of the human language. Algorithm[s] cannot navigate the complexities and subtleties of our communications.” 16% expressed awareness that algorithms can be programmed with built-in bias, suggesting that support can depend on specific details of how and why an algorithm is created. Additionally, 32% of respondents made performance based arguments (as rationale or qualification) about algorithms, markedly higher than for contractors (2%), digital juries (9%), and expert panels (8%). Even many participants who expressed support for algorithmic moderation had reservations. Although nearly one third of respondents believed the algorithm was the most trustworthy process, this support was made conditional at the highest rate of all processes, depending on factors like the algorithm being constructed fairly and impartially (25%) and decisions being subject to checks and balances (11%) and appeal to humans (7%), with similar rates for impartiality.

4.1.3 H1.3. Quantitative and qualitative results definitively refute H1.3, and both show a strong preference for the expert panel. We anticipated that the greater democratic legitimacy of digital juries and skepticism of claims to expertise in content moderation would override other considerations. While we did find some support for these phenomena in free response, by and large participants viewed expert panels as legitimate, trustworthy, and impartial. Although we anticipated juries’ democratic nature might be seen as a check on the platform’s ability to impose its own standards on the community (a view articulated by few respondents), participants seemed more concerned that digital juries would impose members’ own viewpoints (expressed by 30%) and that vetting would not be rigorous enough (26%). One participant stated, “It would be very difficult for users who liked a person who posts things that violated the standards to be impartial...” Another commented, “they are randomly chosen and could be just about anybody. If there was some type of selection process from Facebook users, then that would be a little bit different.” One participant even fretted about demographic bias in randomly selected juries, saying “Facebook users tend towards certain demographics – the middle aged and not people like me who are younger.” Additionally 8% expressed the idea that regular users are inherently unsuited to the task. Some participants went as far as to reject the legitimacy of juries in the justice system, in one instance, stating, “[The unfairness of juries of users] is similar to how ineffective an actual jury is at trial.” By contrast, 25% of participants showed appreciation for the expert panel members’ training and expertise, suggesting that their perceived greater formal education and experience would help mitigate bias.

4.2 Importance of Outcome-Preference Alignment
To assess H2, we consider both qualitative and quantitative factors. Qualitatively, the magnitude of the fixed effect of ALIGNMENT, as well as its significance lends support to H2—that users will report higher perceived legitimacy when content moderation systems make decisions that align with their individual preferences. Since ALIGNMENT is on a five-point scale, the maximal variation in perceived legitimacy due to ALIGNMENT is approximately 7.4 points out of 20, larger than that of any other variable (see Figure 3).
Comparing The Perceived Legitimacy of Content Moderation Processes

Quantitatively, we find that comparing the model with alternative models with a single variable removed, the largest regression occurs when removing ALIGNMENT according to the Akaike information criterion (AIC), an information theory based measure that balances goodness of fit with model complexity. Performing an ANOVA comparison between the primary model and a model without ALIGNMENT, we calculate $p < .001$. Additionally, we calculate the marginal $R^2$ value [105] for a reduced model using ALIGNMENT as the sole predictor variable as 0.27, suggesting that 27% of the variance in perceived legitimacy is explainable by outcome-preference alignment, assuming the modeled random effects.

Our model estimates that interactions between ALIGNMENT and Process are small and not statistically significant. Parameter estimates of these interactions are presented in Table 2.

In free response, many participants’ explanations for their legitimacy ratings rested solely on their opinion of the moderated post and the random decision shown to them in the study. In addition, when participants were asked to assess the overall trustworthiness and impartiality of moderation processes, a significant proportion made arguments based on the survey examples.

5 DISCUSSION

5.1 Implications of Process Effects

Our quantitative and qualitative results build a strong case that the Expert Panel is perceived as the most legitimate process by our participants. This result might be considered surprising in light of common criticisms that platforms are undemocratic and biased in favor of unpopular views [77]. The result is especially notable given the limited information participants were given regarding panelist selection, ideological alignment, and the nature and relevance of their expertise. Moreover, our results’ direct refutation of H1.3 seems to show that expertise, rather than the
body’s independence from the platform or other characteristics, was what participants appreciated. Perhaps, despite the popular notion of a crisis of mistrust in expertise [42], mistrust of peers is stronger still [90], though different results may be obtained in high trust societies [33] or those whose cultures are poorly represented by the expert body. As one participant noted, “they are experts, they know how to deal with things like this better than anyone. They can be trusted more to make the right decisions.”

