Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms

Jie Cai
Pennsylvania State University
University Park, USA
jie.cai@psu.edu

Aashka Patel
New Jersey Institute of Technology
Newark, USA
asp263@njit.edu

Azadeh Naderi
New Jersey Institute of Technology
Newark, USA
an57@njit.edu

Donghee Yvette Wohn
New Jersey Institute of Technology
Newark, USA
yvettewohn@gmail.com

ABSTRACT

Social media users may perceive moderation decisions by the platform differently, which can lead to frustration and dropout. This study investigates users' perceived justice and fairness of online moderation decisions when they are exposed to various illegal versus legal scenarios, retributive versus restorative moderation strategies, and user-moderated versus commercially moderated platforms. We conduct an online experiment on 200 American social media users of Reddit and Twitter. Results show that retributive moderation delivers higher justice and fairness for commercially moderated than for user-moderated platforms in illegal violations; restorative moderation delivers higher fairness for legal violations than illegal ones. We discuss the opportunities for platform policymaking to improve moderation system design.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

Social Media, Content Moderation, Justice and Fairness, Platform Governance, Policymaking

ACM Reference Format:
Jie Cai, Aashka Patel, Azadeh Naderi, and Donghee Yvette Wohn. 2024. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613905.3650882

1 INTRODUCTION

To fight harmful content and maintain a safe online space, platforms use algorithms to automatically or human labor to remove harmful content and sanction offenders manually [14], which is termed "content moderation." Content moderation is defined as "the practice of screening of user-generated content (UGC) posted to Internet sites, social media platforms and other online outlets, to determine the appropriateness of the content for a given site, locality, or jurisdiction" [33]. Social media platforms have developed complex and platform-specific content moderation policies to regulate harmful content [32]. This paper uses the term "policies" to indicate formalized statements such as rules, standards, terms of services, and community guidelines. These content moderation policies guide human moderators and algorithms to remove offensive content and sanction offenders [6]. Many users complain their content has been removed by the platform without a clear explanation and frequently express confusion about what action has triggered a moderation action or account suspension [11, 23, 24], leading users to feel frustrated and leave [30].

As moderation strategies vary across different contexts and platforms, users may perceive different levels of justice and fairness in the moderation decision. Many HCI scholars have explored the platform's moderation mechanisms from various perspectives, such as human labor [40] and transparency [25, 27]. From the end-user perspective, some work explores the bystander effect [1], and others explore victims of harmful content [36, 45]. For example, recent work shows that explanation of content removal increases a user's perceived fairness and engagement in the community again, and a user's perceived education and interaction are more effective than blocking [4], suggesting that removal explanation is positively associated with users' perceived justice and fairness.

In line with working to understand the user's perception of moderation decisions, we ran an online experiment to understand how the explanation in different contexts with different harmful content can influence users' perception of fairness and justice. We extend the line of research about justice and fairness in content moderation. The findings can potentially benefit social media platforms by furthering their understanding of how their users perceive moderation decisions and enabling them to better current moderation practices with policymaking.

2 RELATED WORK

2.1 Dependent Variables: Justice and Fairness

Computing systems are embedded with different biases regarding context, user populations, and technical constraints, further leading

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650882

to unfairness and injustice in use [13, 39]. Content moderation as a sociotechnical process can privilege normalized groups while pushing other user groups with different races, genders, and religions, to the margins [10].

The justice lens from criminal justice systems has been applied to the content moderation domain; if users perceive moderation decisions to lack justice and fairness, they are likely to stop abiding by the rules of the platforms or even seek their own ways to punish the offenders, resulting in punishment that may be indeterminate, uncalibrated, or inaccurate [1]. Moderation decisions considering different contexts and offender's characteristics can potentially increase the perceived fairness and justice [5, 35]. As such, in order to keep online communities safe and civil, we aim to assess users' perceptions of justice and fairness on social media platforms.

Justice is the perceived adherence to rules that reflect appropriateness in decision contexts [8], including consistency, accuracy, bias suppression, and correctability [42]. Justice has been discussed broadly in online harassment and platform governance [5, 20, 34]. In our study, we center the general users'/bystanders' perspective of justice in content moderation. We view it as adherence to rules of conduct regarding online content violation and moderation strategies applied. Fairness is an individual's moral evaluation of the rules of conduct [7]. Fairness is an important factor for online content moderation. For instance, providing explanations for content removal can increase an offender's perceived fairness and desire to keep participating in the community [17]. In this study, users' perceived fairness of moderation decisions is the users' moral evaluations of rule enforcement regarding online content violations and moderation strategies.

