AppealMod: Inducing Friction to Reduce Moderator Workload of Handling User Appeals

SHUBHAM ATREJA, University of Michigan School of Information, USA
JANE IM, University of Michigan School of Information and the Division of Computer Science & Engineering, USA
PAUL RESNICK, University of Michigan School of Information, USA
LIBBY HEMPHILL, University of Michigan School of Information and ICPSR, USA

As content moderation becomes a central aspect of all social media platforms and online communities, interest has grown in how to make moderation decisions contestable. On social media platforms where individual communities moderate their own activities, the responsibility to address user appeals falls on volunteers from within the community. While there is a growing body of work devoted to understanding and supporting the volunteer moderators' workload, little is known about their practice of handling user appeals. Through a collaborative and iterative design process with Reddit moderators, we found that moderators spend considerable effort in investigating user ban appeals and desired to directly engage with users and retain their agency over each decision. To fulfill their needs, we designed and built AppealMod, a system that induces *friction* in the appeals process by asking users to provide additional information before their appeals are reviewed by human moderators. In addition to giving moderators more information, we expected the friction in the appeal process would lead to a selection effect among users, with many insincere and toxic appeals being abandoned before getting any attention from human moderators. To evaluate our system, we conducted a randomized field experiment in a Reddit community of over 29 million users that lasted for four months. As a result of the selection effect, moderators viewed only 30% of initial appeals and less than 10% of the toxically worded appeals; yet they granted roughly the same number of appeals when compared with the control group. Overall, our system is effective at reducing moderator workload and minimizing their exposure to toxic content while honoring their preference for direct engagement and agency in appeals.

CCS Concepts: • **Human-centered computing** → **Field studies**; *Natural language interfaces*; **Social media**; **Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: online content moderation, moderation tools, contestability, collaborative design, field experiment, effort asymmetry, friction, self-selection

ACM Reference Format:
Shubham Atreja, Jane Im, Paul Resnick, and Libby Hemphill. 2024. AppealMod: Inducing Friction to Reduce Moderator Workload of Handling User Appeals . *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 19 (April 2024), 35 pages. https://doi.org/10.1145/3637296

CONTENT WARNING: *This paper contains offensive language, including misogynistic slurs, that readers may find disturbing.*

Authors' addresses: Shubham Atreja, satreja@umich.edu, University of Michigan School of Information, Ann Arbor, Michigan, USA; Jane Im, imjane@umich.edu, University of Michigan School of Information and the Division of Computer Science & Engineering, Ann Arbor, Michigan, USA; Paul Resnick, presnick@umich.edu, University of Michigan School of Information, Ann Arbor, Michigan, USA; Libby Hemphill, libbyh@umich.edu, University of Michigan School of Information and ICPSR, Ann Arbor, Michigan, USA.

ACM 2573-0142/2024/4-ART19
https://doi.org/10.1145/3637296

1 INTRODUCTION

As content moderation becomes a central aspect of all social media platforms and online communities, interest has grown in how to make moderation decisions contestable [55, 56]. Particularly as certain decisions (e.g., permanently banning users) can have long-term consequences, moderation systems need to ensure they allow users to appeal individual decisions in case of error or injustice. For platforms and communities, appeals provide a mechanism to evaluate individual decisions. For users, they offer opportunities to understand the moderation policies and modify their behavior [43].

On social media platforms that empower individual communities to manage and moderate their own activities, the responsibility to address user appeals falls on volunteers within the community [14, 49]. These communities include groups on Facebook, subreddits on Reddit, and channels on Twitch and YouTube. Within these communities, volunteer moderators perform all moderation tasks from formulating and enforcing rules to explaining their decisions and considering appeals [52, 57]. Consequently, a growing body of content moderation research [5, 6, 23, 32, 50, 60] focuses on understanding and supporting volunteer moderators' work. For instance, some research aims to help moderators identify and remove problematic content [6, 24]. However, other aspects of volunteer moderators' work, particularly how they handle user appeals, have received little attention.

To understand moderators' handling of user appeals, we turned to prior work on social media users' contestability needs [55, 56]. For instance, Vaccaro et al. [55] found that when users contest moderation decisions, they almost always request a human review and prefer to directly engage with the moderators. However, directly communicating with users on every appeal can amount to a significant workload for volunteer moderators who are already overwhelmed with their work, and suffer occupational stress and burnout [9, 49]. Directly engaging with users also puts moderators at an increased risk of being exposed to toxic content from community members who are angry or upset about their decision [13, 57]. Ideally, a successful appeal process would honor both users' and moderators' needs.

In this paper, we investigated Reddit volunteer moderators' current practices of handling appeals from users who are banned from their community, and worked with them to design a system to help them process appeals fairly and efficiently. On Reddit, volunteer moderators cannot control who joins their community, but they have the power to ban members. As a result, moderators actively use banning as a strategy to keep problematic users away from their community [52]. Furthermore, given the long-term consequences of a permanent ban, Reddit requires that moderators must allow users to appeal bans.

Through a collaborative and iterative design process with Reddit moderators, we identified their strategies and needs for addressing appeals from banned users. Our interviews and design sessions with moderators revealed an effort asymmetry between moderators and users. While users could submit an appeal with minimal effort using Reddit's current system, moderators described spending considerable time and effort investigating and reviewing these appeals. They assessed whether a banned user had reflected on their behavior and tested the user's awareness of community norms. Moderators also desired to directly engage with banned users, despite the increased risk of receiving toxic messages, and to retain their agency over final decisions.