With respect to digital juries, other work studying online communities finds similar mistrust of peers and resistance to peer judgment as we observed [44, 85]. Digital juries might offer benefits in certain scenarios, by aligning content moderation enforcement with users’ preferences. However, there remains debate around the ability of digital juries to scale effectively and carry out moderation decisions on platforms that lack diversity like Parler [114, 132].

Recent rulings by the Facebook Oversight Board, in particular its rulings about President Trump’s posts following the 2021 Capitol Riot [6] show that the exercise of expert authority in content moderation can be fraught in ways that go beyond the process factors examined in this study. Firstly, it is clear that the composition of the body, a variable not manipulated in this study, strongly colors decisions. The board, with heavy representation from lawyers and judges, has couched its reasoning within the framework of judicial review, self-imposing significant limits on the scope of its powers [106]. Secondly, experts may face challenges in claiming and exercising authority, like authority to craft policy, and may be tempted to take a middle of the road approach in controversial cases to safeguard their own perceived legitimacy in the short term [37]. Lastly, while limiting the body’s scope may help avoid controversy, in practice it may push important work like determining how international human rights law applies to content moderation [37], to platform-internal processes with less legitimacy and transparency.

Despite inconclusive results for H1.2, qualitative analysis does support many phenomena regarding algorithmic decision making discussed in prior work—these phenomena suggest widespread belief in algorithmic objectivity but also show several factors limiting trust. The most widely articulated of all attitudes toward moderation processes in the free response was that algorithms don’t apply their own beliefs and agenda to decisions, and the fourth most common was that algorithms make decisions based on logic and rules, not feelings. However, study results show that belief in impartiality does not necessarily translate into a high level of trust or perceived legitimacy—a similar number of respondents gave the algorithm as the least trustworthy process as had given the paid contractor. Prior work notes that positive sentiments toward algorithmic decision making are tempered by factors like the subjectivity of the domain [95, 124], opaqueness of function and deployment [41, 79], and performance [35], elements that can be seen in free response.

5.2 Implications of Outcome-Preference Alignment
Quantitative results firmly support H2, showing that outcome-preference alignment strongly determines perceived institutional legitimacy. This influence far outstrips that of the process variables manipulated in this survey. This result arguably poses an intractable problem for platforms, discussed further in Section 5.5, and raises important questions about the perceived legitimacy of majoritarian decision making.

While it is difficult to disentangle outcome-preference alignment and performance in qualitative analysis, outcome favorability bias is a well documented phenomenon in both criminal justice [94] and algorithmic decision making [144]. In contrast to these contexts, however, outcome favorability for bystanders in content moderation is driven less by personal interest and more by beliefs, ideology, and community norms. We can expect, therefore, for personal experience and political discourse to be especially important in shaping presences and the subjective experience of content moderation, as discussed below.
5.3 Familiarity, Understanding, and Experience

Perceived legitimacy is a sociological phenomenon, and can only be meaningfully studied in the context of a society and the attitudes of individuals therein; however, it naturally follows that levels of perceived legitimacy will vary with the nature of and degree of public awareness, understanding, and idiosyncratic experience.

In comparative free response, five respondents gave a rationale or qualification that they did not understand a process well enough. Some respondents expressed this skepticism forcefully, for example, writing, “...my question is, what are the experts experts in? How do we verify their expertise, and ensure they are operating in an unbiased manner?” and “I do not trust the algorithm because I’m unsure how it was made and what it is looking for in a content in order to determine if it should be removed or not.” However, transparency doesn’t necessarily confer trust [9], especially for algorithmic decision making [24]. As users gain more understanding of the true capabilities of algorithms, they may instead grow more skeptical [32]. Because content moderation today remains opaque to users [60, 116], it is important to ask the question how more knowledge might affect users’ attitudes. By the same token, steps taken by platforms to help build legitimacy can only be effective when users know about them.

