Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
December 2022

Algorithms and the Perceived Legitimacy of Content Moderation

Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein

SOCIAL MEDIA PLATFORMS ARE NO STRANGERS TO CRITICISM, especially with respect to their content moderation policies. More speech is taking place online. Simultaneously, online and offline speech are becoming increasingly entangled. As these trends continue, social media platforms' content moderation policies become ever more important. How companies design their algorithms and determine what speech should and should not be removed are just some of the decisions that impact users, the platform, and how the platform is perceived by policymakers and the public.

The public perception of legitimacy is important. Research in other fields underscores that institutions often depend, at least partly, on people accepting their authority. For courts, whether or not citizens buy into a court's ruling can impact how many people will respect the law, adhere to it, and trust the court system to operate effectively. The same idea of legitimacy applies to social media companies. If their content moderation processes—from human reviews to algorithmic flagging—are not perceived as legitimate, it will impact how users and policymakers view and engage with the platform. It can also shape whether users believe they must follow platform rules.

In our paper, "Comparing the Perceived Legitimacy of Content Moderation Processes," we dive into this problem by surveying people's views of Facebook's content moderation processes. We presented U.S. Facebook users with content

Key Takeaways

- Policymakers play an important role in shaping the future of online speech and content moderation, but so does the public. Understanding people's perceptions of content moderation legitimacy—such as concerns about algorithmic fairness and individual moderator's political and personal biases—is essential to designing better online platforms and improving online content moderation.
- We conducted a survey on people's views of Facebook's content moderation processes and found that participants perceive expert panels as a more legitimate content moderation process than paid contractors, algorithmic decision-making, or digital juries.
- Responses from participants also showed a clear distinction between impartiality and perceived legitimacy of moderation processes. Although participants considered algorithms the most impartial process, algorithms had lower perceived legitimacy than expert panels.

moderation decisions and randomized the description of whether paid contractors, algorithms, expert panels, or juries of users made those decisions. Their responses, given this information, provide a window into how individuals perceive the legitimacy of moderation decisions.

We also studied whether the decision itself—and respondents' agreement with it—shaped their answers. The more social media companies' content moderation policies shape popular discourse, and the more algorithms play a role in that moderation, the more essential it is to understand how to make those content moderation processes as legitimate as possible.

Introduction

Improving platform design will not boost user trust if users distrust the platform itself—or its processes. Criticism leveled at social media companies has described content moderation policies and processes as opaque, unrepresentative, and occurring without meaningful oversight. Facebook, YouTube, Twitter, and other platforms have been subject to these critiques at various points. All told, policymakers, scholars, and the general public are clearly concerned about how they view social media companies' content moderation processes.

When the public believes institutions are highly legitimate, they are more likely to accept unpopular decisions and to cooperate and comply with those institutions; much empirical and sociological research bears this out. Legitimacy can be understood from a normative or descriptive perspective. Constitutional legitimacy, democratic legitimacy, and other normative frameworks understand legitimacy as "some benchmark of acceptability or justification of political power or authority and—possibly—obligation." Descriptive legitimacy refers to the acceptance of authority by people. More commonly called perceived legitimacy, it is a measurable phenomenon. We use perceived legitimacy here that refers to how people view an *organization's* legitimacy. This definition comprises five parts: satisfaction (with how the organization handled the decision), trustworthiness (of the organization), fairness and impartiality (of the organization), commitment to continue having the organization, and belief in maintaining the scope of the organization's decision-making powers (i.e., decisional jurisdiction).

When the public believes institutions are highly legitimate, they are more likely to accept unpopular decisions and to cooperate and comply with those institutions.

Our paper focuses on the perceived legitimacy of social media content moderation processes. We did not study community or artisanal moderation, such as when a platform relies on its community members to moderate content (e.g., volunteer moderators at Reddit). Instead, we limited our scope to corporate moderation practices since they operate at a large scale. In particular, we looked at paid individual

contractors (whom the company hires and trains), automated systems (often, databases combined with machine learning algorithms), digital juries (ad hoc groups of users), and expert panels (experts from content moderation, law, human rights, journalism, and other fields). The first two—paid individual contractors and automated systems—are widely used. The last two, digital juries and expert panels, are emerging features of social media content moderation.

We then compiled Facebook posts covering a wide range of topics common in takedown decisions—racism, protest, vaccination, electoral fraud, and more—with liberal and conservative viewpoints. The posts we picked could all be viewed as violating Facebook restrictions on content inciting violence, hate speech, and misinformation. We narrowed the group to nine posts and sent them to 100 U.S. Facebook users in a survey distributed through the Amazon Mechanical Turk platform.

Each participant was presented with four moderation decisions. These "decisions" each contained one of the Facebook posts, one of the four moderation processes, a random decision outcome (kept up or taken down), and an indication of which content restriction it (supposedly) violated. The individuals then shared their perceived legitimacy of the moderation process and were asked to select which process they saw as most trusted, least trusted, most fair and impartial, and least fair and impartial.