To fulfill these needs, we created AppealMod, a system that helps Reddit moderators to process appeals from banned users. AppealMod shifts the onus to banned users to complete their appeal by answering additional questions about their past behavior and their understanding of the community norms. To reduce the burden on moderators, appeals are hidden from their view until after users complete the AppealMod process. Completing the process restores the direct engagement channel between banned users and moderators, who then make the final decision. By asking users to put

more effort into their appeal initially, AppealMod effectively induces *friction* in the appealing process. We expected this friction to lead to a selection effect among users as some users would be discouraged by the AppealMod process, ideally exactly those users who had insincere or toxically worded appeals. We also hoped that the extra information provided by those appellants who completed the process would save effort for the moderators in assessing the appeals.

We conducted a field experiment to evaluate AppealMod in r/pics, a subreddit of over 29 million users.[1] The experiment lasted for four months. Appealing users were randomly assigned to a control or a treatment condition. Under control, users followed Reddit's existing process to submit their appeal. Users under treatment were subject to the AppealMod process for submitting their appeal. Results from the experiment show that roughly 70% of appealing users under treatment were discouraged by the AppealMod process and abandoned their appeal. More specifically, we found that users making insincere appeals or using toxic language were more likely to abandon the process. Therefore, while moderators reviewed only 30% of the appeals under treatment, they still granted roughly the same number of appeals under control and treatment. AppealMod also offered moderators some protection from toxic content. 91.3% of the toxic appeals subject to the AppealMod process were abandoned and remained hidden from the moderators. We conclude the paper by discussing implications for inducing friction in moderation processes to reduce the workload of volunteer moderators and potential next steps for improving the design of AppealMod.

2 BACKGROUND AND RELATED WORK

We first review prior research on volunteer moderators' experiences and needs, as well as existing work on designing tools for moderators. We then review research on contestability of content moderation decisions. Lasty, we discuss costs and friction-based techniques to reduce volunteer moderators' workload.

2.1 Volunteer Moderator Workload and Experiences

Schöpke-Gonzalez et al. [49] define volunteer content moderators as "individuals who uphold their own online community's standards". Many social media platforms empower members of individual communities to manage and moderate their own activities [26]. This includes groups on Facebook, subreddits on Reddit, and streaming channels on Twitch and YouTube. Matias [40] has described the work of these volunteer moderators as a form of "civic labor". They are often members of the community who care about its growth and development [51]. These volunteers formulate the rules and standards of their community and also enforce them, which involves reviewing and removing potentially problematic content and users, explaining their decisions, and considering appeals on their decisions [14, 32, 52, 57]. As a result, volunteer moderators often find themselves overwhelmed with the varied demands of their work [9, 49].

In addition to being labor-intensive, moderating online communities can also take an emotional toll on the moderators [47, 49, 53]. When reviewing content, moderators have to look at some of the most toxic content the Internet has to offer, leading to significant psychological distress [13, 57]. Therefore, volunteer moderation has also been classified as a form of "emotional labor" [9]. Furthermore, as volunteer moderators make decisions about what is permitted, they face harassment and abuse from community members who are angry or upset about a moderation decision [35, 57].

While platforms are responsible for providing tools for volunteers to be able to perform moderation adequately, these tools often fall far short of addressing the moderators' needs [24, 32]. For example, the prolonged neglect of moderation software on Reddit was so frustrating that it pushed

[1]As of January 2023, the community had 29.7 million users.

moderators to join together in protest against the platform [39]. In light of this, any tool that helps volunteer moderators manage their workload or offers protection from toxic content will likely both benefit moderators and help the community grow. Our work builds on the existing literature to design and deploy a system that can address volunteer moderators' needs.

2.2 Designing Tools for Volunteer Moderators

Given the inadequacy of platform-provided moderation tools, volunteer moderators increasingly rely on third-party tools, which include those developed by researchers. For instance, Automod, the most popular tool on Reddit, was first developed by a volunteer moderator and later adopted by Reddit [23]. Automod can be independently programmed by moderators to automatically detect (and act against) content that violates the rules of their community [59]. Relying on external tools is not unique to volunteer moderators on Reddit. Cai and Wohn [5] found that streamers on Twitch also relied on third-party applications to accomplish their moderation tasks. The most common use of these tools included identifying potentially problematic content and taking action against problematic users, e.g., banning them [5].

Consequently, there is a growing thread of content moderation research focusing on the study and design of tools for volunteer moderators. This body of work has mainly resulted in an assortment of techniques and tools for detecting and acting against problematic content and users (e.g., word-filters [24], machine learning classifiers [6], forecasting tools [34]). In contrast, other aspects of volunteer moderators' work have received little attention. One notable exception is PolicyKit, developed by Zhang et al. [60] to support moderators' task of managing and formulating their community rules. However, there has not yet been research on designing tools for supporting moderators' handling of user appeals, an essential task that volunteer moderators perform [52]. In this work, we focus on moderators' handling of appeals from banned users and present a system aimed at supporting this workload. Banning is an important strategy that moderators use to remove problematic users from their community [52], and given the long-term consequences of a permanent ban, it is essential that moderation systems allow users to appeal against their ban.

In order to be successful, tools for volunteer moderators must attend to how they are situated relative to moderators' existing work, processes, and control. Automating one aspect of the moderators' work may create manual work in other aspects of their role. For instance, Jhaver et al. [23] found that moderators' use of automated bots created new challenges of training and coordination among moderators and also added the new task of maintaining these bots [23]. Automation is also in tension with agency [4, 17, 27]. Moderators have resisted replacing rule-based tools with machine learning approaches [24], in part because they want to maintain agency over a system's decisions and impacts. They also find the ML systems difficult to understand and control, even when they outperform rule-based systems [6].