Personal experience can also play a major role in shaping attitudes. For algorithmic moderation, as performance improves, positive personal experiences with algorithms would be the likeliest path to changing attitudes. For unfamiliar, emerging content moderation processes, like the digital jury, initial user experiences with the system will be especially important. Although this study did not directly measure or control for familiarity or experience with content moderation mechanisms, future work can explore a single individual’s subjective experience of legitimacy due to personal experience.

5.4 Political Discourse

Despite inconclusive quantitative results on the effect of political affiliation, it is clear that the role of political discourse in shaping attitudes toward content moderation processes cannot be ignored. Anecdotally, multiple participants complained that moderators would be chosen to reflect a political viewpoint, and one participant consistently voiced mistrust of the platform due to its liberal bias. In such cases, political affiliation appeared to play a strong role and some weak patterns emerged—experts were presumed by some users to have liberal bias, and digital juries were presumed to be more tolerant of harmful content. In the United States, content moderation has become a flashpoint, and is viewed by many conservative-leaning individuals as an illegitimate attempt to regulate speech. Rhetoric from partisan opinion leaders, for example the Republican-led FCC’s announcement that it would try to reduce liability protections for platforms that moderate content [4], both reflects and shapes public opinion. Moreover, prior work shows that reactions to hypothetical interventions taken by social media platforms can be heavily influenced by party ideologies [63], and that more generally, liberals and conservatives place a differing degree of importance on components of perceived legitimacy, such as fairness [61]. In this study, qualitative responses show evidence of systemic skepticism on the part of those identifying as conservatives or independents. In any case, the prominence of content moderation as a political issue adds an element of volatility to any attempt to build legitimate moderation systems.

5.5 Design Implications

Although the strong effect of outcome-preference alignment appears to pose a daunting challenge for platforms, our findings suggest platforms have procedural levers at their disposal to build perceived legitimacy. We outline several such suggestions below, synthesizing our findings with
analysis of prior work, industry developments, and speculation. However, we note that in general these should be implemented as part of a tiered, hybrid system that not only optimizes for perceived legitimacy but also allows for fast response times in cases where there is a likelihood of immediate harms (e.g., 2019 Christchurch terrorist attack [84]), and accommodates the challenges of scale. Moreover, adopting these measures is only a first step for platforms—indeed, some are already in use. Perceived legitimacy cannot exist without both transparency and public awareness of these efforts.

Because our findings offer clear support for expert panels, we recommend that such bodies be incorporated into moderation procedures. In practice, however, it would be impractical for such panels to make a large proportion of moderation decisions. Platforms should explore alternative means to incorporate expert judgment into hybrid processes. As a first step, we suggest that a publicly visible and independent expert panel be responsible for drafting moderation guidelines. A next step would be to allow the expert panel to handle appeals of the most controversial cases [78]. By contrast, the Facebook Oversight Board has focused first on deciding borderline cases, and does not have the authority to set policy (though it may recommend policy changes when solicited to do so)[5]. Such an appeals body is especially important when algorithmic moderation is used, given the perceived importance of oversight among study participants. However, expertise can be brought to other places. Digital juries might, for example, include an expert member to facilitate deliberation. More broadly, an independent expert group might be given authority over the overall moderation process. Finally, experts might play a visible role in training rank-and-file moderators, assessing the performance of and appropriate scope for automated systems, and educating the public about the content moderation process.

Because large proportions of our participants displayed wariness of individual moderator biases as well as groupthink in deliberative bodies, we recommend that platforms incorporate multiple perspectives into all processes. While our results might seem to imply majoritarian decision making can be seen as a legitimate in a utilitarian sense, we believe diverse perspectives are even more critical in a divided environment. Platforms should, for example, make clear to users that posts are reviewed by multiple contractors, assuaging our participants’ fears that contractors apply their own biases and opinions to moderation. While Facebook is known to monitor agreement between contractors [107], our results show this is not part of the public consciousness. Deliberative bodies could also employ pre-screening to encourage more diverse composition.