2.2 Independent Variables: Platform Types, Violation Types, Moderation Types

2.2.1 Commercially moderated Versus User-moderated Platforms. While specific platforms practice a bottom-up governance model that relies on community members to enforce policies, others practice top-down governance in which "officials implement a relatively detailed set of rules over a given community" [3]. Examples of such platforms include Facebook and Twitter, where online harassment is dealt with centrally rather than relying on volunteer moderators [6]. Commercial content moderators for these platforms are paid and contingent [33]. Users of such platforms believe that the company should bear more responsibility for content moderation rather than having the responsibility fall on the users themselves [31]. On the other hand, with user-moderated platforms such as Twitch, creators of the user-generated content have moderating privileges and can appoint their followers to be additional moderators [44]. Often, volunteer moderators find personal meaning in their roles [37] and want to strengthen their online communities by guiding and developing offenders rather than simply "cleaning up" misbehavior [38, 44]. Overall, there are apparent differences between how commercial and user moderators treat offenders [9]; consequently, the different treatment may affect users' perceived justice and fairness of the punishment. As such, we ask our research question:

- 1) Will users' perceived justice and fairness of online moderation decisions be higher or lower for commercially moderated versus user-moderated social media platforms?

2.2.2 Illegal versus Legal Content Scenarios. Social media platforms face increasing pressure from users and lawmakers to "clean up their platforms" [22]. The legality of content posted online matters for some users, and these users generally believe that platforms should not have the power to remove this content "as long as their posts are not illegal and do not incite illegal assembly, destruction of property or violence" [29]. This study considers illegal violations as content-inducing crimes or public safety concerns. Users believe social media platforms should have the ability and fairness to remove and moderate illegal material, such as child pornography, from their platforms [19, 21].

Online abuse such as racial slurs, bullying, sexual harassment, spam, trolls, and hate speech toward a specific group or individuals are also considered violations [2, 19]. This abuse is mainly handled by the platform or even the specific entities like end-users and human moderators [6]. In this study, by community guidelines of various social media platforms, we consider this online abuse targeting a specific group or individual without severe public impact as a legal violation. These violations are broadly defined and contingent on different platforms and communities. Because the belief that illegal content should be moderated on social media platforms echoed in existing literature and media articles, we developed the following hypotheses:

- H2a. Perceived justice is higher in illegal compared to legal content scenarios.
- H2b. Perceived fairness is higher in illegal compared to legal content scenarios.

2.2.3 Restorative Versus Retributive Moderation Strategies. Retributive justice is "a theory of punishment in which individuals who knowingly commit an act deemed morally wrong receive a proportional punishment for their misdeeds" [1]. Restorative justice is "a process whereby all the parties with a stake in a particular offense come together to collectively resolve how to deal with the aftermath of the offense and its implications for their future" [28]. Generally, platforms remove offensive content and the offender from their communities. This content moderation practice is believed to echo the American criminal justice system and retributive justice [15]. Through restorative justice, the offenders are meant to acknowledge their wrongdoings, accept responsibility for their transgressions, and demonstrate remorse [35]. As such, through restorative content moderation, social media platforms can build healthier, resilient, and long-term online communities [15].

We adapt the retributive and restorative perspectives and divide the moderation strategies into retributive and restorative moderation, similar to the moderation styles punishing and nurturing [18]. Retributive moderation strategies in this study refer to banning offenders for rule-breaking and incapacitating offenders' community participation. Restorative moderation strategies in this study refer to light punishment compared with the same violation (e.g., warning offenders with rule explanation and potential consequences if offenders keep behaving similarly) to maintain the community.

Users of social media platforms may prefer one form of moderation strategy over the other. As such, we developed the following hypotheses:

- H3a. Perceived justice is higher for retributive compared to restorative moderation strategies.
- H3b. Perceived fairness is higher for retributive compared to restorative moderation strategies.

3 METHODS AND MATERIALS

We choose Twitter (now "X") and Reddit as our research context. Twitter is a typical platform applying commercial moderation, while Reddit is known for user moderation. Both platforms provide thread-style communication with sharing and commenting features. The similar affordances and the different moderation policies indicate they fit well for design experiments with various scenarios.

3.1 Online Experiment Deployment

We conducted an online experiment with participants recruited from Qualtrics. We limited the recruitment of participants to the United States to ensure that all participants had a similar understanding of the topic. We also set the gender as 50% male and 50% female to control the gender bias. Our sample size was 200 (100 with either legal or illegal scenarios). To ensure our sample was as representative as possible, we asked Qualtrics to set these quota guidelines. If potential participants were under 18 and did not utilize Twitter and/or Reddit, they were terminated from the online experiment. If online experiment takers answered either "I will not provide my best answers" or "I can't promise either way" to our question "Do you commit to providing your thoughtful and honest answers to the questions in this online experiment?", they were also terminated. In the consent form at the beginning of the survey, we stated that participants would be exposed to violent and controversial content with relevant mental health resources to mitigate the negative impact of this study. The median time for participants to complete the online experiment was 7.4 minutes. If participants completed the online experiment, they were compensated per their agreement with Qualtrics, the panel provider. We paid Qualtrics $5.00 (USD) for each participant. Our participants were 18 to 65 or older, mostly between 30-49 (50.5%) and White (69%).