In our research, we adopted a collaborative and iterative design approach to create a tool informed by moderators' existing practices and needs so that it can seamlessly integrate into their workflows. We first built connections with volunteer moderators from different communities and invited them to participate in a research collaboration. We conducted a series of synchronous and asynchronous sessions to uncover their needs for addressing user ban appeals, including their need to maintain agency over each individual decision. We then designed a system to address their needs and conducted a field experiment to evaluate its effectiveness.

2.3 Designing for Contestability of Moderation Decisions

As we noted earlier, there is no prior work on understanding volunteer moderators' workload of handling user appeals. Designing for contestability from the users' point of view, however, is an important area of research [55, 56]. Moderation systems can frequently make incorrect decisions

[18, 28], and it is crucial for users to be able to appeal individual decisions as incorrect moderation decisions can cause significant negative outcomes for users [8] and social media platforms [31] alike. Facebook [58], Instagram [2], and Twitter [46] all recently introduced or updated their processes for appealing content moderation decisions. However, these systems are subject to frequent criticism from the users [43, 55]. For instance, most appeal systems do not provide a space for users to explain their behavior [55]. Some users describe the appeal process as "speaking into a void" [43] for its lack of direct human interaction.

Recently, researchers have started to explore alternate designs of appeal process. Vaccaro et al. [55] found that when users have the option to add an explanation to their appeal, they actively use that space to defend their behavior, share additional context, or even raise questions about moderation policies. Researchers have also argued for "scaffolding the appeal" [55], i.e., the appeal process should make clear what information moderators would consider, for instance by providing a structured form, because many users do not know what to say in their appeals or how to persuade the moderators [55]. Furthermore, by laying out factors that are considered in the decision-making, scaffolding can bring transparency to the process and reduce perceived inconsistencies in moderation decisions [56]. At the same time, Vaccaro et al. [55]'s study did not find any differences in the subjects' fairness perceptions toward different appeal processes. However, it is worth noting that in their online experiment, participants were given hypothetical scenarios in which the participants' written appeals could not change the final outcome—which remained undesirable for the user. Therefore, it is still worth evaluating alternative appeal processes in a real-world setting, especially as prior research has shown that moderation outcomes can impact people's satisfaction with the process [45].

Prior work also notes that social media users considered communication, especially direct communication with human moderators, as one of the top three avenues for improving their abilities to contest content moderation decisions [56]. When writing their appeals, many users explicitly requested a human review [55]. The requirement of a human review is becoming increasingly important as more decisions are taken via automated systems. Both scholarly and legal frameworks have underlined the importance of human review in their conceptualization of contestability [1, 36]. For instance, Sarra [48] advocates for contestability in the form of dialectical exchanges between users and decision systems that allow users to ask questions and collect any additional information needed to contest their decisions. However, challenges still remain in terms of the scaling of human review while also tackling bias and decision fatigue [36].

In sum, prior research has underlined the importance of allowing users to contest moderation decisions and provided multiple design recommendations for supporting users' contestability needs. While we designed AppealMod to fulfill moderators' needs while addressing user appeal, our system nevertheless implements many of the design recommendations that can support users' contestability needs as well. In particular, AppealMod provides scaffolding to users by sharing a webform with questions that moderators would consider in their decision making. Furthermore, all users who complete the AppealMod process get an opportunity to interact with the moderators and have their appeals individually reviewed and responded to by a human.

2.4 Cost and Friction for Reducing Problematic User Behavior in Content Moderation

Part of what makes handling user appeals challenging for moderators is the volume of appeals submitted. For instance, in 2021, moderators permanently banned over 4.9 million users from various communities on Reddit [22]. Even a small proportion of users (say 5%) appealing their decision will result in roughly 250,000 appeals for the moderators. One way to reduce the volume of a particular behavior is to make it costly via direct monetary payments [11, 44]. For instance, an online community could require an entry fee to join the community or a fee to submit an appeal of

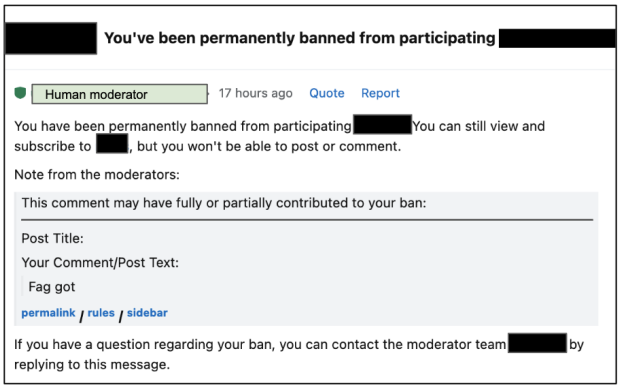

Fig. 1. An example ban message shared with the banned user.

a decision. This is different from the fees charged for an online service, such as streaming platforms [54], in that the appeal fee is intended to make users recognize the cost of their actions, and thereby reduce problematic behaviors [15].

Alternatively, one can think of the costs more broadly as a form of *friction* [26], perhaps by a specific task(s) that requires time and effort [15]. For example, in an attempt to reduce email spam, researchers recommended requiring email senders to compute a moderately expensive function before their message would be routed to a recipient [10]. An example more relevant to social media is Reddit's decision to quarantine problematic subreddits (e.g., r/the_donald). Quarantined subreddits still function, but they are no longer publicly accessible and users must explicitly opt in to access the community [7, 25]. Platforms also employ a mix of both monetary cost and time or effort-based friction. An online community, MetaFilter, charged a $5 entry fee and imposed a waiting period before new members could make new posts [29, p. 200]. In addition to raising a modest amount of revenue, the monetary cost and waiting period were acceptable to members who were a good match for the community, but not for those who might have disrupted it.