To address concerns about members of digital juries applying personal biases to decisions, we suggest exploring public reputation systems to improve accountability for decisions. Our results suggest that anonymity and lack of vetting hinders accountability. Reputation systems could range from publishing jury deliberation and justifications of decisions to a numerical rating system driven by peer reviews. Similar methods could also improve the accountability of contractors.

We can also look to prior work for solutions to the outcome-preference alignment problem. It is informative to consider prior work on the US Supreme Court, an institution that is forced to make polarizing, politically charged decisions in the public eye. Gibson, for example, suggests legitimacy can arise through a social learning process [53]. It is reasonable, therefore, to conclude that platforms may be able to improve perceived legitimacy over time through sustained public education efforts. Platforms should publish information like how automated systems are constructed and how moderators are selected and trained. Gibson also writes about the negative effects of politicization for perceived legitimacy [21], an outcome platforms should take care to avoid.

We can also look to the literature on procedural justice—there is evidence, for example, that perceptions of legitimacy are enhanced when authorities take extra time to explain how they reached decisions [138], that having the opportunity to express views and opinions, as a user might have during an appeal, can enhance feelings of procedural fairness irrespective of outcome [94],
and that mere knowledge of such a right can have beneficial effects even if not availed [138]. An analogous phenomenon has also been described in the context of online content moderation [74], and it is likely that enhancing the quantity and quality of communication between the user and platform can improve perceived legitimacy.

5.6 Limitations

While this study provides a novel comparative perspective on content moderation processes, the study design has several limitations. First, the study attempts to measure and analyze prevailing public attitudes toward content moderation processes. However, we recognize that the formation of attitudes is a multi-faceted social and experiential process, and our study design does not allow rigorous claims about attitude formation. Furthermore, our study measures attitudes at a single snapshot in time—we did not provide an opportunity for participants to gain experience with each process, instead exposing participants to a single decision per process. In addition, the study focused on highly disagreed-upon posts, and results may not generalize to all types of moderated content. Additionally, since the study only investigated one possible version of each process type, results may not generalize to all possible versions of these processes. A future study could not only examine more versions of these processes, but also identify which attributes (e.g., jury rules) contribute to perceived legitimacy.

Since our study was scoped to only include industrial content moderation, we did not investigate moderation and artisanal moderation, which gives rise to two limitations. The first is that we know less about the perceived legitimacy of the excluded approaches. Second, this work only investigates how moderation processes impact the perceived legitimacy of rule enforcement, not rule creation. Thus, the results may not generalize well to the perceived legitimacy of moderation processes involved in rule creation.

Biases in our user population may also limit generalizability of results. While the study was conducted among Facebook users, the demographics of AMT workers do not exactly match that of Facebook’s US user base. The survey population overrepresented higher-educated and non-Hispanic white individuals and underrepresented multiple minority groups. Furthermore, the technical and digital nature of AMT work may mean that our survey respondents had a different relationship with online platforms than average social media users. Additionally, our study was limited to one social media platform (Facebook) in one country (United States). In addition, since attitudes and perceptions vary upon their existing knowledge of content moderation, the results may not generalize to populations with highly expert populations. A future cross-cultural study may be needed to determine which drivers of perceived legitimacy are more universal and which are more specific to the United States, its present cultural moment, and the present level of knowledge about content moderation.

Additionally, two potentially significant factors of moderation processes we do not consider in this work are when moderation is carried out (i.e., pre-moderation vs. post-moderation) [29, 142], and tiered or hybrid moderation processes. Qualitative results suggest knowledge of oversight mechanisms and appeals processes can influence perception of legitimacy, and holistic assessment of perceived legitimacy of governance mechanisms in practice requires consideration of the entire system.