3.2 Experimental Design

Our online experiment used a between-subjects design by legality (Group 1 illegal, Group 2 legal). Then, we used a within-subjects design for each group with four scenarios (1a, 1b, 1c, 1d, 2a, 2b, 2c, 2d). We showed each group a total of 4 mock scenarios designed by the team, as shown in Appendix A. The first group (Group 1) of participants viewed four scenarios in which illegal content was posted to Reddit and Twitter. The second group viewed four scenarios where legal content was posted to Reddit and Twitter, as shown in Table 1. We ran a series of mixed ANOVA with Bonferroni post-hoc tests to answer RQ1 and test our main hypotheses, H2s and H3s.

Before creating the scenarios, the team read Twitter's Community Guidelines and Reddit's Content Policy. We did this to create illegal and legal content from scratch that would violate the two platforms' policies and would likely be removed from the platform. Team members wrote the illegal and legal content of all eight scenarios for the online experiment design. We presented the scenarios in the lab meeting to collect feedback and modify them. The illegal scenarios are consistently about terrorism inducing public safety concerns, and legal scenarios are consistently about sexism and racism targeting a specific group or individual. Each scenario states either a Reddit or Twitter user posted the message or tweet to either Reddit or Twitter. The scenario then states the message or tweet the fake Reddit or Twitter user posted. Following this, the scenario states the removal from the platform by the moderator on the respective platforms for violating the platform's policy. The scenario states the policy that the message or tweet violated. Additionally, the scenario states the moderator either warned the user or permanently banned the user. After each scenario, participants answered questions about perceived justice and fairness. Scenarios and survey questions are in Appendix B.

3.3 Measurements

The participants' perception of justice for each scenario was measured with a single item: "Is it necessary to punish the (Reddit or Twitter) user for their post to deliver justice?". Participants answered on a 5-point Likert scale from 1 (Absolutely Not Necessary) to 5 (Absolutely Necessary). This item was extracted and adapted from the Punishment Orientation Questionnaire [46]. The participants' perception of fairness for each scenario was measured with a single item: "How fair do you perceive the (Reddit or Twitter) moderator's decision to be?". Participants answered on a 5-point Likert scale from 1 (very unfair) to 5 (very fair).

4 RESULTS

4.1 Perceived Justice

Mauchly's Test of Sphericity indicated that sphericity was met (W = .98, $\chi^2(5)$ = 4.02, p = .547). The tests of within-subjects effects show a significant main effect among the scenarios ($F(3,594)$ = 21.45, p < .001, η^2 = .10). There was also a significant interaction between scenarios and groups ($F(3,594)$ = 21.05, p < .001, η^2 = .13). The partial eta squared indicated that interaction had a stronger impact than the main effect on justice. All the descriptive statistics for each scenario and the p values between groups are shown in Table 2. The interaction effect is shown in Figure 1a.

RQ1 tries to understand if users' perceived justice was higher or lower for user-moderated platforms (i.e., Reddit) versus commercially moderated platforms (i.e., Twitter); we need to specifically compare scenario_1a and 1b, scenario_1c and 1d, scenario_2a and 2b, and scenario_2c and 2d. The Bonferroni posthoc tests with simple effects showed a perceived justice difference in scenario_1c and 1d (p < .001). Overall, perceived justice is higher for commercially moderated than for user-moderated platforms in illegal scenarios with retributive moderation strategies.

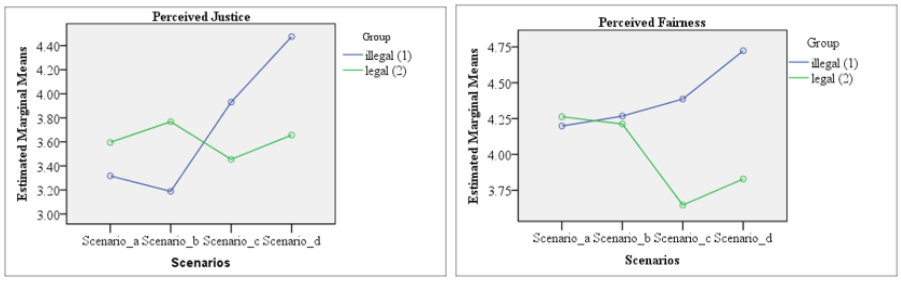
H2a stated that perceived justice is higher in illegal than legal scenarios. The tests of between-subjects effects showed that, in general, there is no significant difference in perceived justice between illegal and legal scenarios (p = .487). Bonferroni posthoc tests comparing justice in scenarios revealed a significant difference in scenarios_b, c, and d; specifically, in scenario_b, perceived justice was lower in the illegal group compared to the legal group (p =