However, inducing friction may unintentionally discourage all user behavior, including legitimate user appeals or reports. As a major example, research has shown that users found filing reports of online harassment on platforms to be burdensome and ineffective [3, 37, 38]. Thus, the design challenge here is to induce friction that can lead to a selection effect among users and deter only undesirable participation. In this work, we present a system that reasonably increases the effort cost of submitting a ban appeal. Appealing users must provide additional information for their appeals to be reviewed by moderators. However, for those with legitimate appeals the effort is not wasted; the process solicits information that moderators normally use to evaluate appeals, and successful users would have provided it anyway. Users with insincere appeals or those who want to send toxic messages to moderators are more likely to be discouraged by the effort and abandon their appeal.

3 STUDY CONTEXT: BAN APPEALS ON REDDIT

In this section, we provide an overview of Reddit's current ban appeals process. On Reddit, moderators of individual communities are empowered to manage their own activities. Each individual community is called a subreddit. In terms of membership, moderators cannot control who joins their subreddit but they can ban users from participating in it. Moderators report banning users is

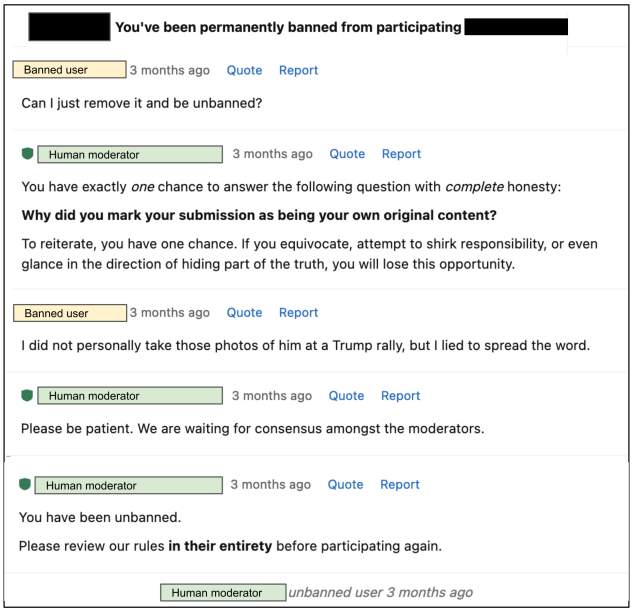

Fig. 2. An example interaction between an appealing user and human moderators following the user's ban.

an effective strategy for reducing spam and trolling behavior in their community [52]. According to Reddit's transparency report, over 4.9 million users were banned in 2021 across all subreddits [22].

When a user is banned from a subreddit, they receive a message from the moderators of that subreddit, informing them of their ban (and the ban reason) via modmail. Modmail is a shared messaging system on Reddit that moderators can use to communicate with individual members of their subreddit and vice versa. Figure 1 shows an example ban message that was sent to one of the users from the study subreddit. The user was banned for intentionally misspelling the word fa***t to try and get around the subreddit's moderation bot which automatically deletes any comments containing that word.

Reddit allows users to appeal against their ban by responding to their ban message. The first message sent by a user in response to their ban is automatically classified as their ban appeal. Ban appeals often lead to follow-up conversations between moderators and the banned user. Figure 2 shows an example conversation between a banned user and moderators of our study subreddit. First, the banned user responds to their ban message to initiate their ban appeal process. The subreddit has a strict policy on when users can claim a picture to be their "original content", and the user was banned for breaching that policy. The moderator responds to the user's appeal by asking them to further explain their behavior. After some back and forth with the banned user, the moderators decide to grant their appeal.

Generally, moderators may adopt any of the following course of action in response to a user appeal – 1) immediately deny or grant the appeal, 2) engage with the user in a follow-up conversation (and then deny or grant the appeal), 3) ignore and hide their appeal, or 4) mute the user and prevent them from sending any more messages for up to 28 days. Users may be banned temporarily (for a fixed period of time, ranging from weeks to a month) or permanently. For our study, we only consider permanent bans, as users are much more likely to appeal permanent bans due to their long-term consequences. While users can create multiple accounts on Reddit, using an alternate account to

circumvent their ban is prohibited, and usually results in the user account being suspended from Reddit as a whole[2].

We iteratively designed and built a system called AppealMod to support moderators' work on reviewing such ban appeals. We evaluated the system by deploying it in the subreddit called r/pics. r/pics is one of the largest communities on Reddit with over 29 million total members[3]. It is the most popular place for sharing photos on Reddit and frequently features on Reddit's front page, which is visible to all Reddit users. The r/pics community is an attractive avenue for users who want to share their visual art with a broader audience; at the same time, its large membership and popularity also invites spam and other problematic behaviors. Overall, this results in a significant workload for the volunteers who are responsible for moderating the community. The moderators have drafted strict policies to govern the kind of content that is allowed in the community and often use automated tools to enforce some of these policies. The large number of moderation decisions that are made everyday (many of which are automated) invite further complaints and appeals from the users. Moderators' preference to manually review these complaints further adds to their workload. Through AppealMod, we aim to reduce the volunteer moderators' workload of reviewing user appeals. All screenshots provided in the paper are actual conversations between users and moderators of r/pics. Moderators provided permission to use the conversations in our paper. The screenshots have been anonymized and edited for clarity.