5.7 Future Work

Future studies should more rigorously examine tiered processes, the impact of oversight, and the appeals process. Our qualitative results indicate that including these processes may be especially significant for algorithmic moderation, where participants indicated a desire for human oversight. Additionally, future work should examine how the wide array of artisanal and community-driven
moderation models found in platforms like Reddit, Vimeo, Patreon, Wikipedia, and League of Legends affect perceived legitimacy when employed together with or in place of industrial content moderation processes. Given that these approaches tend to have more community participation in governance, future work comparing them needs to be careful to separate the investigation of moderation legitimacy from that of governance.

A second area for future work is to investigate how hybrid moderation processes can better incorporate expertise. While we can hypothesize the benefit of expertise for perceived legitimacy will diminish the further removed experts are from day-to-day decision making, hybrid models are the only practical solution to scaling challenges. A promising direction might be to combine elements of the expert panel and digital jury—for example, including an expert facilitator or introducing credentialing for jury members. Additionally, future work should examine the importance of specific types of expertise and representation of diverse viewpoints.

A third critical area is to investigate the impact of political affiliation and political debate on perceived legitimacy. While quantitative and qualitative results hinted that conservatives and independents may be less trusting of content moderation processes in general, the power of the study was not sufficient to establish this. Future studies can not only investigate this effect, but also examine the interaction of political affiliation with elements of moderation process design. Moreover, while this study did establish the importance of normative preferences, it did not attempt to distinguish political or closely held preferences from other preferences, and did not specifically distinguish content with a significant political dimension from content without this dimension. An especially important topic for future work is studying how sticky factors like institutional commitment and decisional jurisdiction are in the face of politically unpalatable decisions, and what role normative concepts like democratic legitimacy play in politically charged environments.

6 CONCLUSION

As online platforms and their governance mechanisms increasingly resemble digital polities, platforms must focus greater attention on user perceptions of legitimacy. However creating a legitimate content moderation process appears to be a nearly intractable problem as long as people with different views continue to occupy the same digital spaces. Not only is the scale of the task daunting, but this study also highlights the degree to which individual outcome preferences can dominate perceptions of legitimacy, regardless of how platforms design their processes. Content for which opinions differ wildly, therefore, poses a “catch-22” to platforms—goodwill generated with one segment of the user population may be met in equal measure with feelings of illegitimacy by another.

Nevertheless, our quantitative and qualitative results illuminate potential paths forward. We find the strongest support for a robust role for experts in content moderation processes, with participants perceiving the expert panel as having high levels of trustworthiness, fairness and impartiality, and overall perceived institutional legitimacy. Our qualitative results also indicate a preference among users for group decision making over decisions made by individuals, supporting future work on processes that synthesize multiple views. Our results are also consistent with prior work on attitudes toward algorithmic decision making, showing that while algorithms can be perceived as legitimate decision makers, their performance and users’ experience with them will significantly shape attitudes.

Today, content moderation stands at an inflection point. While platforms are accountable for their content moderation practices, academics and policymakers are increasingly vocal participants in shaping the future of content moderation, and the public will have the final say. Criticism of existing mechanisms abounds, but so do proposals and experiments seeking to build better systems.
Studies of perceived legitimacy can be a powerful tool for all groups to ensure these systems are trusted and respected by the public.

ACKNOWLEDGMENTS
We thank the reviewers for their helpful comments and suggestions. We also thank Michael Sklar of the Stanford Department of Statistics for statistical advice. The Stanford Computer Science (CS) department provided funds to assist this research. This project was supported by the Office of Naval Research (N00014-21-1-2839).
REFERENCES

Comparing The Perceived Legitimacy of Content Moderation Processes


Comparing The Perceived Legitimacy of Content Moderation Processes

References


7 APPENDICES

7.1 Appendix A

<table>
<thead>
<tr>
<th>Content Moderation Process</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paid Contractor</td>
<td>The content moderation decision was made by a human contractor employed by Facebook. The human contractor was trained in a workshop with examples of posts that violated Facebook’s Community Standards.</td>
</tr>
<tr>
<td>Algorithm</td>
<td>The content moderation decision was made by an algorithm that was built by software engineers at Facebook. The algorithm was trained on examples of posts that violated Facebook’s Community Standards.</td>
</tr>
<tr>
<td>Digital Jury</td>
<td>The content moderation decision was made by a jury of 6 randomly-selected Facebook users. Jury members received training on enforcing Facebook’s Community Standards, and after structured deliberation in a video conference session, reached a unanimous decision.</td>
</tr>
<tr>
<td>Expert Panel</td>
<td>The content moderation decision was made by a panel of 6 experts selected for their expertise in content moderation, human rights, and digital rights. After structured deliberation in a videoconference session, they reached a unanimous decision.</td>
</tr>
</tbody>
</table>

