"Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit

SHAGUN JHAVER, Georgia Institute of Technology
DARREN SCOTT APPLING, Georgia Institute of Technology
ERIC GILBERT, University of Michigan
AMY BRUCKMAN, Georgia Institute of Technology

Thousands of users post on Reddit every day, but a fifth of all posts are removed [45]. How do users react to these removals? We conducted a survey of 907 Reddit users, asking them to reflect on their post removal a few hours after it happened. Examining the qualitative and quantitative responses from this survey, we present users' perceptions of the platform's moderation processes. We find that although roughly a fifth (18%) of the participants accepted that their post removal was appropriate, a majority of the participants did not — over a third (37%) of the participants did not understand why their post was removed, and further, 29% of the participants expressed some level of frustration about the removal. We focus on factors that shape users' attitudes about *fairness* in moderation and *posting again* in the community. Our results indicate that users who read community guidelines or receive explanations for removal are more likely to perceive the removal as fair and post again in the future. We discuss implications for moderation practices and policies. Our findings suggest that the extra effort required to establish community guidelines and educate users with helpful feedback is worthwhile, leading to better user attitudes about fairness and propensity to post again.

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

Keywords: content moderation; content regulation; removal explanations

ACM Reference Format:
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. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 192 (November 2019). ACM, New York, NY. 33 pages. https://doi.org/10.1145/3359294

> *"I feel sad that my effort in making that post was for nothing, and that no one will see it and no one will reply with any help or advice." - P254*

Authors' addresses: Shagun Jhaver, jhaver.shagun@gatech.edu, Georgia Institute of Technology, 85 5th Str. NW, Atlanta, GA, 30308; Darren Scott Appling, scott.appling@gatech.edu, Georgia Institute of Technology, 85 5th Str. NW, Atlanta, GA, 30308; Eric Gilbert, University of Michigan, 105 S State St, Ann Arbor, MI, 48109, eegg@umich.edu; Amy Bruckman, asb@cc.gatech.edu, Georgia Institute of Technology, 85 5th Str. NW, Atlanta, GA, 30308.

2573-0142/2019/11-ART192 $15.00
https://doi.org/10.1145/3359294

1 INTRODUCTION

How do users feel when their content is removed from online communities? Does it deter them from posting again? Does it change their attitude about the community? Individuals have a range of motivations for posting, and this shapes their reactions to content removal. In some cases (like P254 above), a user might really need advice. In others, as we will see, a user might annoy the moderators on purpose, intending to provoke a removal. How does the level of effort made in creating content affect the way users perceive its removal, and does receiving an explanation of why content was removed matter? In this paper, we address these questions through a survey of 907 Reddit[1] users whose posts were removed.

Our research is concerned with understanding content moderation from the perspectives of end-users in cases where the user has likely broken a rule or a community norm. For this article, we use Grimmelmann's definition of "content moderation" as the "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [38]. We focus specifically on content moderation processes that determine which user-generated content to allow on the site and which to remove, as well as how to handle removals. The goal of this research is to offer theoretical and practical insights for community managers to moderate their communities in ways that are considered fair by the end-users and that encourage users to continually submit productive contributions.

In recent years, the fields of Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW) have actively engaged in research on various aspects of content moderation, highlighting the differences in moderation policies across various social media platforms [66] and exploring the design challenges of creating automated tools for enacting efficient content regulation [44]. However, much of the research in this area is theoretical [34, 38, 51] and existing empirical work usually takes a data-centered [12, 14, 55] or moderator-centered [58, 63, 77] perspective. Limited prior research has investigated the perspectives of moderated end-users on Twitter and Facebook [86], but to the best of our knowledge, no prior work has explored how users react to content removals on sites like Reddit that rely on volunteer, community-driven moderation. We seek to address this gap in research with our analysis of moderated Reddit users using large-scale survey data.

End-users are the central actors in online social systems. Sites like Reddit don't usually create their own content. Instead, they rely on a constant stream of user-generated content [34]. Therefore, end-users are not just consumers who bring in the ad revenue to sustain these platforms but they are also the content creators. As a consequence, it is crucial for these platforms to have users who are invested in the online community and who feel valued for their content contributions.

Although many users on these platforms create information goods that are appreciated by the community, there are others whose posts are promptly removed by the community managers before they can be seen by the community[2]. We do not know what happens to individual users after they invest time in creating content only to have it discarded. It is also unclear how the different elements of submission process (e.g., the existence of community guidelines) and the subsequent removal process (e.g., whether or not the user is provided a removal reason) affect users.

Understanding the concerns and experiences of such users may open up opportunities for identifying and nurturing users who have the potential to become valuable contributors in the

[1] https://www.reddit.com

[2] We do not take the view that all content removal decisions on social media sites are legitimate. For example, automated moderation tools may miss the contextual details of the post and remove it based on the presence of specific keywords in the post [44] or human moderators may be motivated by personal biases [47]. We did not (and as individuals external to the community, and therefore, unaware of its social norms, could not) independently verify whether the removals of our participants' posts were legitimate. Rather, our analysis focuses on whether or not end-users perceive these removals as fair.

community. From a wellness perspective, the effects of content moderation on the levels of stress experienced by millions of end-users is also important to consider [57]. Therefore, it is imperative that we study the users who experience content removals.

Our study focuses on moderated users' perceptions of fairness in content moderation and how content removals shape their attitude about posting in the future. We consider these two factors as important to understanding the users' current orientation and future outlook towards online communities. Our analysis is guided by the following research questions:

- RQ1: How do users perceive content removals?
- RQ2: In what ways do the contextual factors of post submission and content moderation, such as community guidelines and removal explanations, shape users' perceptions of fairness of content removals?
- RQ3: How do these contextual factors affect users' attitude about posting in the future?

To answer these questions, we conducted a survey of users (N=907) who have experienced content removals. We chose to conduct this study on the popular social media platform Reddit. Reddit is made up of more than a million communities called "subreddits." We use the terms communities and subreddits interchangeably in this paper. Each subreddit has its own independent content regulation system maintained by volunteer users — the subreddit's moderators. These communities show a variety of approaches to enacting content curation. For example, some communities highlight a set of community guidelines or subreddit rules that users should follow while others don't provide any such guidelines [29]. To take another example, moderators on some communities provide explanations for post removals while others choose to silently remove posts. Therefore, the Reddit platform provides a rich site to study how the differences in the contexts of post submissions and subsequent moderation actions affect the attitudes of users.

We triangulate quantitative and qualitative data from our survey to present a rich overview of moderated users, their concerns, and their interactions with various elements of the moderation process. As might have been predicted, a majority of our participants expressed negative attitudes about their content removals. However, analyzing their dominant affective responses and interactions with moderation processes reveal insightful nuances. For example, our qualitative data indicate that the absence of notifications about removals results in users creating folk theories of how content moderation works, and this reinforces their negative attitudes. Our quantitative analyses show that having community rules and receiving removal explanations are associated with users perceiving the content removal as fair and having better attitudes about future posting.

This work sheds light on the needs of users whose content has been removed. We add support to prior research that calls for taking an educational approach rather than a punitive approach to content moderation [86]. We offer recommendations for designers to build tools and community managers to adopt strategies that can help improve users' perceptions of fairness in content moderation and encourage them to become productive members of the community. Since moderation resources are often scarce, our empirical findings can help moderators make informed choices about where to invest their effort related to providing community guidelines, offering removal explanations, and using automated tools.

2 STUDY CONTEXT: REDDIT MODERATION

Reddit is one of the most popular social media sites today. As mentioned above, Reddit is made up of millions of independent subcommunities called subreddits, with each subreddit focused on a separate topic. Users can subscribe to multiple subreddits in order to see the latest content from those subreddits on their front pages.

Each subreddit allows users to post submissions which can be text posts, images, videos or links to other sites. Other users can then comment on these submissions and thereby engage in threaded conversations. Users can also upvote or downvote each submission and comment. Each Reddit user (or Redditor) accumulates a digital score called 'karma' that reflects the net votes her posts have received. Many subreddits also have a list of rules or community guidelines that informs users what type of content is acceptable to post on the community. These rules are placed in a sidebar that appear on all pages of the subreddit.

For the purpose of this study, we focus on moderation of submissions on Reddit. When moderators remove a submission on Reddit, the site doesn't automatically notify the author of the submission about the removal. While on some communities, moderators explicitly inform the poster that their submission has been removed, most communities choose not to inform the poster. When signed in, posters can still access all their submissions they have posted on their user profile page, regardless of whether they have been removed on Reddit. Therefore, posters may only come to know about the removal if they check the front page of the subreddit and do not notice their post.

When a submission is removed, moderators on some communities also choose to provide posters an explanation of why the removal occurred. This can be done in a number of different ways. Moderators can (1) comment on the removed post with a message that describes the reason for removal, (2) flair[3] the removed post, or (3) send a private message to the submitter. Moderators can either choose to post the removal explanation themselves, or they can configure automated tools (e.g., AutoModerator [44]) to provide such explanations when the submission violates a rule.

In this study, we evaluate how these different design mechanisms mediate user responses to content removals on Reddit.

3 RELATED WORK

3.1 Content Moderation

Since the early days of the internet, scholars as well as community managers have deliberated over how to manage content online and how to enable constructive conversations among end users [6, 24]. Although modern social media platforms operate in new contexts, they still struggle with some of the same problems that afflicted the early incarnations of online communities. For example, they have to address the challenges of dealing with "trolls" who escape easy detection [40, 47] and making subjective decisions that distinguish controversial speech from online harassment [46].

As online communities grow large, curating content and enforcing local social norms of the community [13] become increasingly difficult [34]. To address these challenges, platforms like Facebook, Twitter and Reddit have developed complex, multi-layered content-moderation systems that involve "visual interfaces, sociotechnical computational systems and communication practices" [86]. The black-boxed nature of many social media platforms [56] makes it difficult to study their moderation systems. Yet, over the past few years, a rich body of research has made significant initial forays into unpacking the complexities of how content moderation is enacted [19, 34, 49, 54, 55, 58, 62, 63, 71, 77]. We refer the reader to Jhaver et al. [44] for an in-depth review of this literature.

The line of content moderation research most closely related to the current work focuses on the perspectives of end-users of centralized moderation[4] platforms like Facebook and Twitter.

[3]Flairs are short tags that can be attached to users' submissions. Only the moderators on each subreddit have access to assign removal explanation flairs to the posts on that subreddit.

[4]We define centralized moderation systems as those systems in which moderators may be assigned to regulate any part of the site using a common set of global criteria. On the other hand, we define distributed moderation systems as those systems in which moderators regulate only a specific community (or specific communities), and make individual decisions using the criteria relevant to that community.