Table 1: Scenario Descriptions in the Experiment

Scenario 1a	Scenario 1b	Scenario 1c	Scenario 1d
a restorative moderation strategy for illegal content on a user-moderated platform	a restorative moderation strategy for illegal content on a commercially moderated platform	a retributive moderation strategy for illegal content on a user-moderated platform	a retributive moderation strategy for illegal content on a commercially moderated platform
Scenario 2a	Scenario 2b	Scenario 2c	Scenario 2d
a restorative moderation strategy for legal content on a user-moderated platform	a restorative moderation strategy for legal content on a commercially moderated platform	a retributive moderation strategy for legal content on a user-moderated platform	a retributive moderation strategy for legal content on a commercially moderated platform

Table 2: Descriptive Statistics for Each Group and Scenario in Terms of Perceived Justice

	Justice-a	Justice-b	Justice-c	Justice-d	Justice-overall
Group1: Illegal	Scenario_1a M = 3.32, SD = 1.30	Scenario_1b M = 3.19, SD = 1.29	Scenario_1c M = 3.93, SD = 1.35	Scenario_1d M = 4.48, SD = 1.10	M = 3.73, SE = 0.11
Group2: Legal	Scenario_2a M = 3.60, SD = 1.33	Scenario_2b M = 3.77, SD = 1.34	Scenario_2c M = 3.45, SD = 1.49	Scenario_2d M = 3.66, SD = 1.44	M = 3.62, SE = 0.11
	p = .136	p = .002	p = .019	p < .001	p = .487



(a) Justice **(b) Fairness**

Figure 1: Interaction between scenarios and groups regarding perceived justice and fairness.

.002); in scenario_c, perceived justice is higher in the illegal group compared to the legal group (p = .019); and in scenario_d, perceived justice is higher in the illegal group compared to the legal group (p < .001). Overall, the perceived justice was higher in illegal scenarios compared to legal scenarios with retributive moderation strategies and had either no difference or was lower in illegal scenarios compared to legal scenarios with restorative moderation strategies. Thus, H2a was partially supported.

H3a stated that perceived justice was higher for retributive moderation strategies than restorative moderation strategies. Specifically, we compare scenario_a and c to see the different moderation strategies on commercially moderated platforms and compare scenario_b and d to see the different moderation strategies on user-moderated platforms. The simple effects for the interaction revealed that the perceived justice increased significantly from scenario_a to scenario_c in the illegal group (p < .001) but showed no significant difference in the legal group (p = 1.00). The simple effects for the interaction revealed that the perceived justice increased significantly from scenario_b to scenario_d in the illegal group (p < .001) but showed no significant difference in the legal group (p = 1.00). Overall, the perceived justice was higher for retributive compared to restorative moderation strategies in the illegal group but had no difference in the legal group. H3a was partially supported. In other words, in illegal scenarios, retributive moderation is more just; in legal scenarios, there is no justice difference regarding retributive/restorative moderation.

4.2 Perceived Fairness

Mauchly's Test of Sphericity indicated that sphericity was not met (W = .79, $\chi^2(5)$ = 46.50, p < .001). The tests of within-subjects effects show a significant main effect among the scenarios (F(2.59, 512.25) = 4.16, p = .009, η^2 = .02). There was also a significant interaction between scenarios and groups (F(2.59, 512.25) = 17.38, p < .001, η^2 = .08). The partial eta squared indicated that interaction had a stronger impact than scenarios on fairness. All the descriptive statistics for each scenario and the p values between groups are shown in Table 3. The interaction effect is shown in Figure 1b.

Table 3: Descriptive Statistics for Each Group and Scenario in Terms of Perceived Fairness

	Fairness-a	Fairness-b	Fairness-c	Fairness-d	Fairness-overall
Group1: Illegal	Scenario_1a M = 4.20, SD = 1.00	Scenario_1b M = 4.27, SD = 0.96	Scenario_1c M = 4.39, SD = 0.99	Scenario_1d M = 4.72, SD = 0.74	M = 4.39, SE = 0.08
Group2: Legal	Scenario_2a M = 4.26, SD = 0.89	Scenario_2b M = 4.21, SD = 1.01	Scenario_2c M = 3.65, SD = 1.39	Scenario_2d M = 3.83, SD = 1.33	M = 3.99, SE = 0.08
	p = .630	p = .692	p < .001	p < .001	p < .001

RQ1 tries to understand if users' perceived fairness was higher or lower for user-moderated platforms versus commercially moderated platforms. We specifically compare scenario_1a and 1b, scenario_1c and 1d, scenario_2a and 2b, and scenario_2c and 2d. The Bonferroni posthoc tests with simple effects showed only a perceived fairness difference in scenario_1c and 1d (p < .001). Overall, perceived fairness is higher for commercially moderated platforms than for user-moderated platforms in illegal scenarios with retributive moderation strategies.