Table 7. Descriptions of each of the moderation processes shown to survey participants with decision outcomes.
7.2 Appendix B

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Subject to single person bias</td>
</tr>
<tr>
<td>2</td>
<td>Abuses power</td>
</tr>
<tr>
<td>3</td>
<td>Doesn’t abuse power</td>
</tr>
<tr>
<td>4</td>
<td>Applies own beliefs and agenda</td>
</tr>
<tr>
<td>5</td>
<td>Doesn’t apply own beliefs and agenda</td>
</tr>
<tr>
<td>6</td>
<td>Allows for multiple perspectives to mitigate bias</td>
</tr>
<tr>
<td>7</td>
<td>Suffers from group-think or peer pressure</td>
</tr>
<tr>
<td>8</td>
<td>Has necessary formal training and experience</td>
</tr>
<tr>
<td>9</td>
<td>Lacks necessary formal training and experience</td>
</tr>
<tr>
<td>10</td>
<td>Takes work seriously because it’s their job</td>
</tr>
<tr>
<td>11</td>
<td>Doesn’t take seriously because not compensated (enough)</td>
</tr>
<tr>
<td>12</td>
<td>Subject to control by the platform</td>
</tr>
<tr>
<td>13</td>
<td>Independent from the platform</td>
</tr>
<tr>
<td>14</td>
<td>Subject to (improper) influence by third parties</td>
</tr>
<tr>
<td>15</td>
<td>Not subject to (improper) influence by third parties</td>
</tr>
<tr>
<td>16</td>
<td>Has rigorous and fair selection process</td>
</tr>
<tr>
<td>17</td>
<td>Lacks rigorous and fair selection process</td>
</tr>
<tr>
<td>18</td>
<td>Selection controlled by the platform</td>
</tr>
<tr>
<td>19</td>
<td>Accountable for decisions</td>
</tr>
<tr>
<td>20</td>
<td>Unaccountable (e.g., lacks oversight)</td>
</tr>
<tr>
<td>21</td>
<td>Makes consistent decisions</td>
</tr>
<tr>
<td>22</td>
<td>Makes inconsistent decisions</td>
</tr>
<tr>
<td>23</td>
<td>Faithfully adheres to moderation guidelines</td>
</tr>
<tr>
<td>24</td>
<td>Doesn’t faithfully adhere to moderation guidelines</td>
</tr>
<tr>
<td>25</td>
<td>Makes a good-faith attempt to consider all factors and sides</td>
</tr>
<tr>
<td>26</td>
<td>I don’t understand process well enough</td>
</tr>
<tr>
<td>27</td>
<td>Performed well in the survey examples</td>
</tr>
<tr>
<td>28</td>
<td>Performed poorly in the survey examples</td>
</tr>
<tr>
<td>29</td>
<td>Generally performs well</td>
</tr>
<tr>
<td>30</td>
<td>Generally performs poorly</td>
</tr>
<tr>
<td>31</td>
<td>Makes decisions based on logic, data, and/or rules, not feelings</td>
</tr>
<tr>
<td>32</td>
<td>May have relationship with defendant</td>
</tr>
<tr>
<td>33</td>
<td>No relationship to defendant</td>
</tr>
<tr>
<td>34</td>
<td>Has human factors of cognition</td>
</tr>
<tr>
<td>35</td>
<td>Lacks human factors of cognition</td>
</tr>
<tr>
<td>36</td>
<td>Can be programmed with biases</td>
</tr>
<tr>
<td>37</td>
<td>Can be programmed with rules</td>
</tr>
<tr>
<td>38</td>
<td>Can be trained or programmed poorly</td>
</tr>
<tr>
<td>39</td>
<td>Can be optimized or improved over time</td>
</tr>
<tr>
<td>40</td>
<td>Costly or impractical</td>
</tr>
<tr>
<td>41</td>
<td>Composed of regular users (who are well equipped to make decisions)</td>
</tr>
<tr>
<td>42</td>
<td>Composed of regular users (who are unsuited to make decisions)</td>
</tr>
<tr>
<td>43</td>
<td>I do not trust</td>
</tr>
<tr>
<td>44</td>
<td>Resembles criminal justice system</td>
</tr>
<tr>
<td>45</td>
<td>I trust</td>
</tr>
<tr>
<td>46</td>
<td>Considers broader social context</td>
</tr>
<tr>
<td>47</td>
<td>Does not consider broader social context</td>
</tr>
<tr>
<td>48</td>
<td>Concerned with upholding individual rights</td>
</tr>
</tbody>
</table>