Jhaver et al. interviewed Twitter users to understand the use of third-party blocking mechanisms on Twitter for addressing the problem of online harassment [47]. They showed that although adopting these tools helps improve the user experiences of previously harassed users, it also raises legitimate concerns among users who are mistakenly blocked by these tools without being given a chance to redress. We found similar concerns among moderated Reddit users who felt that their content was removed unfairly and who reported being frustrated about their communications with the moderation team. In another recent study, West analyzed users whose posts are removed on Facebook, Twitter and Instagram, and surfaced these users' folk theories of how content moderation works [86].

Our work complements this prior research on centralized moderation platforms by highlighting the impact of content moderation on users in the distributed moderation system of Reddit. We expect that the multi-community environment of Reddit, along with its locally crafted guidelines and volunteer-driven moderation, makes moderation on Reddit a much different experience from moderation on other platforms such as Facebook and Twitter. Our research attempts to highlight the concerns that arise in the distributed moderation system of Reddit. We use the user-centered perspectives we obtain in our findings to suggest guidelines on designing moderation tools and strategies that may improve the health of distributed online communities like Reddit.

Moderators on many sites use explicit community guidelines to make the community norms more visible, especially to newcomers [29]. However, how these guidelines affect user attitudes and behaviors remains unclear. Kiesler et al. hypothesize that while explicit guidelines may help users understand what is acceptable to post on a community, they may also discourage user contributions if the users feel stifled [51]. Building upon their theoretical work, we provide empirical insights into how the presence of community guidelines and the perceptions of users about these guidelines shape their attitudes about the community.

Our research extends prior work on community-created rules on Reddit. Fiesler et al. [29] recently presented a description of the rules ecosystem across Reddit, highlighting that the activity on Reddit is guided by multiple layers of rules. First, there exists a user agreement and content policy similar to the terms and conditions of many websites. Second, a set of established rules defined by Reddit users, called Rediquette, guide site-wide behavior. Finally, many subreddits also have their own set of rules that exist alongside site-wide policy and lay out expectations about content posted on the community [29]. Thus, Reddit users operate in an ecosystem of governance where even the explicit guidelines for expected behavior, let alone the implicit norms, derive from multiple sources. Prior research has shown that negotiating multiple sources of rules can result in confusion among end-users [28]. We explore in our analysis how Reddit users negotiate posting in multiple communities, each having its own set of rules, and the challenges they face in the process.

Our work also contributes to the growing body of research on understanding bad actors online [3, 4, 18, 46, 47, 68]. Coleman [18] and Phillips [68] both conducted deep ethnographic investigations to understand the subculture of internet trolls. We add to their research by surfacing the perspectives of bad actors[5] on Reddit and discussing various factors that motivate them to post. More recently, in their analyses of communities accused of perpetrating online harassment, Jhaver et al. pointed out the challenges of distinguishing sincere users from bad actors, and the problems that emerge when sincere users are mistakenly moderated [46, 47]. We discuss how the difficulties in identifying bad actors complicate the process of allotting moderation resources to nurturing sincere users.

[5]We do not considers all users whose posts have been removed as "bad actors." Here, we refer to only those participants who described themselves as "trolls."

3.2 Fairness and Transparency in Content Moderation

Social media platforms play a decisive role in promoting or constraining civil liberties [20]. They make day-to-day judgments about which content is allowed and which is removed, and intervene in public disputes over intellectual property and controversial speech [20, 46]. How platforms make these decisions has important consequences for the communication rights of citizens and the shaping of our public discourse [34].

Prior research has suggested a number of different frameworks in which platforms can ground their policy decisions, each with its own merits and challenges [32, 35, 80–82]. Each of these frameworks emphasize different sets of normative principles, ranging from rights-based legal approaches [32] and American civil rights law [16] to principles of social justice [35, 81, 83]. Another framework that has been proposed for platform governance is the 'fairness, accountability, and transparency' (FAT) model, that has recently seen a lot of interest in the HCI community, especially as it applies to algorithmic systems [11, 23, 35]. Although this model has faced some critiques, such as its slowness to fully incorporate the lessons of intersectionality [41] and its tendency to overstate the power of technology [67], we draw from this model as a starting point, and focus on the principles of fairness and transparency in our study because this allows us to engage with other relevant HCI literature that concerns fairness and transparency in sociotechnical systems [27, 48, 70].

Over the past few years, social media platforms have often been criticized for how they moderate their content. Media sources have frequently reported how platforms fail to remove disturbing content [17, 65, 72], block content that is considered important to be circulated in the public sphere [37, 50, 76], promote content that are related to conspiracy theories [1, 84] and show political biases in content curation [5, 10]. Some scholars have begun to raise questions about the appropriate limits on the private exercise of power by social media platforms [33, 81]. They have highlighted the human rights values that must inform the evaluation of fairness in content moderation. Such values include "freedom of expression," "due process," and "transparency and openness" [81]. We add to this emerging literature by providing empirical insights into end users' perspectives of what they consider "fairness" in content moderation. We also study how different elements of content moderation and the context of submissions affect users' perspectives of fairness in moderation decisions.

In a related line of work, some researchers have reflected on the importance of transparency in content moderation [47, 77, 81, 82, 86]. For example, West noted that moderation systems have the opportunity to serve an educational, rather than a punitive, role by providing moderated users an understanding of why their content was removed [86]. We add to this research by investigating whether the design mechanisms of community guidelines and removal explanations can serve an educational role on Reddit. Many scholars have proposed that platforms should ground their policy decisions in the principles of meaningful democratic accountability and transparency [31, 35, 43]. Suzor et al. have called for a range of digital platforms to issue transparency reports about how they enforce their terms of service [81]. Although these previous studies have begun to raise questions about transparency and explanations, there still exists a gap in our understanding of how transparency in distributed moderation systems like Reddit affects user attitudes and behaviors.

We begin to fill this gap by focusing on explanations for content removals, an important aspect of transparency in moderation on social media platforms. While the utility of explanations in providing system transparency and thereby increasing user acceptance has been demonstrated in many domains [2, 9, 53, 69, 85], some scholars have raised questions about the limits and effectiveness of such transparency based strategies [27, 52, 70]. In the context of content moderation, explanations are difficult to design effectively [8, 36, 39], and they require time and effort on the behalf of both

moderators who provide them as well as users who consume them. Thus, it is unclear whether they are effective and worth implementing. This paper examines the effectiveness of explanations from the perspective of end-users. We focus on how users perceive such explanations, how they feel about the lack of any explanations, and the ways in which the content, modes and sources of explanations affect user perceptions.

3.3 Folk Theories of Sociotechnical Systems

DeVito et al. define folk theories as "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" [22]. In recent years, HCI and CSCW researchers have begun exploring how users of sociotechnical systems develop folk theories about their operations and the ways in which such folk theories affect users' interactions with these systems [21, 22, 26, 30, 48].

DeVito et al. analyzed how Facebook users seek and integrate different information sources to form folk theories of algorithmic social media feeds and how these folk theories interplay with users' self-presentation goals [21]. We discuss the need for similar efforts to understand folk theory formations in content moderation systems.

Jhaver et al. investigated the folk theories developed by Airbnb hosts about the operation of Airbnb search algorithm and found that hosts' beliefs in some of these theories created anxiety among them, often forcing them to engage in wasteful activities as part of their coping strategies [48]. Similar to this, we found folk theories in our data that caused anxieties and frustrations among end-users in the context of Reddit moderation. Eslami et al. studied the folk theories of Facebook News Feed curation algorithm and concluded that implementing structured transparency or "seams" into the design of these systems may help improve human-algorithm interaction and benefit human agency in complex systems [26]. We add to this literature by exploring the interplay between folk theories and transparency in the domain of content moderation. We highlight the folk theories of content moderation that Reddit users develop in order to make sense of their content removals, and discuss how these theories may inform governance practices.

4 METHODS

We designed a survey to directly ask users who receive different types of moderation responses questions about their experiences (see Appendix A). We used a modified form of experience sampling approach [75] to collect our participants' responses right after their posts got removed on Reddit. The survey contained 24 questions (mostly multiple choice, with a few free-response). Our goal in this analysis was to gain a deep understanding of how variations in moderation affect users. We study how the users perceive the feedback mechanisms (e.g., subreddit rules, comment describing the reason for removal) that are implemented by the moderators. Most importantly, we investigate how the users' experiences with content removal shape their attitudes about fairness in content moderation and future interactions on the community.

4.1 Survey Instrument

The survey questions were based on the tension points around content moderation that have surfaced in prior work [29, 44, 47, 58, 59] and workshop discussions [4, 7, 60]. To increase the validity of the survey, we conducted an in-person cognitive pretest of the survey with four students at the authors' institution. These students were not involved with the project and they provided feedback on wording of the questions and survey flow, which we incorporated into the final survey design. We also piloted the survey with a small subset of the sample (28 participants). During this field test, we included this question in the survey: "Q - This survey is currently in pilot stage. Do

you have any suggestions for how we can improve this survey?" These survey pretests resulted in several rounds of iteration before our questionnaire reached the desired quality.

The questions in this survey measured the attitudes and perceptions of users concerning posts they made that had recently been removed on Reddit. We asked users how they perceived the fairness of the post removal. Our questions captured users' awareness and impression of different features of Reddit moderation system such as subreddit rules and removal explanations. We also included open-ended feedback questions in this survey to understand the relative frequency of key satisfactions and frustrations with moderation systems. These questions asked users: "Please explain how you felt about the removal" and "Is there anything else you'd like to tell us about your view of this removal?"

Our questionnaire used skip logic, i.e., we asked a different set of questions to different respondents based on their answers to previous questions. For example, we first asked users whether they noticed any rules on the subreddit they posted to. Only if the participant answered 'yes' to this question did we ask them follow-up questions about whether they read the rules and whether the rules were clear. This was done so as to remove questions that may be irrelevant for some respondents and reduce the time they needed to complete the survey. We used Google Forms to implement this survey.

Following the guidelines described in Müller et al. [64], we took several steps to avoid the common questionnaire biases in designing our survey instrument. For example, to minimize satisficing bias, we avoided questions that required an excessive amount of cognitive exertion. To minimize social desirability bias, we allowed participants to respond anonymously [42]. However, we asked participants to submit their Reddit username so that we could later merge survey responses for each participant with their behavioral data obtained using the Reddit API via the username identifier. To minimize question order biases, we ordered questions in a funnel approach, i.e., from broad to more specific. Earlier questions were easier to answer and were more directly related to the topic of the survey whereas sensitive questions (e.g., about education levels and age) were placed towards the end of the survey so as to build rapport and avoid early drop-off [25]. We grouped related questions together to reduce context switching, and we presented distinct sections on separate pages of the survey for easier cognitive processing. Furthermore, we avoided including broad questions, leading questions and double-barreled questions in this survey [64].