H2b stated that perceived fairness was higher in illegal than legal scenarios. The tests of between-subjects effects showed that, in general, there was a significant difference in perceived fairness between illegal and legal scenarios (p < .001). Bonferroni posthoc tests comparing fairness in scenarios revealed a significant difference in scenarios_c and d; specifically, in scenario_c, perceived fairness is higher in the illegal group compared to the legal group (p < .001), and in scenario_d, perceived fairness is higher in the illegal group compared to the legal group (p < .001). Overall, the perceived fairness was higher in illegal scenarios compared to legal scenarios with retributive moderation strategies, and there was no difference in illegal scenarios compared to legal scenarios with restorative moderation strategies. Thus, H2b was partially supported. In other words, moderating illegal scenarios was fairer than legal scenarios in scenarios with retributive moderation, but there was no fairness difference regarding legality in scenarios with restorative moderation.

H3b stated that perceived fairness is higher for retributive than restorative moderation strategies. We compared scenario_a and c to see the different moderation strategies on commercially moderated platforms and compared scenario_b and d to see the different moderation strategies on user-moderated platforms. The simple effects revealed that perceived fairness showed no significant difference in the illegal group (p = .760) but decreased significantly in the legal group (p < .001) from scenario_a to c. The simple effects revealed that the perceived fairness increased significantly from scenario_b to d in the illegal group (p < .001) but decreased significantly in the legal group (p = .012). Overall, the perceived fairness was higher or no difference for retributive compared to restorative moderation strategies in the illegal group but lower in the legal group. H3b was partially supported. In other words, retributive moderation is fairer in illegal scenarios, but in legal scenarios, restorative moderation is fairer.

5 DISCUSSION

5.1 Not Only Restorative But Retributive Moderation Can also Improve Justice and Fairness

Recent research proposes to move from retributive justice to restorative justice to mediate the harm and resolve the conflict between violators and victims [34]. This line of research criticizes that retributive justice can even cause severe sequential harm to the stakeholders [35, 36]. However, they often consider online toxicity as a whole and do not consider the nuanced difference of violation types. In this study, we separate violation from the legality perspective. Our results consistently show that users consider retributive moderation strategies to deliver higher justice and fairness for illegal violations than legal ones, and restorative strategies deliver higher justice for legal violations than illegal ones. In this sense, we contribute to a nuanced understanding of violation and perceived justice differences. First, we supplement the line of restorative justice research [35] and point out that it is more about the legal violation related to online toxicity and harassment.

Second, retributive moderation is still preferred to deal with illegal violations. Users still think it is fair for illegal violations to receive severe punishment, such as terrorism and public violence. Speculatively, an illegal violation is more likely to cause severe consequences for public safety or society, like content inflaming civil unrest or terrorism. These types of violations should be punished severely online or even have the police involved offline. Legal violation is more likely to cause psychological harm for individual entities, like harassment towards a specific group related to gender, race, identity, or disability [23, 43]. Consequently, we suggest that both types of moderation strategies should be incorporated into the platform's moderation policymaking, and it is difficult to weigh which one is better and should be emphasized. Policymakers should try to differentiate the types of violations. Though the potential legal violation can sometimes lead to illegal violations at scale, it is essential to explore the transition between these two types of violations further and identify the appropriate boundary to intervene with relevant agencies.

We can also get some clues from the tech giants' moderation policies about their different attitudes toward legal or illegal violations. For example, Facebook, Microsoft, Twitter, and YouTube established the "Global Internet Forum to Counter Terrorism" in 2017 to coordinate content removal about "violent terrorist imagery and propaganda". However, there is little or no collaboration about daily online harassment. Different platforms have different policies regarding online harassment, such as community guidelines and

codes of conduct. Such situations also cause challenges from the policymaking at the platform level and hinder platform collaboration. It seems there are no clear collaborative solutions for the legal violations since different platforms hold different values, such as the Stormfront white nationalist website [16]. This study also sheds light on collaborative policymaking to deal with legal violations from some commonly shared values for most platforms, such as racism and sexism, with zero tolerance.

5.2 Explanation Can Improve Justice and Fairness in Certain Scenarios But Not All

In line with work about moderation explanation increasing perceived justice and fairness [17], we extend and show how the explanation of moderation decisions across different platforms affects users' perceived justice and fairness. Regarding illegal violations, retributive strategies on commercially moderated platforms deliver higher justice and fairness than user-moderated platforms. Many scholarships have highlighted the importance of moderation transparency to improve justice and fairness [12, 41]. Most of them focus on one platform's moderation policies. For example, many criticize that commercial moderation lacks transparency and that community moderation should keep increasing transparency with clear rules and norms with active engagement [24, 26]. We contribute to the comparison of these types of moderation with an explanation. Users' perception of the platform difference after viewing explanations also plays a significant role in perceived justice and fairness only in illegal violations with retributive moderation. Offering explanations with restorative moderation strategies seems to have no platform difference, even retributive moderation under legal scenarios. In this sense, we provide a nuanced understanding of the moderation strategies in different scenarios with platform differences and show that transparency is essential to justice and fairness in a specific situation, but not all. Such findings supplement prior work regarding the tension of punishing and nurturing [18] and offer naunaced differences to consider moderation resource allocation. For example, with scenarios where the explanation makes no or little difference, the platform should invest less human labor and resources.