4 DESIGN PROCESS AND MODERATOR NEEDS

We used a collaborative and iterative design process to understand Reddit moderators' current practices of handling user ban appeals and design a system that addresses their needs. We began forging connections with Reddit moderators by interviewing nine moderators from different communities and inviting them for a future research collaboration. Then, we conducted three interviews to understand moderators' current practices and needs for handling user ban appeals. Finally, we focused on one community as a use-case and conducted design sessions with the moderation team to iterate over the design and evaluate it in a real world setting. The moderators generously shared their time, expertise, and ideas with us, and this deep, continuous engagement is a hallmark of our project. In the next two subsections, we describe our design process, provide an overview of the moderators' current practice of handling user ban appeals, and summarize their key needs that informed the design of our system.

4.1 Method: Collaborative and Iterative Design Process with Moderators

4.1.1 Building connections with moderators. Fostering connections with community moderators is integral for developing a system that will be used by and deployed within the community. We started by interviewing nine moderators who collectively managed over 100 different subreddits. We selected these moderators via purposive and snowball sampling. Given our emphasis on building partnerships with moderators, we customized recruitment messages and focused on communities where our first author could demonstrate commitment and interest. This allowed the first author to express their shared interests in the growth and welfare of the community while inviting the moderators to participate in the study. An example recruitment message is provided in the Appendix A.1

Details about our participants are provided in Table 1. We were broadly interested in understanding the major challenges moderators faced, and the interview scope was not restricted to their handling of user ban appeals. The first author, who is an active participant in several Reddit

[2]https://www.reddit.com/r/modnews/comments/wrnnvb/piloting_a_new_ban_evasion_tool/
[3]The subreddit had roughly 29.8 million users as of January 2022

Table 1. Details of moderators who participated in the design process

Participant identifier	Subreddit	Subreddit topic	Subreddit size	# of subs moderated	Moderation experience	Gender	Country	Participated in:		
								Building connections	Understanding mods' needs	Iterative design sessions
P1 (SOC_M1)	r/soccer	sports	3.8 m	1	6 yrs	M	Argentina	x	x	
P2	[redacted]	hobbies	19m	12	3 yrs	F	India	x		
P3 (POL_M1)	r/politicalhumor	satire	1.5m	56	10 yrs	M	US	x	x	
P4	[redacted]	lifestyle	21m	26	4 yrs	F	US	x		
P5 (PIC_M1)	r/pics	pictures	29m	26	11 yrs	M	Scotland	x	x	x
P6	[redacted]	history	1.6m	7	12 yrs	M	Sweden	x		
P7	[redacted]	education	40k	1	9 yrs	M	US	x		
P8	[redacted]	racial identity	5.6m	24	8 yrs	M	US	x		
P9	[redacted]	lifestyle	8m	21	3 yrs	F	UK	x		
P10 (PIC_M2)	r/pics	pictures	29m	19	7 yrs	M	Israel			x
P11 (PIC_M3)	r/pics	pictures	29m	2	6 yrs	M	UK			x
P12 (PIC_M4)	r/pics	pictures	29m	34	8 yrs	M	N/A			x
P13 (PIC_M5)	r/pics	pictures	29m	42	7 yrs	M	N/A			x

communities, conducted all interviews between April and July 2021. The interviews were conducted in English over Zoom and lasted for approximately 1 hour. During the interviews, moderators elaborated on the major issues they face, their coping strategies, and the scope for technology to support their work. At the end of the interviews, we asked moderators if they would be interested in a future research collaboration. Among the nine participants, three moderators expressed interest in a future collaboration and signed up for another round of interviews. These included moderators of r/pics, r/soccer, and r/politicalhumor.

4.1.2 Understanding moderators' handling of users ban appeals. In the second round of interviews, we focused on the particular task of handling ban appeals on Reddit. Moderators during the first round repeatedly brought up their workload of addressing ban appeals from users. Furthermore, as we reviewed prior studies that focused on volunteer moderators, we found little evidence on understanding or supporting the moderators' current practices for addressing user ban appeals (see Section 2.2 for a detailed review of prior work). All three moderators who signed up for a potential collaboration further expressed their interest in supporting the design of a moderation tool for addressing user ban appeals.

As we conducted a second round of three interviews with these three moderators, we centered the narrative around previous instances of ban appeals. For instance, when moderators described their potential approaches to address a ban appeal, we asked them to find examples of previous appeals that demonstrate this approach. We also asked follow-up questions about their use of technological tools and other resources during this process. Lastly, we asked them for any potential design ideas to reduce their workload. The interviews lasted between roughly an hour to an hour-and-a-half. All interviews were conducted in English over Zoom and transcribed by the first author.

We analyzed the interview data using interpretive qualitative analysis [41] approach as outlined in [24]. In particular, we began with open coding [19] to categorize our data into relevant patterns and then group them into appropriate themes. Next, we engaged in multiple subsequent rounds of coding and memo-writing, conducting a continual comparison of codes and associated data. The codes and emerging themes were discussed in weekly project meetings among coauthors. The first-level codes were specific, such as moderators looking for an apology, low effort appeals from users, etc. After several rounds of iteration, the codes were condensed into high-level themes, such as getting more information from users and effort asymmetry between moderators and users. Once authors agreed upon the themes, we used them to generate the design ideas that were used in the iterative design process. The themes are described in Section 4.2.

4.1.3 Iterative design process. During interviews, all moderators emphasized the need to solicit additional feedback from their community's entire moderation team before finalizing the design and evaluating it in a real setting. Therefore, for our iterative design process, we decided to focus on designing for one of the three communities, r/pics. r/pics is one of the largest subreddits with over 29 million total members. It is the most popular place for sharing photos on Reddit and frequently features on Reddit's front page, which is visible to all Reddit users.