We took several steps to discourage disruption from those who might seek to manipulate the data. First, we recruited our participants through private messages instead of publicizing the link to our survey webpage on public forums. Second, we made most questions "required" so that the survey could not be completed until a response was submitted for each question. Third, following the method implemented by West [86], we adopted a page-by-page design so that users had to click through multiple pages of content in order to complete the survey. We also asked participants to describe in their own words how they perceived the content removal. Although dedicated actors may still manipulate the data, these measures were intended to act as disincentives to providing falsified data in large quantities.

This research was approved by the Georgia Tech Institutional Review Board (IRB). Subjects were not compensated for their participation.

4.2 Data Collection

Our sampling frame consisted of all Reddit users whose recent post(s) was removed. To minimize selection bias, we used a random sampling approach to select participants: we randomly drew the sample from users in our sampling frame and invited every user in the sample in the same way [64]. Our strategy for determining when to contact users in this sample was motivated by two goals:

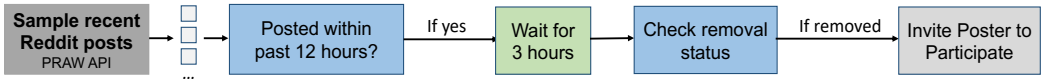

Fig. 1. Flowchart depicting the data collection process. We created a Python script to automate this process.

(1) Enough time should have passed after a submission is posted for human moderators to review and in some cases remove that submission or allow moderators to reverse the removal decisions incorrectly made by automated tools.
(2) Submission should be posted recently so that the submitter would easily recall the circumstances around the posting and provide appropriate responses to the questions asked in the survey.

Achieving both of these goals required attaining a balance between sending the survey request soon enough so that the users would have a sufficient recall of their submission but not so soon that moderators haven't had enough time to review that submission. For the purpose of this survey, we configured our data collection such that at least three hours elapsed after the time of submission so that the moderators had enough time to review it. At the same time, we only collected data against submissions that were posted less than 12 hours ago. This selection of time bounds for contacting participants was not based on any empirical tests but it was based on our experiences as Reddit moderators on a variety of subreddits over the past four years. Subreddits usually have multiple moderators, often located in different time zones, and they review submissions much more promptly than they review comments in order to avoid community engagement with submissions that are subsequently removed.

We began by collecting a random sample of 10,000 (allowed as well as removed) submissions recently made on Reddit using PRAW Reddit API[6]. This API does not provide a facility to directly retrieve 10,000 random submissions across all subreddits. Therefore, to collect our data, we started with randomly sampling a subreddit and then retrieved the most recently posted submission on this subreddit. We stored this submission if it was posted in the past 12 hours. Next, we repeated this process until we got 10,000 random submissions posted in the past 12 hours (Figure 1).

After we retrieved the 10,000 submissions in the previous stage, we waited for three hours so that the moderators had sufficient time to review and moderate those submissions. At the end of this waiting period, we tested the removal status of the collected submissions, again using PRAW Reddit API. Next, we retrieved the authors of submissions that were removed and sent them a customized invitation message to participate in the survey, with each message containing the link to the corresponding user's removed submission. We created a Python script to automate this entire process, which included using Reddit API to send out a customized survey invitation to moderated users through a Reddit private message. Using this script allowed us to ensure that the messages were promptly sent to all the participants.

Because we randomly sampled subreddits in this process, regardless of their popularity or subscriber size, our sampling strategy allowed us to have a broad coverage of subreddits in our sample. It was important for us to have this diversity of communities and to not have our data representing only a few, large subreddits because we wanted to measure users' responses to a wide range of moderation practices in our survey (Appendix A). Yet, since we selected only those

[6]https://praw.readthedocs.io/en/latest/

posts that were submitted in the past 12 hours, our sampling favored subreddits that were at least somewhat active.

We repeated this process for seven days at which point we got the target number of responses to our survey. 8.2% of all users we contacted submitted our survey. We considered this to be a suprisingly high response rate given that we sent our survey requests through online private messages. We attribute this high response rate to the customized private messages we sent using our Python script. Many of our invitees replied back to our invitation message with clarifying questions about the survey. We answered such questions as often as we could, and we believe, this also helped boost the survey response rate. We ensured that we did not send invitation message to any user more than once. Furthermore, we initiated our Python script for data collection at different times of the day everyday so that users posting at different times and in different time zones could be selected.

4.3 Data Preparation

After the data collection was completed, we proceeded with data preparation and cleaning. First, we removed all duplicate responses by looking for multiple entries that contained the same username. As mentioned before, our data collection was configured such that each user was sent only one invitation to respond to the survey. Therefore, our records contained a unique subreddit for each user where the post removal occurred. Our survey asked respondents to type in the name of the subreddit that their removed submission was posted to. This was an attention check question. We manually matched the answer posted to this question against our records to verify that the participant was responding about the removed submission we had on our records, and removed all survey responses where a mismatch occurred.

We read all responses to the open-ended questions in this survey and manually removed the obvious garbage answers such as "abcd." We also examined other answers from the same respondent to determine whether all answers from that respondent warrant removal [64]. It is, of course, possible that some participants were not truthful about their experiences of content removals. We acknowledge, however, that response bias is inherent in any self-report dataset obtained, and our results should be interpreted with this in mind.

4.4 Participants

In total, 1,054 users clicked through the consent form and submitted the survey. We filtered out inappropriate survey responses using the data cleaning process described in the last subsection. Our final sample consisted of 907 responses. Participants came from 81 different countries, although there was a heavy skew toward North America. The four countries with the most respondents were the U.S. (61%), Canada (7%), U.K. (5%), and Australia (3%). Seven respondents elected not to provide their place of origin. A majority of the participants were male (81%) and under 25 years old (55%). Comprehensive demographic information is reported in Table 1.

We used the Reddit API to get additional information about our participants related to their activity level on Reddit. Participants had a median of 3,412 karma points (μ = 45K, σ = 421K). Their account age had a median of 436.6 days (μ = 804.93, σ = 496.6). Participants had posted a median of 35 submissions on Reddit (μ = 258.06, σ = 1195.7).

4.5 Variables

The survey asked respondents about their agreement with the statements (1) "I think that the removal was fair" on a five-point Likert scale (1=Strongly Disagree, 5=Strongly Agree), and (2) "How likely are you to post again on this subreddit after this experience?" also on a five-point Likert scale (1=Very Unlikely, 5=Very Likely). We used the answers to these questions as dependent variables

Table 1. Participant demographics (N = 907). Note that the percentage of each factor does not always sum to 100% because of rounding. A majority of participants were from North America. The most frequent age group was 18-24. 81% of participants identified as male.

Factor	Category	% (N)
Country	United States	61% (554)
	Canada	7% (68)
	United Kingdom	5% (46)
	Australia	3% (31)
	India	2% (14)
	Other	21% (187)
	Prefer not to Answer	0.8% (7)
Age	18-24	55% (502)
	25-34	23% (210)
	35-44	11% (100)
	45-54	5% (45)
	55-64	1% (10)
	>65	0.3% (3)
	Prefer not to Answer	4% (37)
Education	Less than High School	15% (133)
	High School	18% (160)
	Some College, No Degree	20% (185)
	Bachelor's Degree	22% (200)
	Master's Degree	9% (81)
	Associate degree	6% (57)
	Doctorate Degree	2% (18)
	Prefer not to Answer	8% (73)
Gender	Male	81% (738)
	Female	13% (121)
	Another Gender	1% (14)
	Prefer not to Answer	4% (34)

in each of our regression analyses (Table 2). We refer to these variables as *Fairness* and *PostAgain*, respectively through the rest of this paper. We note here that the variable *PostAgain* measures our participants' *perception* of their likelihood of future posting, and not whether they actually posted again on the subreddit. While analyzing the actual future behavior of moderated users is an important research direction, and we pursue that direction in [45] with a larger dataset, the current work focuses on understanding users' attitudes. Therefore, using the *PostAgain* dependent variable, we seek to explore users' beliefs about their future posting just after they experience content moderation.

For each user, we gathered features that we hypothesized would be related to our dependent variables. These features can broadly be categorized into three different buckets: (1) Posting context, (2) Community guidelines, and (3) Removal explanations. Table 2 shows the list of variables we included in our models.

We included the following features that related to the context of the posting as independent variables in our analyses: (a) The amount of time spent in creating submission, (b) whether the participant suspected the post would be removed before submission, and (c) whether the participant noticed the post removal before starting the survey. Through these variables, we wanted to test whether the amount of time users spend in composing their posts has an effect on their attitudes about content removals. We were curious to see whether the participants who did not expect a post removal before submission have different reactions to content removals than users who suspected a removal. Finally, we wanted to analyze whether the users who were not notified about the removal

Table 2. Control variables (including demographic and prior history variables), independent variables (including post context, community guidelines and removal explanations variables), and dependent variables included in our regression analyses.

Control Variables
Demographics
(1) Age
(2) Education
(3) Gender
Prior History
(1) Reddit karma
(2) Time on Reddit (in days)
(3) Number of submissions posted on Reddit

Independent Variables
Posting Context
(1) Time spent in creating submission
(2) Suspecting removal before posting
(3) Noticing post removal
Community Guidelines
(1) Reading the rules
(2) Understanding the rules
Removal Explanations
(1) Explanation providing new information
(2) Explanation mode
(3) Explanation source

Dependent Variables
(1) *Fairness*
(2) *PostAgain*

of their post and only came to know about it through our survey request have different reactions to content moderation than users who were notified about the removal. We test for these associations in our regression analyses.

In the context of community guidelines, we used responses to two questions on a Likert scale as our independent variables: (a) "I read the rules of the subreddit before posting," and (b) "The rules on this subreddit are clear." Limited prior research shows that highlighting the rules has an effect on improved user behavior [59]. We wanted to test whether reading the rules has an association with participants' *attitudes* as well. We also sought to study whether the clarity of rules improves users' perceptions.

As we discussed in Section 2, on Reddit, moderators can provide explanations for post removals in a variety of ways. They can comment on the removed post, flair the post, or send a private message to the poster. Moderators can either compose these messages themselves or they can configure automated moderated tools to automatically send explanations whenever a submission matches a known pattern. Given this context, we wanted to test these factors for their associations with our dependent variables: (a) Whether the explanation provided new information to the participant, (b) Source of explanation ('human', 'bot' or 'unsure') and (c) Mode of explanation ('comment to submission', 'private message' or 'flair'). Testing for these associations allowed us to explore how the ways of providing explanations and the novelty of information in explanations affect users' attitudes.

We selected these independent variables because they were open to direct interpretation and we wanted to test their relationships with attitudes about fairness in moderation and future postings. Our selection of these variables as factors of interest was based on intuitions developed through serving as moderators and content contributors on many Reddit communities over the past four years. We note that our analytic approach is designed to be primarily exploratory, as these constructs have not yet been examined in literature.