6 LIMITATION AND FUTURE WORK

Although this study has important implications for content moderation policies, it is subject to a few limitations. Firstly, since many people in the US come from diverse cultures, their cultural biases may be influenced by living or growing up in the US. Thus, people from the same culture in the US may hold different cultural views compared to those in their country of origin. Therefore, future studies should take into account the cultural differences in users of different countries [19] and investigate how different perceptions of social media use in other regions affect content moderation. Secondly, we used a single-item measurement for fairness and justice, whereas most other fairness and justice research employs qualitative methods. It is crucial for HCI scholars to develop a scale for future content moderation studies. Additionally, the preliminary study did not explore action variables, so it would be beneficial to investigate how users would like platforms to take action regarding content moderation to make their decisions more just and fair. Last, there are also potential opportunities to explore how to incorporate the policy into moderation system design [47] and to reshape the governance and platform identity to make the platform inclusive and diverse [10].

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation (Award No. 1928627).

REFERENCES

[1] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). https://doi.org/10.1609/icwsm.v12i1.15036 Number: 1.

[2] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–19. https://doi.org/10.1145/3134659

[3] Ben Bradford, Grisel, Florian, Meares, Tracey L, Hamilton, Walton Hale, School, Yale Law, Owens, Emily, Pineda, Baron L, Shapiro, Jacob N, Tyler, Tom R, Fleming, Macklin, Law, Yale, Danieli, School, and Peterman, Evans. 2019. *Report Of The Facebook Data Transparency Advisory Group.* Technical Report. Yale Law School, The Justice Collaboratory. 44 pages. https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report

[4] Jie Cai and Donghee Yvette Wohn. 2019. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '19 Companion).* Association for Computing Machinery, New York, NY, USA, 166–170. https://doi.org/10.1145/3311957.3359478

[5] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 410:1–410:25. https://doi.org/10.1145/3479554

[6] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences (IMX '21).* Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3452918.3458796

[7] Jason A. Colquitt and Jessica B. Rodell. 2015. Measuring Justice and Fairness. (July 2015). https://doi.org/10.1093/oxfordhb/9780199981410.013.0008

[8] Jason A. Colquitt and Kate P. Zipay. 2015. Justice, Fairness, and Employee Reactions. *Annual Review of Organizational Psychology and Organizational Behavior* 2, 1 (2015), 75–99. https://doi.org/10.1146/annurev-orgpsych-032414-111457 _eprint: https://doi.org/10.1146/annurev-orgpsych-032414-111457.

[9] Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. *Frontiers in Human Dynamics* 3 (Feb. 2021), 626409. https://doi.org/10.3389/fhumd.2021.626409

[10] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. "Jol" or "Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 473:1–473:25. https://doi.org/10.1145/3479860

[11] Houda Elmimouni, Yarden Skop, Norah Abokhodair, Sarah Rüller, Konstantin Aal, Anne Weibert, Adel Al-Dawood, Volker Wulf, and Peter Tolmie. 2024. Shielding or Silencing?: An Investigation into Content Moderation during the Sheikh Jarrah Crisis. *Proceedings of the ACM on Human-Computer Interaction* 8, GROUP (Feb. 2024), 1–21. https://doi.org/10.1145/3633071

[12] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics* 26, 6 (Dec. 2020), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

[13] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

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

[15] Amy A. Hasinoff, Anna D. Gibson, and Niloufar Salehi. 2020. The promise of restorative justice in addressing online harm. https://www.brookings.edu/techstream/the-promise-of-restorative-justice-in-addressing-online-harm/

[16] Libby Hemphill. 2022. *Very Fine People: What Social Media Platforms Miss About White Supremacist Speech.* Technical Report. The Anti-Defamation League. https://www.adl.org/resources/report/very-fine-people

[17] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 192:1–192:33. https://doi.org/10.1145/3359294

[18] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (March 2023), 3:1–3:34. https://doi.org/10.1145/3534929

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

[20] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 334:1–334:21. https://doi.org/10.1145/3476075

[21] Kyle Langvardt. 2018. Regulating Online Content Moderation. *Georgetown Law Journal* 106, 5 (June 2018), 1353–1389. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00168092&v=2.1&it=r&id=GALE%7CA548321177&sid=googleScholar&linkaccess=abs Publisher: Georgetown University Law Center.

[22] Kalev Leetaru. 2018. Is Social Media Content Moderation An Impossible Task? https://www.forbes.com/sites/kalevleetaru/2018/09/08/is-social-media-content-moderation-an-impossible-task/ Section: AI & Big Data.

[23] Yao Lyu, Jie Cai, Anisa Callis, Kelley Cotter, and John M. Carroll. 2024. "I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3642148 arXiv:2401.11663 [cs].