In September 2021, we shared a detailed proposal with all moderators of r/pics via modmail, introducing and explaining the design of our new tool and the evaluation process. The moderation team gave largely positive feedback and expressed interest in supporting the design and evaluation of the tool. At the time, r/pics had 8 moderators, and we invited all of them to participate in the design process. Between October and November 2021, we held two synchronous design sessions on Zoom. The protocol we followed for these design sessions is provided in the Appendix A.2.1. Three of the moderators attended these sessions, and each lasted for approximately one hour. In between these sessions, we also held a series of asynchronous discussions (over modmail) with the r/pics moderators. Five out of the eight moderators provided feedback during these discussions. For our design process, we used multiple low and mid-fidelity prototypes to present the design ideas generated from our earlier qualitative analysis. For instance, we first used a flowchart to provide moderators with an overview of the process. The flowchart highlights that appealing users will receive automated messages in response to their requests and that some of the users' requests may remain hidden from moderators. A snapshot of the flowchart is provided in the Appendix (Figure 8). In addition to the flowchart, we also shared a potential typology for questions to be asked of the users during the process (see Figure 9 in Appendix). Finally, to solicit additional feedback on our proposed automated bot, we provided moderators with a PowerPoint mockup conversation between appealing users and the bot (see Figure 10 in Appendix). These materials were sent to the moderators in advance and used during the design sessions. To keep this section concise, we included our design materials in the Appendix.

All moderators were offered a compensation of USD $20 (or an equivalent amount in their local currency) for each session or interview they participated in. One moderator respectfully declined the compensation. The entire design process was reviewed and determined to be exempt from oversight by our Institutional Review Board (IRB).

4.1.4 Pilot deployment. In January 2022, we carried out a pilot deployment of the tool for 3 weeks in r/pics under the complete supervision of its moderators. During this time, 88 banned users were subject to the AppealMod process. 34 out of 88 users completed the process. The research team analyzed the log data from the pilot deployment to resolve any bugs in the system, and we asked moderators for any feedback. The log data that we analyzed included 1) messages sent by the users and our bots' response to these messages, and 2) any messages or actions taken by the moderators, including their final decision. Moderators provided their feedback via modmail.

4.2 Findings: Understanding Moderators' Practices and Needs

In this subsection, we report the findings from our design process. To provide our readers with a comprehensive understanding of moderators' current practices and their needs, we report the combined findings from the interviews with mods of r/pics, r/soccer, and r/politicalhumor and the design sessions with r/pics mods under relevant themes below. To help our readers contextualize the changes we made in our design after the pilot deployment, we summarize these changes in Section 5.3 after describing the design features of our system.

Reviewing users' past behavior: When reviewing a user's appeal, moderators first looked for some evidence of positive behavior from the user. For example, they would try to find out if the

user is active in other communities on Reddit and how they are behaving there. Regularly engaging in other communities was viewed favorably, whereas further instances of problematic behavior were detrimental to the user's appeal. As PIC_M3 put it, *"Are they a 4 or 5 year old account that is heavily active all over Reddit? Or is all of their history talking about Ivermectin and whether Covid vaccines are fake?"* Reviewing a user's past behavior also allowed moderators to gauge the intent behind their behavior. According to POL_M1:

> *"I would say most of the time, especially with bans, I like to know what's going on with that account a little bit. One of the reasons is sarcasm as sometimes it is very hard to tell people being sarcastic, so [I] look at their account and see if they're a normal person, or if they're just a horrible person."*

In some cases, moderators also considered the norms of other communities a user is participating in as a useful signal in evaluating their appeals. As PIC_M2 pointed out:

> *"You can do the same thing if a racist comment is what they were banned for. There are certain subreddits that are known to be more racist, and if they have a lot of comments in those subreddits, then yeah, it's one indicator at least."*

To support the moderators' review, Reddit allows them to directly access the information on a user's past behavior from their modmail interface. However, many banned users do not have such old accounts, or they may not be as active in other communities. Furthermore, Reddit allows users to delete their contribution history, making it impossible for the moderators to access that information.

Getting more information from users: In case information on a user's past behavior was absent or insufficient, moderators reported directly engaging with the user to try and find out more about them. For example, when investigating the ban appeal in Figure 2, moderators were interested in finding out if the banned user was willing to accept responsibility for their actions. More generally, moderators reported asking users to explain why they think they were banned and then observed the user's reaction to their question.

Moderators explained they were more likely to grant appeals from users who were apologetic or who put effort into explaining the circumstances that led to their behavior. Users who did not accept their mistake, blamed the moderators, or sent toxic messages to the moderators were less likely to have their appeals granted. As PIC_M2 pointed out about their interactions with banned users:

> *"They might be more polite about it, which could be built into a questionnaire workflow, or they might double down and be like, go f**k yourself, you f*****g snowflake, you know? And then that's just where the conversation ends."*

Moderators also gauged the user's awareness of their community norms, usually by asking them to read the rules and then explain whether or how they broke the rule(s). Moderators felt that users who demonstrated an awareness of the rules would be less likely to violate them again in future. As noted by SOC_M1:

> *"We usually ask direct and open ended questions like, why were you banned? Or why do you think you were banned? So you know, as I told you before, we want users to have self awareness but also have awareness about our rules."*

The overall approach was largely similar across all moderators and communities. In a few cases however, moderators took an approach that was more specific to a user's past behavior. For instance, if the user was banned for spreading misinformation, a moderator may ask the user if they were willing to remove their contributions and stop participating in communities that are known for discussing conspiracy theories.