In each of our models, in keeping with prior literature on perceptions of social media users [73, 74], we used the participants' demographic characteristics (age, education level, and gender) and prior history on Reddit (as measured by their karma score, number of days since they created their account, and number of submissions on Reddit) as control variables (Table 2). We treated gender as a nominal variable, and age and education as ordinal variables in our regression analyses.

Further, we treated all the 'Prefer not to Answer' entries as missing values in our models. We note although the control and independent variables discussed above capture many important factors that may mediate user reactions to content removals, there are other factors related to user demographics (e.g., race) and subreddits (e.g., topic) that we do not control for in our analyses. Therefore, we see our models as reasonable simplifications of the Reddit sociotechnical system.

4.6 Data Analysis

We used linear regression models for their easy of interpretability after checking for the underlying assumptions. We created separate regression models for evaluating the effects of posting context variables, community guidelines variables and removal explanations variables (Table 2) on our dependent variables. We built separate models because given the skip-logic nature of our questionnaire, only certain subsets of participants had responses to some lines of questions (Appendix A). In addition to these analyses, we also conducted separate tests to understand the associations of noticing community guidelines and receiving removal explanations with our dependent variables. When building each regression model, we performed listwise deletions of the cases where any of the input variable value was missing.

For the open-ended questions, we iteratively developed a set of codes based on an inductive analysis approach [79]. We coded for these questions together because each question had responses that pertained to themes about perceptions of content moderation. This process resulted in a codebook with ten codes (Table 7). The first and second authors coded all open-ended responses side-by-side in order to iterate on the codes and to double-check the results. All disagreements between the two coders were resolved through discussions.

5 QUANTITATIVE FINDINGS

We first calculated the descriptive statistics for the dependent variables *Fairness* and *PostAgain*. Overall, we found that 10.3% of all participants strongly agree and 13.6% agree that the removal was fair whereas 33% of participants strongly disagree and 26.9% disagree that the removal was fair. 16.3% of participants felt "neutral" about the fairness of moderation (Figure 2). We also found that 19.7% of all participants considered it very likely and 23.3% considered it likely that they will post again on the subreddit where their submission was removed while 13.3% of participants felt it very unlikely and 21.1% considered it unlikely that they will post again. 22.6% of participants felt "neutral" about this factor (Figure 2).

In the rest of this section, we explore how these attitudes are linked to different aspects of user experience on Reddit such as posting context, submission guidelines, and removal explanations.

5.1 Posting Context

To identify how posting context is associated with our dependent variables, we conducted a linear regression for each of the two dependent variables. We used posting context variables as independent variables in these analyses. Table 3 shows the results of these regressions. Our baseline models, that used only the control variables (n=738), explain only 0.6% (adj. R^2 = .006) and 0.8% (adj. R^2 = .008) of variance in *Fairness* and *PostAgain* respectively (Table 4). Adding the posting context variables (n=738) increases the adjusted R^2 values to .219 and .033 respectively.

We asked participants how much time they spent in creating their submission. Our survey data show that 31% of participants took less than a minute and only 8% of participants spent more than 10 minutes to create their submissions (Figure 3). As Table 3 shows, time spent in creating submissions is significantly related to both *Fairness* and *PostAgain* even after controlling for other factors. Results indicate that as time spent increases, participants are less likely to consider the removal as fair (β = -.118) and less likely to consider posting in the future (β = -.136). One possible

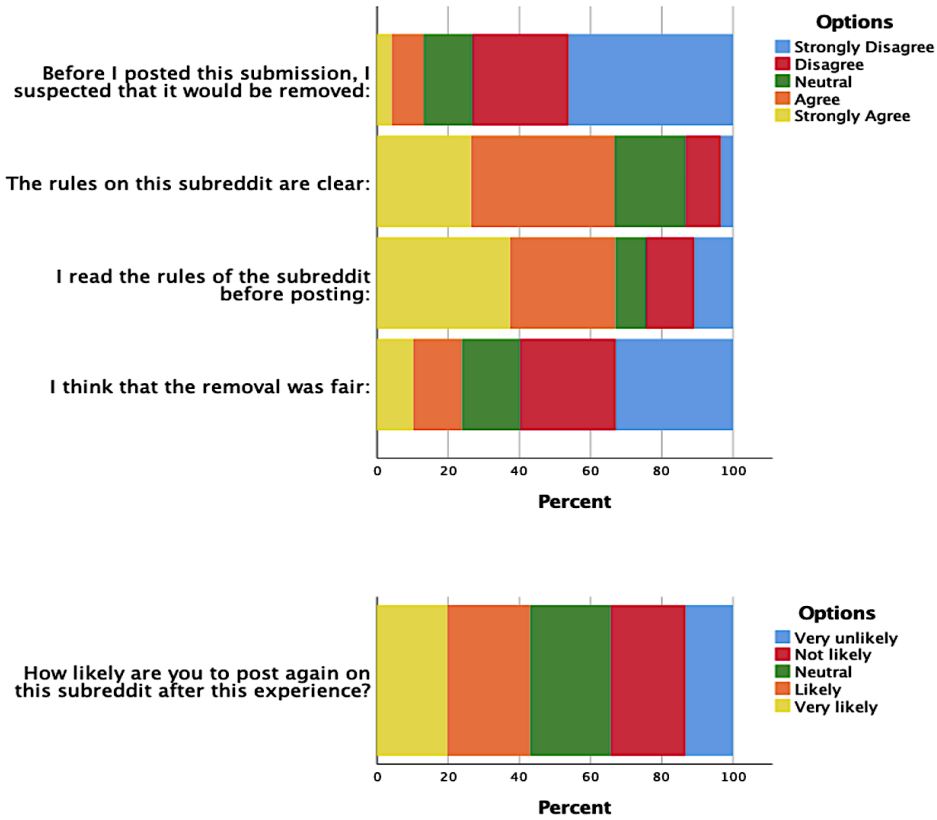

Fig. 2. Frequency of participants' responses to various survey questions, measured in percentage.

Table 3. Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users are likely to post again on the corresponding subreddit (*PostAgain*). This model includes posting context variables as independent variables in addition to the control variables as inputs.

		Fairness				*PostAgain*			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	1.857	0.219		.000	3.606	.240		.000
Control Variables	**Age**	-0.039	.052	-.029	.451	-.016	.057	-.012	.780
	Education	-0.041	.030	-.052	.169	.005	.033	.007	.867
	Gender	0.165	.130	.043	.204	-.252	.142	-.066	.076
	Reddit Karma	1.3E-8	.000	.004	.921	-6.2E-9	.000	-.002	.965
	Reddit Age	0	.000	.073	.041	4.5E-5	.000	.031	.431
	Submission Count	2.5E-5	.000	.021	.630	8.3E-5	.000	.071	.140
Posting Context Variables	**Post Creation Time**	-0.18	.050	-.118	.000	-.203	.055	-.136	.000
	Noticed removal	0.362	.045	.265	.000	.051	.049	.038	.297
	Expected removal	0.396	.039	.334	.000	.096	.043	.083	.025

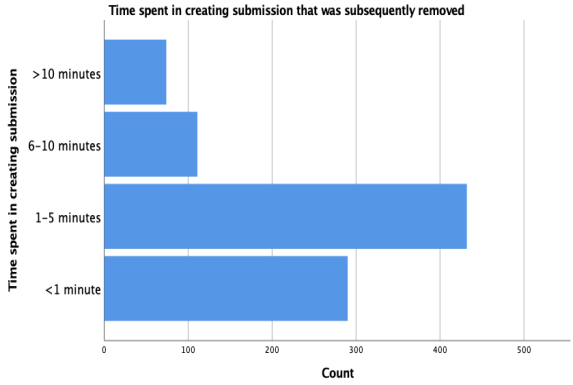

Fig. 3. Time spent in creating submissions that were subsequently removed, as reported by participants.

explanation for this is that users who spend more time crafting their post are more invested in their post, and therefore, they feel more aggrieved at the removal, and less motivated to post again.

Next, we explored how often the users even notice it when the content they post on Reddit is removed. To our surprise, 41.8% of our respondents (n=379) reported they did not notice that their post was removed until they received our invitation message to participate in the survey. We received dozens of replies to our invitation messages where users expressed surprise about their post removals. Many users we contacted also complained about not receiving any notification about the removals. Our regression analyses (Table 3) show that when participants notice their removal, they are more likely to consider the removal as fair (β = .265). This suggests that when users become aware of their post removals, perhaps through communication by the moderation team, they are more likely to consider the moderation of their post as fair than when they are not made aware. Noticing removals, however, did not have any statistically significant effect on whether participants consider it likely they will post in the future.

We also asked in our survey whether users suspected that their submission would be removed before they posted it on Reddit. Our results showed that 73.2% of participants "disagree" or "strongly disagree" that they suspected a post removal whereas only 13.1% of participants "agree" or "strongly agree" that they expected a removal (Figure 2). Our regression analyses (Table 3) suggest that suspicion of post removals before submission is positively associated with *Fairness* (β = .334) as well as *PostAgain* (β = .083). This indicates that users who expect a removal prior to posting their submissions are more likely to consider the removal as fair and less likely to be deterred from future posting by the moderation process.

5.2 Community Guidelines

As we discussed in Section 2, on each Reddit community, moderators can provide community members with explicitly stated guidelines called subreddit rules. These rules appear in the sidebar of each subreddit and they describe the injunctive norms[7] of the community. They are one of the key design elements on Reddit for allowing community managers to encourage voluntary compliance with behavior norms. Kiesler et al. suggest that when social norms are clearly stated through explicit rules rather than being left for users to reasonably infer for themselves, users are more likely to comply with those norms over a variety of situations [51]. However, not all

[7]Injunctive norms are norms that set behavioral expectations by prescribing acceptable community practices [15].

Table 4. Summary of regression analyses for (1) whether users will perceive removal as fair (*Fairness*) and (2) whether users consider it likely that they will post again on the subreddit (*PostAgain*). Asterisk levels denote p<0.05, p<0.01, and p<0.001.

	Fairness		*PostAgain*	
Model factors	**Adj. R^2**	**F**	**Adj. R^2**	**F**
Baseline	.006	1.755 (6, 732)	.008	2.007 (6, 732)
Baseline + Posting context	.219	24.06 (9, 729) ***	.033	3.842 (9, 729) ***
Baseline + Community guidelines	.124	10.307 (8, 518) ***	.055	4.852 (8, 518) ***
Baseline + Removal explanations	.044	2.459 (9, 296) **	.000	.989 (9, 296)
Baseline + Posting context + Community guidelines	.298	21.305 (11, 515) ***	.067	4.451 (11, 515) ***
Baseline + Posting context + Community guidelines + Rem. explanations	**.324**	8.279 (14, 212) ***	**.123**	3.262 (12, 212) ***

subreddits choose to create and display community guidelines. In a recent analysis of a sample of Reddit communities, Fiesler et al. found that only half of the communities had explicit rules [29].