[24] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 425:1–425:28. https://doi.org/10.1145/3555150

[25] Renkai Ma and Yubo Kou. 2023. "Defaulting to boilerplate answers, they didn't engage in a genuine conversation": Dimensions of Transparency Design in Creator Moderation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 44:1–44:26. https://doi.org/10.1145/3579477

[26] Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581097

[27] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 278:1–278:30. https://doi.org/10.1145/3610069

[28] H. Messmer and H. U. Otto (Eds.). 1992. *Restorative Justice on Trial: Pitfalls and Potentials of Victim-Offender Mediation — International Research Perspectives —*. Springer Netherlands. https://doi.org/10.1007/978-94-015-8064-9

[29] Peter Morici. 2021. Facebook and Twitter should not be in the censorship business. https://www.marketwatch.com/story/facebook-and-twitter-should-not-be-in-the-censorship-business-11611871551 Section: Economy & Politics.

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

[31] Aashka Patel, Christine L. Cook, and Donghee Yvette Wohn. 2021. User Opinions on Effective Strategies Against Social Media Toxicity. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2021.366

[32] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, Sanibel Island Florida USA, 369–374. https://doi.org/10.1145/2957276.2957297

[33] Sarah T. Roberts. 2017. *Content moderation*. https://escholarship.org/uc/item/7371c1hf

[34] Sarita Schoenebeck and Lindsay Blackwell. 2021. *Reimagining Social Media Governance: Harm, Accountability, and Repair*. SSRN Scholarly Paper ID 3895779. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3895779

[35] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *New Media & Society* 23, 5 (May 2021), 1278–1300. https://doi.org/10.1177/1461444820913122

[36] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 2:1–2:18. https://doi.org/10.1145/3449076

[37] Joseph Seering and Sanjay R. Kairam. 2022. Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (Dec. 2022), 18:1–18:18. https://doi.org/10.1145/3567568

[38] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316 Publisher: SAGE Publications.

[39] Jessie J. Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. 2023. The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1652–1663. https://doi.org/10.1145/3593013.3594106

[40] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445092

[41] Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13, 0 (March 2019), 18. https://ijoc.org/index.php/ijoc/article/view/9736 Number: 0.

[42] J. Thibaut and L. Walker. 1976. Procedural Justice: A Psychological Analysis. https://doi.org/10.2307/448155

[43] Jirassaya Uttarapong, Jie Cai, and Donghee Yvette Wohn. 2021. Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences (IMX '21)*. Association for Computing Machinery, New York, NY, USA, 7–19. https://doi.org/10.1145/3452918.3458794

[44] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

[45] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3517614

[46] Susan Yamamoto and Evelyn M. Maeder. 2019. Creating the Punishment Orientation Questionnaire: An Item Response Theory Approach. *Personality and Social Psychology Bulletin* 45, 8 (Aug. 2019), 1283–1294. https://doi.org/10.1177/0146167218818485

[47] Qian Yang, Richmond Y. Wong, Thomas Gilbert, Margaret D. Hagan, Steven Jackson, Sabine Junginger, and John Zimmerman. 2023. Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3573827

A EXPERIMENTAL DESIGN PROCESS

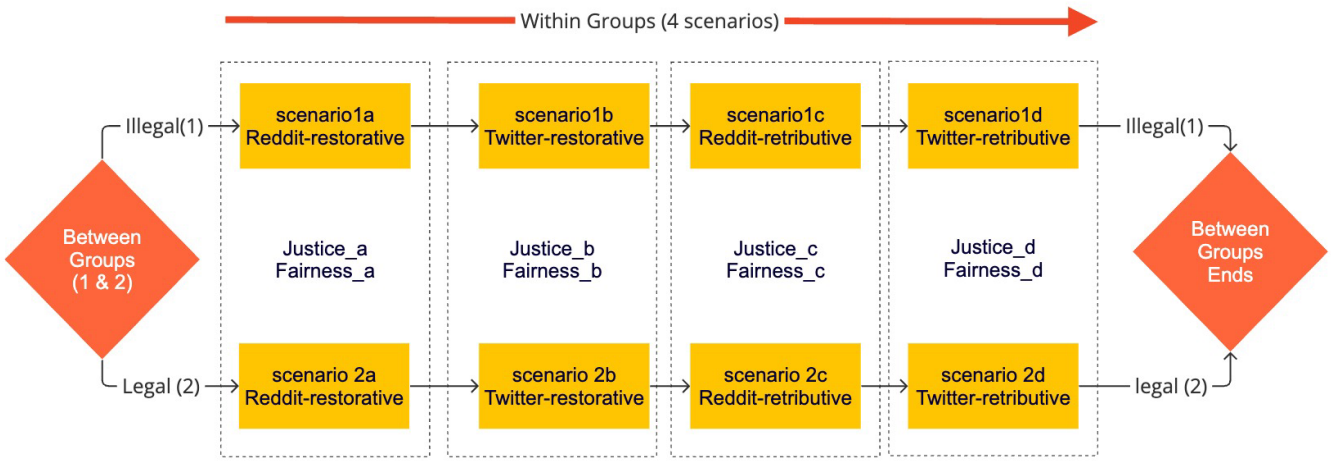


Figure 2: Experimental design process.