Effort asymmetry and increased workload: When reviewing ban appeals, moderators considered it crucial to carry out this investigative work given the long-term consequences of a permanent ban. However, the investigation process significantly increased their workload. In particular, it resulted in an *effort asymmetry* between moderators and users – users could submit their appeal with minimal effort while moderators bore the burden of this investigative work. For instance, in Figure 2, the user wrote a single sentence in their appeal; in response, moderators had to review their past behavior, ask further questions, and then make a final decision based on what they found. In a few cases, this led the moderators to explicitly ask for more effort from the user. As described by PIC_M2, *"I wanted something that shows they put effort into it. So I asked them to do something like write at least 200 words in their appeal. That would may be show that they are acting in good faith"*.

Exposure to toxic messages: Moderators also reported that doing their investigation increased the risk of being exposed to toxic messages from users. As POL_M1 described:

> *"They [banned users] are kind of like pseudo scientific and smart, and also superior sounding to everybody else. They convince a lot of people in their initial comments but then if you talk to them, at any level, they get super vitriolic and super angry and just kind of offensive to everybody, no matter what."*

Furthermore, users who were angry at moderators' decisions or those who were banned for prior toxic behavior tended to send more toxic messages in their appeal. While moderators temporarily muted such users to prevent them from sending new messages, the action was taken after that fact, i.e., once they had already seen the toxic message. As PIC_M2 described, *"when you check what they were banned for, and it's just them dropping the N word everywhere, them saying terrible things and being a clear troll, it's like they're not worth my time...[They were] muted for a month."*

Interacting with banned users: Despite the added risk of being exposed to toxic messages, moderators believed it was important to directly communicate with a banned user. Sometimes, interacting with users revealed a mistake in the moderators' earlier decision. For example, moderators could have mistakenly banned a user if they were not aware of the full context behind their behavior or, as PIC_M3 described, *"simply did not get the joke"*. As POL_M1 put it, moderators believed it was crucial to have *"that human touch...it is just easier for people to correct the mistakes of people"*.

Agency over final decision: Moderators also wanted to retain their control over each and every decision. Moderators were concerned that automated decisions will result in many false positives. For instance, an automated system may incorrectly classify a genuine appeal as toxic and reject it. As PIC_M2 pointed out:

> *"You will get into false positives where they [user] might go, what's wrong with saying fa***t. We don't want to necessarily nuke that opportunity with that person because they're asking a question about that word. And we don't want the bot to ignore such appeals."*

Moderators were also concerned that automated decisions will encourage adversarial behavior from the users. For example, POL_M1 commented:

> *"The weakness of the [automated] system I would say is once you do it enough, other people kind of figure out what's going on. And then you get that kind of planned response, once there's a way to know the answer, then the value of the investigation is diminished"*.

Adding notes and archiving appeals: To manage their workload and coordinate more effectively, moderators regularly added private notes to a user's appeal as a way to hold internal discussions or request another moderator for a review. These notes are not visible to the user and

allow moderators to keep all information relevant to a user's appeal in one place. As POL_M1 noted,

> *"A majority of the bans I give out are three strikes and you're out kind of bans. We use that toolbox program to make notes on people's accounts. Either my co-mods or I would put a note on an account, saying, 'Oh, this is a repeat offender of this kind of behavior' [...] so if it's an old account, and they have no notes in their account, that's a good sign, that means they've been a consistently good user."*

Upon completing their review of an appeal, moderators archive the appeal to declutter their modmail inbox and focus on active conversations with other users.

Summarizing Needs: We identified several moderator needs to support their handling of user appeals. First and foremost, moderators wanted to reduce their own workload of reviewing ban appeals by addressing the effort asymmetry between them and the banned users (N1). To support their review of a user's appeal, moderators wanted to observe the user's reflection on their behavior that led to their ban (N2) and find out about the users' awareness of their community norms (N3). User's responses to these questions directly influenced their final decision. They also wanted more protection from direct toxic messages sent by angry users (N4). Finally, they wanted an opportunity to directly engage with the appealing user before making their final decision (N5).

5 APPEALMOD

Based on what we learned about moderators' needs from the design process, we built AppealMod, a system that helps moderators process appeals from banned users. AppealMod's main goal is to reduce moderators' workload while maintaining their control over final moderation decisions. The system focuses on one type of moderation decision – whether to maintain user bans – though the framework should extend to other moderation decisions as well.

AppealMod enforces an information collection process that users must complete before their appeals are sent to human moderators. The process requires that users answer questions about their past conduct and their understanding of the community's rules. This process is automated via a bot to reduce human moderators' workload. Completing the process requires that users interact with the bot and put in more effort (toward answering the questions) than the existing processing of submitting their appeal. The process, and the extra effort it requires, is designed to discourage some users from continuing their appeals, especially when the appeals are insincere. Appeals from users who abandon the process are not forwarded to moderators, meaning moderators never see incomplete appeals from users assigned to AppealMod. In the remainder of this section, we outline AppealMod's features and explain how they connect with the moderators' needs described in Section 4.2, provide implementation details, and summarize the changes we made after the pilot deployment.

5.1 Design Features

5.1.1 Automated bot. To address the effort asymmetry between moderators and users (N1), we deploy an automated bot that guides users to complete the AppealMod process. The bot is necessary because Reddit's current appeals process allows banned users to send messages directly to the moderators' inbox. Moderators wanted a way to keep banned users' messages out of their queue until the users had completed the process, and only with automation could that be accomplished. The bot responds to any and all messages from banned users until they complete the process. In line with how Reddit classifies ban appeals, the first message from a banned user is automatically considered their ban appeal. The bot ignores any messages from users who are not banned from the community.