Table 8. Predicates used for qualitative coding of the free responses written by survey participants. Each predicate represents a core idea expressed in the participant’s response, and can be used to represent a rationale or a qualification for the opinion expressed.
### 7.3 Appendix C

<table>
<thead>
<tr>
<th>Variable</th>
<th>Satisfaction</th>
<th>Impartiality</th>
<th>Trustworthiness</th>
<th>Commitment</th>
<th>Jurisdiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>0.66***</td>
<td>0.29***</td>
<td>0.27***</td>
<td>0.34***</td>
<td>0.33***</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>-0.01</td>
<td>-0.18*</td>
<td>-0.07</td>
<td>-0.14</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Expert Panel</td>
<td>0.07</td>
<td>0.22**</td>
<td>0.21**</td>
<td>0.25**</td>
<td>0.36***</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Digital Jury</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.18*</td>
<td>-0.18</td>
<td>-0.27**</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.03</td>
<td>-0.09</td>
<td>-0.06</td>
<td>-0.14</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.15)</td>
<td>(0.16)</td>
<td>(0.15)</td>
<td>(0.17)</td>
</tr>
<tr>
<td>Conservative</td>
<td>-0.22</td>
<td>-0.24</td>
<td>-0.32</td>
<td>-0.33</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.19)</td>
<td>(0.21)</td>
<td>(0.18)</td>
<td>(0.21)</td>
</tr>
<tr>
<td>Independent</td>
<td>-0.37*</td>
<td>-0.38*</td>
<td>-0.67***</td>
<td>-0.46*</td>
<td>-0.34</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.18)</td>
<td>(0.20)</td>
<td>(0.19)</td>
<td>(0.20)</td>
</tr>
<tr>
<td>Unreported Affiliation</td>
<td>-1.13*</td>
<td>-0.38</td>
<td>-0.90</td>
<td>-0.92</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td>(0.45)</td>
<td>(0.52)</td>
<td>(0.56)</td>
<td>(0.52)</td>
<td>(0.63)</td>
</tr>
<tr>
<td>Alignment * Algorithm</td>
<td>-0.05</td>
<td>0.09</td>
<td>0.02</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Alignment * Expert Panel</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.04</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Alignment * Digital Jury</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.59***</td>
<td>3.61***</td>
<td>3.65***</td>
<td>3.61***</td>
<td>3.25***</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.14)</td>
<td>(0.15)</td>
<td>(0.14)</td>
<td>(0.16)</td>
</tr>
</tbody>
</table>

Table 9. Regression coefficients of perceived institutional legitimacy submodels, whose dependent variables are outcome satisfaction, fairness and impartiality, trustworthiness, institutional commitment, and decisional jurisdiction, respectively. Process is modeled using deviation contrasts such that the Constant reflects the mean across processes, and coefficients are comparable between models. Significance levels are indicated for readability purposes only and are calculated with t-tests using Satterthwaite’s method [97]. These levels are not calculated with the Bonferroni correction used for hypothesis tests, and do not constitute formal hypothesis tests. $|p < 0.001^{***}, p < 0.01^{**}, p < 0.05^{*}$

Received January 2021; revised July 2021; accepted November 2021