We analyzed how users' attention to the presence of these rules in a community affect their attitudes. First, our results show that 71% of participants (n=644) claimed that the subreddit they posted to contained rules in its sidebar, 6.1% (n=55) said there were no rules, and 22.9% (n=208) were unsure. We built a regression model for *Fairness* (n = 738) using the independent variable *containsRules* (this measures whether the participants noticed the rules) and adding the six control variables (listed in Table 2). This model showed that noticing the rules was positively associated with perception of the removal as fair (β = .068, p < .05). However, a regression model for *PostAgain* using *containsRules* as an independent variable and including the control variables (n = 738) did not find a significant association for *containsRules* (β = .033, p = .432).

A significant association between *containsRules* and *Fairness* highlights that making rules more prominent may improve users' attitude towards online communities. Thus, creating injunctive norms for the community and nudging users to attend to them is a valuable and underused [29] moderation strategy that more community moderators should consider adopting.

We asked the participants who noticed the subreddit rules (n=644) two additional questions: (1) whether they read the rules just before posting and (2) whether they perceived the rules of the subreddit to be clear. Results showed that of all the participants who noticed the rules, 66.9% of participants (n=431) "agree" or "strongly agree" and 24.4% (n=157) "disagree" or "strongly disagree" that they read the rules. Moreover, 66.8% of participants (n=430) "agree" or "strongly agree" and 13.3% (n=86) "disagree" or "strongly disagree" that the rules are clear (Figure 2).

In order to examine the relationships between users' interaction with community guidelines and their perceptions of content moderation, we created linear regression models for *Fairness* and *PostAgain*, including the degree to which the user perceived the rules to be clear and read the rules, as independent variables, and accounting for the control variables (n=526). These models were built using only the responses from the participants who agreed that they had noticed the rules in the subreddit sidebar, and were therefore shown additional questions that pertained to the rules. Each of these models explained a significant amount of variance and had adjusted R^2 value of .124 (for *Fairness*) and .055 (for *PostAgain*), a notable improvement over the baseline models (Table 4).

Our results (Table 5) show that reading the rules was significantly associated with *Fairness* even after controlling for demographic and prior Reddit history variables. As shown in Table 5, when users read the rules, they are less likely to consider the removal as fair (β = -.314). This is surprising as one would expect that reading the rules would help users understand the expectations of the

Table 5. Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users are likely to post again on the corresponding subreddit (*PostAgain*). This model includes community guidelines variables as independent variables in addition to the control variables as inputs.

		Fairness				*PostAgain*			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	2.81	0.32		.000	2.744	0.317		.000
Control Variables	**Age**	-.056	.066	-.040	.403	.029	.066	.021	.664
	Education	-.072	.038	-.087	.061	-.036	.038	-.046	.342
	Gender	.162	.164	.041	.324	-.23	.163	-.062	.159
	Reddit Karma	-1.3E-7	.000	-.049	.357	-8.2E-8	.000	-.032	.558
	Reddit Age	.000	.000	.081	.068	-1.9E-5	.000	-.013	.770
	Submission Count	7.0E-5	.000	.064	.231	.000	.000	.131	.020
Community Guidelines Variables	**Read Rules**	-.308	.041	-.314	.000	-.063	.040	-.067	.119
	Rules are Clear	.222	.054	.170	.000	.278	.053	.223	.000

community and improve the perceived legitimacy of content removals. However, as we will discuss in Section 6, users sometimes have difficulties complying with the subreddit rules because they are too difficult to follow or subjective. Moreover, some users complain that their posts get removed despite compliance with the community guidelines. We did not find a significant relationship between reading the rules and *PostAgain* after accounting for control variables.

We also found that when users perceive the rules to be clear, they are more likely to consider the removal as fair (β = .170) and more likely to consider posting in the future (β = .223) (see Table 5). This indicates that clarity of rules has positive association with user attitudes.

In sum, these findings suggest that composing clearly written community guidelines can render the content removals more palatable to the users and motivate them to continue posting despite the current removal.

5.3 Removal Explanations

We examined how different aspects of removal explanations affect users' attitudes about content removals and future postings. Our results show that in 39.7% of cases (n=360), participants claimed that they were provided an explanation for their removal by the moderation team.

We built a regression model for *Fairness* (n = 738) using the independent variable *receivedExplanation* (this binary variable measures whether the participants received an explanation for post removal) and adding the six control variables (Table 2). This model showed that receiving an explanation was positively associated with perception of the removal as fair (β = .384, p < .001). Similarly, a regression model for *PostAgain* using *receivedExplanation* as an independent variable and including the control variables (n = 738) found that when participants receive an explanation, they are more likely to consider posting again in the future (β = .088, p < .05). These results suggest that explanations can be a useful mechanism to gain trust with the users.

We asked users who had received explanations (n=360) additional questions about the explanations, and analyzed the relationships between different aspects of explanations and user attitudes. Our results show that 64.2% (n=231) of these participants answered 'yes' to the question of whether the explanation provided them any new information, and the rest answered 'no'.

Next, we found that 57.8% of participants (n=208) received a removal explanation through a comment to their removed submission, 36.1% (n=130) received a private message explaining why their post was removed, and 6.1% of participants (n=22) had their submissions flaired with a short text that explained the post removal. Our results also showed that 20% of our participants who received explanations (n=72) felt that their removal explanation was provided by a human

Table 6. Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users consider it likely that they will post again on the subreddit (*PostAgain*). This model includes removal explanation variables as independent variables in addition to the control variables as inputs.

		Fairness				*PostAgain*			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	2.895	.379		.000	3.332	.382		.000
Control Variables	**Age**	-0.076	.096	-.053	.433	.077	.097	.054	.431
	Education	-0.095	.054	-.116	.077	.066	.054	.081	.227
	Gender	0.299	.225	.078	.185	-.143	.227	-.038	.530
	Reddit Karma	-2.3E-9	.000	.000	.998	5.9E-7	.000	.047	.524
	Reddit age	.000	.000	.174	.005	2.2E-5	.000	.014	.823
	Submission Count	4.7E-5	.000	.040	.588	4.5E-5	.000	.038	.609
Removal Explanation Variables	**Explanation Provides New Information**	0.286	.081	.200	.000	.051	.082	.036	.535
	Explanation mode	0.008	.084	.005	.925	-.089	.085	-.063	.295
	Explanation source	-0.045	.111	-.023	.683	-.054	.112	-.028	.627

moderator, 50.6% of participants (n=182) felt that a bot provided explanation, and the rest (n=106) were unsure.

In order to examine relationships between different aspects of removal explanations and users' perceptions of moderation, we created regression models that included the mode of explanation ('comment', 'flair' or 'private message'), source of explanation ('human', 'bot', or 'unknown'), and whether the user perceived the explanation as informative, as independent variables (n=305). These models were built using only the responses from the participants who agreed that they were provided an explanation about the post removal, and were therefore shown additional questions related to explanations. As in earlier models, we used *Fairness* and *PostAgain* as dependent variables and demographic and prior Reddit history variables as control variables. Table 6 shows the results of these analyses.

Results show that the only explanation factor that explained a significant amount of variance in the perception of removal as fair was considering the explanation as informative (Table 6). When users feel that explanations provide them information that is novel, they are more likely to perceive the removal as fair (β = .200). This is possibly indicative of cases where users mistakenly violate a rule or a community norm they weren't aware of when submitting a post and realize their mistake when receiving explanations. However, perceiving the explanation as informative did not have a significant association with *PostAgain*.

Neither the explanation mode nor the explanation source had a significant relationship with either *Fairness* or *PostAgain* (Table 6). Thus, the mode of explanation does not appear to be important to the users. Moreover, whether a human moderator or an automated tool provides the removal explanation does not seem to matter to the users. Through our experiences as Reddit moderators, we know that moderators often use pre-configured removal explanations in order to expedite moderation tasks. Thus, the text outputs for both human and automated explanations on many communities look quite general and not specific to the submission at hand. This may be the reason why the users seem to have similar responses to both human and bot explanations.

Comparing the different regression models (Table 4), we found that adding either posting context variables or community guidelines variables to the baseline (that includes only the control variables) increases the adjusted R^2 values by substantial amounts for both *Fairness* and *PostAgain*. However, adding removal explanations does not contribute as much increase in adjusted R^2 value for *Fairness*, and in fact, the value for *PostAgain* reduces over the baseline model. Combining different sets of

Table 7. Kinds of responses mentioned by participants (N=849). Total adds up to more than 100% since many participant responses fit into more than one category.

Theme	Frequency
Lack of clarity about why post was removed	36.9%
Frustration at post removal	28.7%
Perception of moderation as unjust	28.5%
Acceptance of removal as appropriate	18.3%
Frustration at (lack of) communication about post removal	15.7%
Indifference to post removal	12.0%
Difficulties complying with the rules	11.9%
Difficulties with use of automated moderation tools	3.9%
Greater understanding of how to post successfully	3.2%
Satisfaction of removed post receiving many responses	2.5%

variables, we found that the best-fitting models for *Fairness* (adjusted R^2 = .324) and *PostAgain* (adjusted R^2 = .123) were generated by including all the input variables (Table 4).

6 QUALITATIVE FINDINGS

In this section, we report the results of qualitative analysis of open-ended responses to our survey. These results add nuances to the findings presented above in Section 5. Our participants revealed often negative yet complex attitudes towards content moderation. Table 7 summarizes the 849 responses to the open-ended questions about perceptions of content moderation in the survey. We note that the total percentages add up to more than 100% since at times there were overlapping themes that emerged from user statements, and we classified such statements into more than one category.

For the remainder of this section, we use excerpts of quotes from respondents to show representative examples of emergent themes from each coded category in Table 7.

6.1 Frustration and Lack of Clarity about the Post Removal

28.7% of respondents felt frustrated at the post removals. A recurring theme among the responses of these participants is that they felt their efforts at content creation were "not appreciated" on the subreddit they posted to. Many of these participants mentioned being "embarrassed" by the removal while others reported feeling dejected and demotivated from engaging with the community. In line with the relationship we found between time spent in creating submission and users' attitudes through quantitative analysis (Section 5.1), participants' open-ended responses reflect that those who had spent considerable time and effort creating a submission felt particularly annoyed about the removals. For example, Participant P893 wrote:

> *"I was confused and angry. It was a good post that I spent half an hour making and there was nothing against reddit's (or the subreddit's) rules! There was no reason to remove the post and honestly, I'm quite furious."*

Participant P254, who stated that she is autistic, and posted on the r/disability subreddit, wrote:

> *"I am autistic and it takes significant effort for me to write up things to communicate effectively and in a way that will be received well. I feel sad that my effort in making that post was for nothing and that no one will see it and no one will reply with any help or advice."*

36.9% of respondents wrote about the lack of clarity in their post removals. A frequent complaint among these respondents was "I have no idea what I did wrong." 64 participants mentioned feeling "confused" about the removal. Many participants argued that they had seen posts similar to their

own removed post appear on the subreddit before, so they felt "non-plussed" why their post was targeted for removal. Some respondents pointed out that they were cautious in ensuring that they adhered to all the community guidelines, and yet, their post was removed. Such removals left users uncertain of how they can make successful submissions. For instance, Participant P779 wrote:

> *"I read the rules and my submission was within the guidelines, so I have no idea why it was removed and I'm a little annoyed about it."*

3.9% of respondents complained that their post was mistakenly removed by an automated moderation tool. Some of these participant expressed frustrations about the excessive reliance of moderators on automated tools that often make mistakes.