Research Outcomes

Overall, respondents believed that expert panels were more legitimate than digital juries and algorithms. Beyond that, there was no clear legitimacy ranking between the other moderation options (digital juries, algorithms, and paid contractors). However, the majority of respondents (51%) said an algorithm was the most impartial moderation process. Several of these respondents (52%) said it was because algorithms make decisions based on logic, data, and rules. Many participants (32%) who ranked algorithms as the most trustworthy moderation process said the same. However, this was not unanimous. One quarter of respondents made comments such as, "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." Further, many of those who ranked algorithms as the most trustworthy included caveats: Respondents said that the algorithm's trustworthiness depended on such factors as the algorithm being constructed fairly and impartially, the decisions being subject to checks and balances, and the opportunity for human appeal.

Overall, respondents believed that expert panels were more legitimate than digital juries and algorithms.

These findings are consequential for several reasons. Legitimacy is not the same as impartiality. Even though algorithms were ranked as the most impartial moderation process, expert panels still received the highest score on legitimacy overall.

This finding expands the knowledge gained from previous research showing that positive views of algorithmic decision-making depend on factors like the subjectivity of the domain, the opaqueness of an algorithm's function and deployment, and the algorithm's performance.

Individuals also placed conditions on their votes of perceived legitimacy—such as wanting a moderation process to have adequate checks and balances—which indicates there is more complexity to people's views of content moderation policies than a yes/no binary. Further, the data showed that users will say they perceive content moderation processes as more legitimate when they personally agree with the content moderation outcome. We did not assess whether or not users realized this fact, but the takeaway is worth noting.

Our study had several limitations. For example, it did not examine perceptions of rule creation, only perceptions of rule enforcement. In addition, it only focused on perceptions of Facebook in the United States. It also did not interrogate how respondents felt about the timing of when moderation is carried out—whether at the point of posting or after a post has been reported.

Policy Discussion

The perceived legitimacy of content moderation processes is an important question for policymakers—and it informs how policymakers themselves think about social media.

Expert panels are still an emerging feature of social media content moderation. Nonetheless, respondents suggested that using expert panels was the most legitimate process compared to paid contractors, algorithms, and digital juries of platform users. As one respondent put it, "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."

Some respondents in our survey presumed that experts would have a liberal bias; there were other respondents who presumed that digital juries would be more tolerant of harmful content.

We would note, though, that expert panels still have many limits. The Facebook Oversight Board's rulings about President Trump's posts following the January 6 attack on the Capitol are a prime example. Panel decisions are influenced by the panel's composition, and members of the Oversight Board with legal and judicial backgrounds drove the Board to couch its decisions in legal verbiage and reasoning. In addition, expert panels might prioritize their short-term legitimacy and adopt a middle-of-the-road, noncontroversial position merely to avoid criticism. Moreover, the use of expert panels can move important societal questions—like how international human rights law applies to content moderation—internal to a company, diminishing legitimacy and transparency.

Policymakers also cannot ignore the role of political views in understanding how people view social media platforms. Some respondents in our survey presumed that experts would have a liberal bias; there were other respondents who presumed that digital juries would be more tolerant of harmful content. This expands on prior work showing that liberals and conservatives, generally speaking, place different values on components of perceived legitimacy, like fairness.

We recommend that social media companies incorporate expert panels into their content moderation decisions. The first step, however, would be to form a publicly visible, independent panel that develops moderation guidelines and handles appeals of the platform's most controversial cases. The panel could help to train rank-and-file moderators, assess how algorithms are used to moderate content, and educate the public about moderation policies and practices. Critical to this effort—and something policymakers must monitor—is whether social media companies build these panels with sets of diverse and representative individuals.

Transparency is particularly important to perceived legitimacy. Research into procedural justice suggests that explaining to people how decisions are reached, and giving them the opportunity to express their views and opinions (e.g., during an appeal), can boost their perceptions of procedural fairness—regardless of the actual outcome.

But the resounding perception that expert panels are a legitimate form of content moderation, and the finding that individuals were more likely to view a process as legitimate when they agreed with the outcome, provide important guideposts for policymakers. Content moderation challenges are not going away. Designing better online speech platforms and improving content moderation processes requires a deeper study of how people view content moderation itself.

The original article is accessible at Christina A. Pan et al., "**Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries**," *Proceedings of the ACM on Human-Computer Interaction 6*, no. CSCW1 (April 2022), https://hci.stanford.edu/publications/2022/ComparingPerceivedLegitimacy.pdf.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Christina A. Pan has an M.S. in computer science from Stanford University. She has also conducted research in human-centered AI.

Sahil Yakhmi has an M.S. in computer science and an M.B.A. from Stanford University.

Tara P. Iyer has an M.S. in computer science from Stanford University.

Evan Strasnick is a research scientist at Reality Labs, Meta, and has a Ph.D. in human-computer interaction from Stanford University.

Amy X. Zhang is an assistant professor of computer science and engineering at the University of Washington, Seattle.

Michael S. Bernstein is an associate professor of computer science at Stanford University and a faculty affiliate at the Stanford Institute for Human-Centered Artificial Intelligence.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**