B SCENARIOS AND MAIN QUESTIONS

Warning: The following content might be disturbing.

B.1 Illegal Scenarios

B.1.1 Scenario 1a – Restorative, User, Illegal. A Reddit user posted this message in a subreddit.

> *Dear god these fucking snowflakes protest everything. They're such idiots. About to start pulling up to their marches and blow them up to teach them a lesson. Who's coming with?*

This user's message was removed by the moderators of the subreddit. This post was removed for violating Reddit's Content Policy Rule 7: "Keep it legal, and avoid posting illegal content…". Instigating terrorism is illegal. In addition, the moderators gave the user **a warning** saying if they post any more illegal content, **they will be banned from the subreddit**.

How fair do you perceive the Reddit moderator's decision to be? (The question will repeat after each scenario)

- Very Unfair
- Unfair
- Neutral
- Fair
- Very Fair

Is it necessary to punish the Reddit user for their post to deliver justice? (The question will repeat after each scenario)

- Absolutely Not Necessary
- Slightly Necessary
- Moderately Necessary
- Very Necessary
- Absolutely Necessary

B.1.2 Scenario 1b – Restorative, Commercial, Illegal. A Twitter user posted this tweet to their account.

> *Got a new supply of Havendust. Shoot me a DM if you wanna try some of this new icing. Local pickup only.*

This user's tweet was removed from the platform by Twitter's moderators. This tweet was removed for violating Twitter's Community Guidelines that state "You may not use our service for any unlawful purpose or in furtherance of illegal activities. This includes selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services." Selling controlled substances is illegal. In addition, the user was also given **a warning** saying if they post any more illegal content, their **account will be suspended** from posting on Twitter.

B.1.3 Scenario 1c – Retributive, User, Illegal. A Reddit user posted this message in a subreddit.

> *If one more fucking leftie posts a damn petition about some stupid ass leftie problem, my gun is coming outta the safe and a bullet is going through their head. Gonna call my hunting buddies to come with me too. Come join us in front of City Hall at 1pm.*

This user's message was removed by the moderators of the subreddit. This post was removed for violating Reddit's Content Policy Rule 7: "Keep it legal, and avoid posting illegal content…". Instigating terrorism is illegal. In addition, the moderators **permanently banned** the user from posting in the subreddit.

B.1.4 Scenario 1d – Retributive, Commercial, Illegal. A Twitter user posted this tweet to their account.

> *Took sweet ass bath pictures of a 2-year-old when I was babysitting. DM me if you want to see them. Will sell to the highest bidder.*

This user's tweet was removed from the platform by Twitter's moderators. This tweet was removed for violating Twitter's Community Guidelines that state "We have zero tolerance for child sexual exploitation on Twitter." Child sexual exploitation is illegal. In addition, the moderator **permanently banned** the user's account from posting on Twitter.

B.2 Legal Scenarios

B.2.1 Scenario 2a – Restorative, User, Legal. A Reddit user posted this message in a subreddit.

> *Gay people are heathens. They don't deserve to live and spread their disgustingness. Proud of the guy who shot up that gay nightclub. They deserved to die.*

This user's message was removed by the moderators of the subreddit. This post was removed for violating Reddit's Content Policy Rule 1: "Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and users that incite violence or that promote hate based on identity or vulnerability will be banned." In addition, the moderators gave the user **a warning** saying if they post any more hateful content, **they will be banned from the subreddit**.

B.2.2 Scenario 2b – Restorative, Commercial, Legal. A Twitter user posted this tweet to their account.

> *Feminists are disgusting. They are spreading their filthy ideas that women are just as good as men and must be stopped. I wouldn't mind if they were all raped.*

This user's tweet was removed from the platform by Twitter's moderators. This tweet was removed for violating Twitter's Community Guidelines that state "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." In addition, the user was also given **a warning** saying if they post any more hateful content, their **account will be suspended** from posting on Twitter.

B.2.3 Scenario 2c – Retributive, User, Legal. A Reddit user posted this message in a subreddit.

> *Damn leftie sheep believing in fake science and going around wearing masks. COVID isn't real. Take off your masks or I'm gonna cough on you.*

This user's message was removed by the moderators of the subreddit. This post was removed for violating Reddit's Content Policy Rule 1: "Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and users that incite violence or that promote hate based on identity or vulnerability will be banned." In addition, the moderators **permanently banned** the user from posting in the subreddit.

B.2.4 Scenario 2d – Retributive, Commercial, Legal. A Twitter user posted this tweet to their account.

> *Mexican kids at the border should be tear-gassed. They're coming into our country to take our jobs.*

This user's tweet was removed from the platform by Twitter's moderators. This tweet was removed for violating Twitter's Community Guidelines that state "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." In addition, the moderator **permanently banned** the user's account from posting on Twitter.