> *"Mods rely on bots too much. Sometimes there is no human to see why it was removed." - P55*

Content moderation is usually focused on the goal of curating the best possible content. However, given the large proportions of content removals [45] and the frustrations of a majority of moderated users as discussed above, community managers must consider how to assuage the users' frustrations that are linked to post removals.

6.2 Perception of Moderation as Unjust

28.5% of participants noted that the moderation was unjust. Some of these users felt that they are unfairly censored despite their adherence to the community guidelines because their posting history indicated an unpopular political affiliation. For example, one participant wrote:

> *"I used to argue with mods but since I participate in some edgy subreddits, lots of mods don't like me and will ignore me, even though I am not rude and my post follows the rules." - P594*

Many participants whose politically charged submissions were removed without notification created their own folk theories about why the removal occurred. Some users felt that the moderators on the subreddit they posted to were politically biased. Others worried that influential online communities often promote a particular worldview and that all the "dissenting voices" are removed. These participants often complained about their inability to exercise their "freedom of speech" and they felt they were being silenced. Some of these participants held that a small number of moderators, who are not elected by the community, had too much power over what gets seen by a large number of users. For example, Participant P23, who did not receive any removal notification, wrote:

> *"It's completely unfair. I didn't break the rules and just got punished for disagreeing with the mods' personal opinions."*

We found that 2.5% of participants justified the validity of their posts by pointing to the positive community response their posts received. They argued that the post removal was unwarranted because other users in the community interacted with the post in a supportive way but the moderators still decided to remove the post. For instance, Participant P815 wrote:

> *"Considering my post had almost 250 upvotes with 99% [up-votes to down-votes] ratio, I'd say everyone in the community enjoyed it and it shouldn't have been removed."*

A few of the participants revealed bigoted generalizations and conspiracy theories about how content moderation happens. For example, one participant wrote:

> *"Criticism of Jews is generally forbidden on Reddit for some reason. It's weird how we can question the teachings of Jesus and the Moon landing and the shape of the Earth, but we must never question the Jew." - P138*

In sum, many users have folk theories of content moderation being shaped by partisan community managers. To what extent these folk theories are accurate and reflect the existing biases of moderators is an interesting empirical question that warrants further research.

6.3 Acceptance of Removal as Appropriate

18.3% of respondents indicated an acceptance of their post removal as appropriate. Participants accepted the removal of their posts for a variety of different reasons. Many of these users acknowledged that they had not read the subreddit rules, and they felt that their post removal was valid because they inadvertently violated a subreddit rule. Some of these users expressed regret about not attending more carefully to the community guidelines. For example, Participant P736, who received a private message from the moderators explaining why his post was removed, said:

> *"I felt bad that I had not read the rules and posted inappropriately."*

3.2% of participants mentioned that the content removal helped them become more aware of the social norms of the community and provided them an understanding of how to post successfully in the future. For example, Participant P151 wrote:

> *"I will no longer post images to that subreddit now that I know not to."*

Some individuals explicitly described themselves as "a troll" and they expected that their content would be removed. These users found value in having their blatantly offensive posts be viewed by the community before it is taken down by the moderators. For example, Participant P857 characterized his own post as "disgusting" and he hoped that the post would generate "confused and funny reactions before it was removed." In a similar vein, Participant P375 wrote:

> *"I frequently participate in so called "shitposting" i.e I post content with little to no purpose or meaning behind it. The reasoning behind this is primarily for personal entertainment."*

Another group of users who accepted the removal as appropriate were those who suspected that their post would be removed but they submitted their posts anyway in order to show their group alignments. For example, one participant pointed out that he continues to post inappropriate content on the r/Patriots subreddit, an online community for supporters of the Patriots, an American football team, despite repeated removals because he hates that team. In a similar vein, Participant P90 described his behavior as motivated by a need to prosletyze, and did not feel bothered by the removals of his posts:

> *"Expected, but even if only one person repents, I feel I did what I wanted."*

Thus, participants who accepted the removals as fair include those who realize their mistakes and show an inclination to improve in the future as well as those who have a need to vent on Reddit and did not feel bothered by removals.

6.4 Communications about Post Removals

15.7% of respondents complained about the communication, or more frequently, a lack of communication from the moderation team about the post removal. One common sentiment was people reporting frustration about the silent removal of their posts. This reflects the relationship between users not noticing the post removal and perceiving removal as unfair that we found through our statistical analysis (Section 5.1). These users often felt cheated upon when they realized that their post was no longer available on the site. For example, Participant P85 wrote:

> *"Whatever I did wrong, the mods should have told me up front. I feel left out of the loop. I probably won't post anything there until I can find out exactly what the problem was."*

Many participants pointed out that they felt frustrated at not receiving any explanations for why their content was removed. This is in line with our findings from Section 5.3 that users who do not

receive removal explanations are significantly more likely to perceive the removal as unfair. Some users reported being more indignant at the lack of transparency about the moderation process rather than at the removal itself. For example, Participant P838 wrote:

> *"The removal is ok, doesn't bother me, but it's not ok that I didn't get informed."*

A few participants reported dissatisfaction with their interactions with the moderation team about the content removals. Some noted that even after they corresponded with the moderators, they could not understand why their post was removed. For example, Participant P737 wrote:

> *"I sent a polite and thoughtful private message to the mods asking how I could repost or avoid the problem in the future and was given a one sentence response... Also, the reply did not help me understand the mods' issue with my post."*

6.5 Difficulties Complying with the Rules

11.9% of respondents mentioned that they had difficulties complying with the community guidelines. Reddit is a decentralized platform and every Reddit community has its own set of posting guidelines, moderators and social norms. Some respondents who engaged in multiple communities found it tiresome to keep track of and comply with the posting guidelines for each new community they post on. For example, one participant wrote:

> *"I rarely post, in part because each subreddit has a long list of rules and I'm always concerned that I'll miss something, like I did this time. In too much of a hurry? Skimmed the list of rules too quickly? Rules can be difficult to locate when on mobile...There are 2 millions subreddits each with their own list of rules about what you're allowed to say...I just can't be bothered with it anymore." – P902*

Some participants did not understand the reasoning behind why certain rules were put in place on the subreddit. For instance, one participant expressed surprise at finding out that his post was removed because it violated the rule of mentioning the name of a subreddit. This participant was confused about why such a rule was put in place. Other respondents disapproved of certain subreddit rules, and chose to violate those rules deliberately despite suspecting that their post might be removed as a result. For example, one participant pointed out that he posed a question in his submission despite knowing that according to the rules of that subreddit, questions are only allowed to be posted in designated submission threads. He elaborated:

> *"[My submission] technically violates the rules. But having a large thread for questions makes it REALLY easy for things to get lost and never answered, which happens a lot in subreddits like this. So, I made a thread. Thought it might get removed, and it did...I think a rule that prohibits questions is a terrible idea. If it makes the subreddit spammy, I can understand, but you're stifling community interaction." – P194*

Many respondents complained that it takes a lot of effort to ensure compliance with certain subeddit rules when they post. For example, one participant said that complying with the rule that "a similar link hasn't been posted before" requires putting in a lot of work that involves searching through the prior posts of the subreddit, and it is easier to simply post the submission and hope that it does not get removed. Another participant wrote:

> *"The technical grounds for removal were accurate. However, the rules that qualify a post for removal are overly broad, arbitrarily enforced, and many times onerous to comply with." – P09*

Some participants felt frustrated that the subreddit rules were too subjective and could be interpreted in multiple ways. This mirrors our finding from Section 5.2 that users who found the

rules to be unclear were more likely to perceive the removal as unfair. A few respondents felt that the moderators deliberately designed unclear rules so that they could defend their removal actions.

> *"Well, my submission was removed because [of the rule] "it didn't fit the aesthetic of the sub" and I was under the impression that it did. In the future I think I may limit my submissions as I disagree with the sub moderators on what vapor wave aesthetic really is." - P132*

> *"There's always one really subtle rule you don't notice and then it gets removed ." - P707*

These instances of dissatisfaction with community guidelines partly explain the surprising finding from our statistical analyses that users who read the guidelines before posting submissions have negative attitudes about the community moderation (Section 5.2).

7 DISCUSSION

In this section, we discuss the implications of our findings, focusing on community guidelines, transparency of explanations, and nurturing of dedicated users.

7.1 Community Guidelines

In prior research, Kiesler et al. hypothesized that "explicit rules and guidelines increase the ability for community members to know the norms" [51]. Our findings add evidence to the value of establishing explicit posting guidelines in online communities. As we show in Section 5.2, participants who posted in subreddits containing rules were significantly more likely to consider their removal as fair. However, only about half of all Reddit communities have explicit rules [29]. Since having a list of posting rules is largely a one-time task, site managers should encourage voluntary moderators to establish rules in their communities. Recent research suggests that community norms and rules often overlap among different communities [13, 29]. Moderators of new communities may benefit by having tools that can suggest them which rules they should create, based on the similarity of their community with existing communities containing rules.

We contribute theoretical insights on the role that community guidelines play in shaping user attitudes. In prior research, Kiesler et al. hypothesized and Matias empirically showed that prominently displaying community guidelines helps increase users' adherence to those guidelines [51, 61]. Our findings add nuance to this result by showing that simply making users read the community guidelines may be insufficient to improve the long-term health of the community. Indeed, we found that when moderated users read the rules before posting, they are *less* likely to consider their post removal as fair (Section 5.2). We bring attention to the attributes of community guidelines that are important to end-users: their size (i.e., number of rules in the guidelines), subjectivity, reason why each rule is created, and effort needed to comply with the guidelines. Next, we discuss how the insights our findings provide about these different attributes of community guidelines may benefit platforms and community managers.

Our statistical analysis indicates that when users perceive the community rules to be clear, they are more likely to consider the removal as fair, and are more likely to consider posting again (Section 5.2, Table 5). In line with this, our qualitative findings suggest that users find it difficult to follow rules that are imprecise, subjective or require a lot of effort to comply with (Section 6.5). Thus, when community managers create rules, they should consider whether the rules are clearly laid out and easy to follow.

Prior research has shown that community guidelines are often created as reactions to short-term events or transitions [77]. New users, however, may not be aware of these past events. As our analysis shows (Section 6.5), some users do not comply with the community rules because they do not understand the reasoning behind why these rules were put in place. Many users do not realize

why certain rules are needed. Therefore, documenting the reasons for rule creations or explaining the need for such rules may help increase the acceptability of rules among new users.

When we evaluate the dynamics of users' compliance with community guidelines, it is important to consider that many social media users engage with multiple platforms [78]. Since different platforms may have different posting guidelines [66], it can be challenging for users to precisely follow these guidelines on each platform. Our findings suggest that adhering to guidelines becomes even more difficult in a multi-community environment such as Reddit where alongside the site-wide policies, each community has its own set of unique rules. As the list of rules in a community becomes longer, it becomes increasingly onerous for new users to attend to all the rules and make a successful submission. One approach to address this problem could be to introduce a pre-submission step where submitters are asked what type of post they are about to submit. Following this step, only the community rules relevant to that post type may be shown to the submitter. This may reduce the burden of verifying compliance with the entire set of rules for the submitters, and make it easier for them to post successfully.

Platforms can also design to make compliance with certain rules easier on the users. In particular, moderation systems can warn the user when they are about to post a submission that violates a rule whose compliance can be automatically verified. For example, if a community requires posts to be in a certain format, users should be alerted if they have the wrong format at the time of submission, and they should be allowed to edit their post to avoid future removal. This would also help assist users' understanding of the rules and the social norms of the community.

7.2 Transparency of Explanations

Our findings clearly highlight that lack of transparency about content moderation is a key concern among moderated users (Sections 5.1, 5.3, 6.1, 6.4). We found that many users felt confused about why their post was removed. Some participants were more frustrated at lack of notification about the removal than about the removal itself. At a broader level, transparency in moderation has important implications for our communication rights and public discourse, as pointed out by prior research [33, 81].

Our survey data show that in absence of information about why removal occurred, users often develop folk theories about how moderation works (Section 6.2). While these folk theories may be inaccurate, they influence how users make sense of content moderation and how they behave on the site. Therefore, more examination is needed of how users consume and integrate different information resources to create folk theories about moderation systems. Further, we must update our theories on best practices in moderation systems to account for how users' folk theories about these systems may influence their behaviors.

In their study of how users form folk theories of algorithmic social media feeds, DeVito et al. showed that most folk theories held by the users are "flexible, as opposed to closely-held, rigid beliefs" [21]. This suggests that designing "seams" into [26, 48] or providing more accessible information about how moderation systems work may influence users to revise their folk theories and improve their attitudes about the community. Yet, in his study of the effects of system transparency on trust among end-users, Kizilcec found that providing too much information about the systems eroded users' trust [52]. Indeed, in recent literature, researchers have begun asking questions about how much transparency is enough [27, 48], and how well transparency in design can serve the outcomes for important values like awareness, correctness, interpretability, and accountability [70]. Jhaver et al. also show that it is not appropriate for community managers to be fully transparent about how moderation works because bad actors may then game the system and post undesirable content that evades removal [44]. Therefore, community managers must carefully attend to the design of explanation mechanisms, scrutinizing what and how much information they reveal through these

processes. Similarly, they should cautiously consider the pros and cons of notifying the user of post removal versus silently removing posts. We also expect that not all post removals are similar. Future research should explore whether it is advantageous to distinguish between sincere users and bad actors when determining whether to provide removal notifications.

While the possibilities of being exploited by bad actors remains a challenge to implementing transparency in moderation, our results do indicate that informing the users *why* their post was removed through an explanation message can be a useful educational experience. Some of our participants pointed out that they felt more prepared to make successful posts in the future after they received explanations for post removals. Our statistical analysis also highlights that when removal explanations provide information that is new to the users, they perceive the removal as more fair (Table 6). Therefore, moderators must focus on improving the content of the explanation message, and make it relevant to the moderated posts. Additionally, the mode or source of removal explanation does not seem to matter to the users (Table 6). This suggests a potential for building automated moderated tools to deliver explanation messages. Such tools may help improve users' perceptions of content moderation without unduly increasing the work load of human moderators.

7.3 Are They Acting in Good Faith?

Our qualitative analysis suggests that there is considerable variance among users whose posts get removed. There are those who describe themselves as "trolls" or who post on Reddit for hateful or bigoted reasons (Section 6.3). Such users are not invested in contributing valuable information or fostering constructive discussions with others on Reddit communities. These users deliberately post content that they know is likely to get removed.

On the other end of the spectrum, moderated users include those who are emotionally engaged in the social life of Reddit communities. They invest substantial amounts of time and effort in contributing content to share with others on Reddit but mistakenly violate a community guideline or social norm and suffer removal. Such users feel embarrassed or unappreciated when their posts get removed (Section 6.1). Some users seek crucial advice from the community in their posts, and they feel dejected when their post is removed, as poignantly expressed by the autistic respondent who sought help from the r/disability community.

Social media platforms like Reddit have opened new avenues through which individuals seek not just informational but also emotional and social support. However, when users get moderated without appropriate feedback, they feel dejected, they are less likely to contribute further, and they may even leave the community. To allow a diverse set of users to participate in the digital public spheres, it is vital that moderation systems support not just the users who are in the know, but also those who may be unaware of the normative practices of online communities.

Just as importantly, we argue that platforms themselves may also benefit from nurturing users who are invested in learning from their mistakes and are just confused about where things went wrong. Users who attempt to post a submission on a community in good faith show a certain amount of dedication to contribute to the community. Therefore, if moderation systems offer them opportunities to participate and grow, they may turn into valuable contributors.

Given the variety of users who post online, moderation systems should find ways to distinguish sincere users from trolls and invest their resources in nurturing the former. Yet, as prior research shows, making such distinctions can pose many challenges [3, 4, 46, 47, 68]. For example, it may be problematic to classify a user as a bad actor if she posts an offensive message in response to another offensive post. Further, we may need to develop different strategies to address different types of bad actors [4, 46]. It might be worthwhile explaining the posting norms to a new user who posts an offensive joke, but not to another user who repeatedly posts disturbing content.

While it is necessary for moderation systems to deter bad actors, it is also important to nurture sincere users. Our statistical analyses show that when users spend more time in creating a post that is subsequently removed, they are less likely to consider posting again (Section 5.1). However, online communities need exactly such dedicated users to foster healthy growth. Therefore, information clues that can help moderators identify sincere users can be constructive. For instance, designing tools that allow moderators to easily notice the posts that took a long time for the submitter to create may help the moderators identify and engage with sincere users.

We note that although supporting users who have the potential to be valuable contributors is a worthy goal, there are other constraints and trade-offs that need to be considered. For example, moderator teams, particularly on platforms like Reddit where voluntary users regulate content, often have limited human resources. Such teams may prioritize removing offensive or violent content to keep their online spaces usable. Additionally, moderators may find more value in consistently enforcing their community guidelines regardless of the motivation of the contributor. Still, our findings suggest that online communities may find it useful to invest their resources towards nurturing users who show dedication to contribute well. For example, even when moderators have to remove sincere posts to consistently enforce community guidelines, they can contact post submitters and provide them actionable feedback on how to post successful submissions in the future.

7.4 Limitations and Future Work

Like all online survey studies, our study suffers from self-selection bias and social desirability bias. Users may be biased to obscure essential parts of the moderation experience that may present them in a negative light. Yet, we believe that the subjective perspectives of moderated users we present here provide important insights about the user experience on Reddit and can guide future design and moderation strategies.

While we account for some key demographic factors such as age, education, and gender in our analyses, we did not collect data on other important factors like race/ethnicity and sexuality that are cultural markers likely to influence how individuals perceive conflicts and silencing, either online or offline. These factors also help us understand who we are and aren't hearing from, and nimbly adjust to seek more responses from groups that are missing. We therefore recommend including these factors in subsequent research in this domain. Further, Reddit communities vary among one another on a wide range of factors such as their topic, norms, goals, and policies, but we could not control for such differences in our analyses. Exploring how these factors affect user responses to content moderation could be a productive direction for future work. Another limitation of our sample is that it skews heavily young and heavily male. While we suspect that this may reflect the actual demographics of Reddit users, this still results in an under-representation of the perspectives of other age groups and genders. Future work that focuses on under-represented age groups and gender may provide valuable insights.

We have only focused on post submissions for the purpose of this paper. However, users also comment on these posts, and moderation of comments may involve a different set of concerns. Analyzing how users perceive comment removals differently from post removals may be a fruitful area for future research.

8 CONCLUSION

We began this study to understand the perspectives and experiences of Reddit users whose posts have been removed. Our findings highlight users' frustrations with various aspects of content moderation: absence of removal notifications, lack of explanations about why posts were removed,

community guidelines that can be interpreted in multiple ways, and mistaken removals by automated moderation tools, among others. Lack of transparency in moderation resulted in many users creating folk theories about how content moderation happens. Suspicions about the political biases of moderators were commonly held among our participants.

Analyzing our findings led us to reflect on the question: How should we think about "fairness" in the context of content moderation? Is fairness of content removals whatever the community moderators think is appropriate to remove? Is fairness that unpopular users deserve to have their posts removed even if they believed in good faith to have been following the rules?

From the perspectives of our participants, fairness is associated with having a clear set of rules, getting informed when content removal occurs, and receiving explanations for the removal. Yet, is it fair to expect content moderators to invest their limited resources into these tasks? Even if the moderators take on these tasks, it is possible that such investments may not be productive in many instances, and they may open up opportunities for trolls to waste the time of moderators.

Still, it may be worth considering how a greater focus on transparency and explanations may affect the communities. Currently, moderation mechanisms remove content at massive scales, often without notifying the users of removal. On a positive note, our findings show that when moderator teams commit to transparency and provide removal explanations, users sometime learn from their mistakes and they feel better prepared to make successful submissions.

Therefore, an emphasis on education, rather than removal, may improve users' outlook towards the community and encourage them to participate constructively. Designing automated tools that can provide removal explanations in specific scenarios can help newcomers become familiar with the norms of community without unduly increasing the work of moderators. Creating community guidelines that are clear and easy to follow can improve users' perceptions of fairness in moderation. Ultimately, moderated users include many individuals who have made a deliberate effort to contribute to the community. Therefore, nurturing these users and attending to their needs can be an effective way to sustain and improve the health of online spaces.

ACKNOWLEDGMENTS

We are grateful to all our participants for being involved in this research. We would like to thank Benjamin Sugar, Josiah Mangiameli, Koustuv Saha, and Stevie Chancellor for their valuable inputs that improved this work. We also appreciate the AC and reviewers of this article for their constructive feedback and encouragement. Jhaver and Gilbert were supported by the National Science Foundation under grant IIS-1553376.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (may 2017), 211–236. https://doi.org/10.1257/jep.31.2.211

[2] Eva Armengol, Albert Palaudaries, and Enric Plaza. 2001. Individual prognosis of diabetes long-term risks: A CBR approach. *Methods of Information in Medicine-Methodik der Information in der Medizin* 40, 1 (2001), 46–51.

[3] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *Twelfth International AAAI Conference on Web and Social Media*.

[4] Lindsay Blackwell, Mark Handel, Sarah T Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding Bad Actors Online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, W21.

[5] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15, 3 (2013), 209–227.

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

[7] Amy S Bruckman, Jennifer E Below, Lucas Dixon, Casey Fiesler, Eric E Gilbert, Sarah A Gilbert, and J Nathan Matias. 2018. Managing Deviant Behavior in Online Communities III. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, panel02.

[8] Andrea Bunt, Joanna McGrenere, and Cristina Conati. 2007. Understanding the utility of rationale in a mixed-initiative system for GUI customization. In *International Conference on User Modeling*. Springer, 147–156.

[9] Giuseppe Carenini and Johanna Moore. 1998. Multimedia explanations in IDEA decision support system. In *Working Notes of the AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems*. 16–22.

[10] Matt Carlson. 2018. Facebook in the news: Social media, journalism, and public responsibility following the 2016 trending topics controversy. *Digital Journalism* 6, 1 (2018), 4–20.

[11] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (Atlanta, GA)*.

[12] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. This Post Will Just Get Taken Down: Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.

[13] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

[14] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[15] Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*. Vol. 24. Elsevier, 201–234.

[16] Danielle Keats Citron. 2009. Cyber civil rights. *BUL Rev.* 89 (2009), 61.

[17] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing Revenge Porn. *Wake Forest Law Review* 49 (2014). http://heinonline.org/HOL/Page?handle=hein.journals/wflr49{&}id=357{&}div=15{&}collection=journals

[18] Gabriella Coleman. 2014. *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso books.

[19] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163 arXiv:https://doi.org/10.1177/1461444814543163

[20] Laura DeNardis and Andrea M Hackl. 2015. Internet governance by social media platforms. *Telecommunications Policy* 39, 9 (2015), 761–770.

[21] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 120.

[22] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. Algorithms ruin everything:# RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3163–3174.

[23] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, HV Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, et al. 2017. Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML* (2017).

[24] Julian Dibbell. 1994. A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.* (1994), 471.

[25] Don A Dillman et al. 1978. *Mail and telephone surveys: The total design method*. Vol. 19. Wiley New York.

[26] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2371–2382.

[27] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 494.

[28] Casey Fiesler, Jessica L. Feuston, and Amy S. Bruckman. 2015. Understanding Copyright Law in Online Creative Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 116–129. https://doi.org/10.1145/2675133.2675234

[29] Casey Fiesler, Jialun Aaron Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*. 72–81.

[30] Megan French and Jeff Hancock. 2017. What's the folk theory? Reasoning about cyber-social systems. (2017).

[31] Archon Fung. 2013. Infotopia: Unleashing the democratic power of transparency. *Politics & Society* 41, 2 (2013), 183–212.

[32] Lex Gill, Dennis Redeker, and Urs Gasser. 2015. A human rights approach to platform content regulation. *Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.* (2015). https://freedex.org/a-human-rights-approach-to-platform-content-regulation/
[33] Lex Gill, Dennis Redeker, and Urs Gasser. 2015. Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights. (2015).
[34] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[35] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* 22, 6 (2019), 854–871.
[36] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.
[37] Ryan Grenoble. 2013. Facebook Reverses Stance On Beheading Videos, But Nipples Are Still A No-No (UPDATE) | HuffPost. https://www.huffingtonpost.com/2013/10/22/facebook-allows-beheading-videos-graphic-content_n_4143244.html
[38] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[39] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.
[40] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society* 18, 5 (2002), 371–384.
[41] Anna Lauren Hoffmann. 2019. Where fairness fails: On data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication, and Society* 22, 7 (2019).
[42] Allyson L Holbrook and Jon A Krosnick. 2009. Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly* 74, 1 (2009), 37–67.
[43] Christopher Hood and David Heald. 2006. *Transparency: The key to better governance?* Vol. 135. Oxford University Press for The British Academy.
[44] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243
[45] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2018. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 27.
[46] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). http://firstmonday.org/ojs/index.php/fm/article/view/8232
[47] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593
[48] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. , Article 421 (2018), 12 pages. https://doi.org/10.1145/3173574.3173995
[49] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for Civil Conversations: Lessons Learned from ChangeMyView.* Technical Report. Georgia Institute of Technology.
[50] Leo Kelion. 2013. Facebook lets beheading clips return to social network - BBC News. http://www.bbc.com/news/technology-24608499
[51] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).
[52] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.
[53] David A Klein and Edward H Shortliffe. 1994. A framework for explaining decision-theoretic advice. *Artificial Intelligence* 67, 2 (1994), 201–243.
[54] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004).
[55] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317 – 326. https://doi.org/10.1016/j.giq.2013.11.005
[56] Ganaele Langlois, Greg Elmer, Fenwick McKelvey, and Zachary Devereaux. 2009. Networked publics: The double articulation of code and politics on Facebook. *Canadian Journal of Communication* 34, 3 (2009).
[57] Gloria Mark, Yiran Wang, and Melissa Niiya. 2014. Stress and Multitasking in Everyday College Life: An Empirical Study of Online Activity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).*

ACM, New York, NY, USA, 41–50. https://doi.org/10.1145/2556288.2557361
[58] Nathan J. Matias. 2016. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference*. Oxford, United Kingdom.
[59] Nathan J. Matias. 2016. Posting Rules in Online Discussions Prevents Problems & Increases Participation. https://civilservant.io/moderation_experiment_r_science_rule_posting.html
[60] Nathan J. Matias. 2018. Gathering the Custodians of the Internet: Lessons from the First CivilServant Summit. https://civilservant.io/civilservant_summit_report_jan_2018.html
[61] Nathan J. Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.
[62] Nathan J Matias and Merry Mou. 2018. CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 9.
[63] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. *ICIS 2016 Proceedings* (dec 2016). http://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23
[64] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. In *Ways of Knowing in HCI*. Springer, 229–266.
[65] Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.
[66] 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 (GROUP '16)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297
[67] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering technology in discourse on discrimination. *Information, Communication & Society* 22, 7 (2019), 882–899.
[68] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.
[69] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 93–100.
[70] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.
[71] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
[72] Aja Romano. 2017. How the alt-right uses internet trolling to confuse you into dismissing its ideology. https://www.vox.com/2016/11/23/13659634/alt-right-trolling
[73] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kiciman, and Munmun De Choudhury. 2019. A Social Media Study on The Effects of Psychiatric Medication Use. In *ICWSM*.
[74] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *ICWSM*.
[75] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180.
[76] Mark Scott and Mike Isaac. 2016. Facebook restores iconic Vietnam War photo it censored for nudity. *The New York Times* (2016).
[77] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019), 1461444818821316.
[78] Aaron Smith and M Anderson. 2018. Social media use in 2018. Pew Research Center [Internet]. *Science & Tech. URl: http://www. pewinternet. org/2018/03/01/social-media-usein-2018/(visited on 04/16/2018)* (2018).
[79] Anselm Straus and Juliet Corbin. 1998. Basics of qualitative research: Techniques and procedures for developing grounded theory.
[80] Nicolas Suzor. 2018. Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media+ Society* 4, 3 (2018), 2056305118787812.
[81] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.
[82] 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 (2019), 18.
[83] Linnet Taylor. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society* 4, 2 (2017), 2053951717736335.
[84] Joseph E Uscinski, Darin DeWitt, and Matthew D Atkinson. 2018. A Web of Conspiracy? Internet and Conspiracy Theory. In *Handbook of Conspiracy Theory and Contemporary Religion*. BRILL, 106–130.

[85] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.
[86] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* (2018).

A SURVEY QUESTIONNAIRE

We asked our survey participants the following questions:

1. How much time did you spend in creating this submission?:
   a. < 1 minute
   b. 1-5 minutes
   c. 6-10 minutes
   d. > 10 minutes
2. Which subreddit was the submission posted to?:
   _
3. What is your Reddit username?
   _
4. Before I started this survey, I noticed that this submission was removed:
   a. Yes
   b. No
5. Before I posted this submission, I suspected that it would be removed:
   a. Strongly agree
   b. Agree
   c. Neutral
   d. Disagree
   e. Strongly disagree
6. (If a or b to Q 5): Why did you think it would be removed? Please explain:
   _
7. I think that the removal was fair:
   a. Strongly agree
   b. Agree
   c. Neutral
   d. Disagree
   e. Strongly disagree
8. Please explain how you felt about the removal:
   _
9. Does the subreddit contain rules in its sidebar?
   a. Yes
   b. No
   c. Unsure
10. (If yes to Q 9): I read the rules of the subreddit before posting:
    a. Strongly agree
    b. Agree
    c. Neutral
    d. Disagree
    e. Strongly disagree
11. (If yes to Q 9): The rules on this subreddit are clear:
    a. Strongly agree
    b. Agree

c. Neutral
d. Disagree
e. Strongly disagree

12. Did you notice a comment, flair or private message indicating why your submission was removed?
    a. Yes
    b. No
13. (If yes to Q12) The subreddit provided the reason for why your submission was removed:
    a. Through a comment to the submission
    b. Through a private message
    c. Through a flair to the removed submission
14. (If yes to Q12) Did the removal reason provide you information that you didn't know before?:
    a. Yes
    b. No
15. (If yes to Q12) The removal reason was provided:
    a. By a human
    b. By a bot
    c. I am unsure
16. This experience changes how I feel about posting on this subreddit in the future:
    a. Strongly agree
    b. Agree
    c. Neutral
    d. Disagree
    e. Strongly disagree
17. (If not c to Q16) Please explain why you feel differently about posting again:
    _
18. How likely are you to post again on this subreddit after this experience?
    a. Very likely
    b. Likely
    c. Neutral
    d. Not likely
    e. Very unlikely
19. Is there anything else you'd like to tell us about your view of this removal?

_ _

20. Which country do you live in?

_ _

21. What is your age?
    a. < 25 years old
    b. 25-34 years old
    c. 35-44 years old
    d. 45-54 years old
    e. 55-64 years old
    f. 65-74 years old
    g. 75 years or older
    h. Prefer not to answer
22. What is the highest level of education you have completed?
    a. Less than high school

b. High school graduate (includes equivalency)
c. Some college, no degree
d. Associate degree
e. Bachelor's degree
f. Master's degree
g. Doctorate degree
h. Prefer not to answer

23. What is your gender?
    a. Female
    b. Male
    c. Another gender
    d. Prefer not to answer
24. Would you be willing to participate in a short follow-up interview?
    a. Yes
    b. No

(If yes to Q 24) Please provide your email ID here so that we can contact you for an interview

– –

Received April 2019; revised June 2019; accepted August 2019