In response to a banned user's appeal, the bot asks them to complete the AppealMod process to have their appeals reviewed by human moderators (see Figure 3-A). As part of the AppealMod process, the user has to answer questions (see Section 5.1.3 and 5.1.4) hosted on a webform. The webform can be easily updated and customized by the moderators without requiring any technical reconfiguration of the bot. If the user sends any subsequent messages before completing the process, the bot reminds them to complete the AppealMod process before they can engage with human moderators (Figure 3-D). Once the user completes the AppealMod process, they are informed that the appeal is handed over to human moderators for their review (Figure 3-C).

5.1.2 Hidden appeals. While a user completes the AppealMod process (or simply ignores it), their appeals and any subsequent messages are archived and remain hidden from human moderators (Figure 3-B). Archiving appeals is crucial for reducing the effort asymmetry (N1) so moderators do not have to put effort into reviewing incomplete user appeals. It also declutters their inbox and allows them to focus on currently active and important conversations with other users and moderators. Moderators regularly archived conversations that were no longer active as Reddit does not allow them to delete any conversations.

The default setting of hiding all messages from banned users also offers protection from toxic messages sent by angry banned users (N4). We expected these messages to remain hidden from human moderators, as users who intended to harass moderators by sending toxic messages were likely not willing to complete the AppealMod process.

5.1.3 Behavior reflection. To allow moderators to gauge a banned user's reflection on their behavior (N2), AppealMod asks two open-ended questions. First, the system asks them to describe their behavior and the circumstances that made them act that way. Then, it asks users to reflect on whether or how they might behave differently in the future. We designed these questions to be open-ended, as writing open-ended responses requires more effort and is a costly signal to replicate, compared to providing close-ended responses. For instance, it would be more difficult to forge regret in an open-text field compared to simply selecting the option that they regretted their behavior. We provide the questions verbatim in Table 6 in Appendix.

5.1.4 Awareness of Community Norms. To gauge a banned user's awareness of community norms (N3), AppealMod first asks them to describe, in their own words, the rule that led to their ban. By asking users to describe the rule they broke, the task nudges users to read and understand the rule that led to their ban—so that they can draft an appropriate appeal.

We also added a second task to evaluate the user's general awareness of the community norms. As part of this comment labeling task, the user receives a set of five comments and must select comment(s) that they think should be allowed in the community. This second task evaluates the user's general awareness of community norms to gauge whether they are likely to break additional rules if their appeals are granted and they are given a second chance. We designed this as a labeling task so users do not have to read or memorize all the written rules but rather demonstrate a practical understanding of the community norms. All the comments were vetted by the moderators who participated in design sessions and they picked two of them as permissible. The set of comments is provided in Table 7 in Appendix.

5.1.5 Moderator Handover. Once a banned user completes the AppealMod process, their appeal is unarchived and appears at the top of moderators' inbox. This restores the direct communication channel between banned users and human moderators (N5). All the user messages are now visible to human moderators, and any subsequent messages in this conversation are ignored by our bot. The bot also adds a summary of the user's responses to this communication channel (see Figure 3-E and Figure 4-E). This additional information is visible only to moderators and is readily accessible

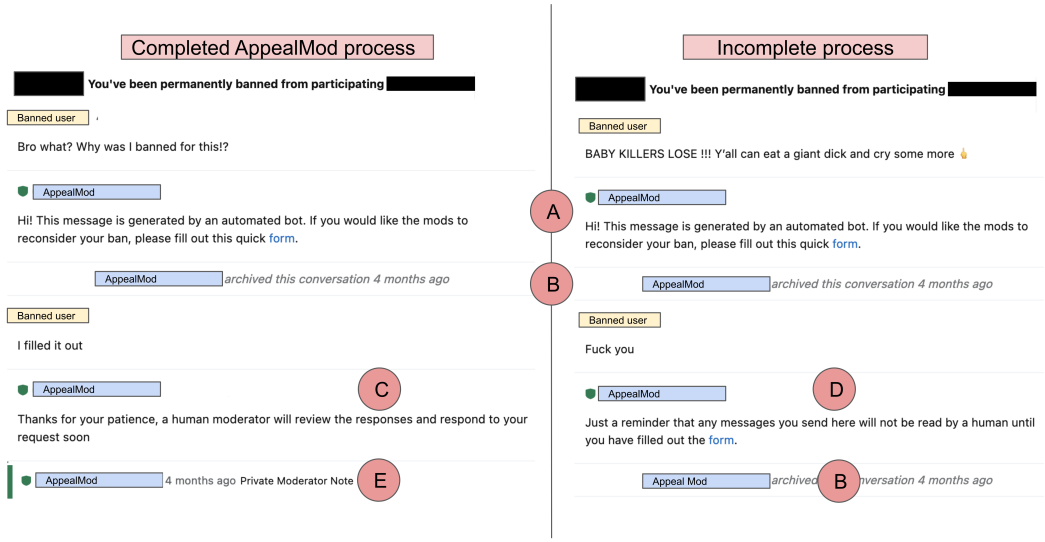

Fig. 3. Two examples demonstrating a complete (left) and incomplete (right) AppealMod process. A) Our bot responds to the user's appeal and shares a link to the form containing our questions. B) Bot archives the conversation to hide it from human moderators. C) Once the user completes the AppealMod process, the bot hands over their appeal to human moderators. D) If the user sends any more messages before completing the process, they are reminded to complete the process. E) For users who complete the process, a private note summarizing their responses is also shared with human moderators.

when they review the user's appeal or engage with them further. The final decision on a user's appeal is always made by a human moderator (see Figure 4-F).

5.2 Implementation Details

The AppealMod bot is implemented in Python. The bot's responses are configured in a fixed dialogue flow, shown in Figure 5. The dialogue flow is triggered in response to any new modmail message from a user. We used the Reddit API (with moderator-level privileges) so that the system continuously listens to any incoming messages from the users and performs other functions such as (un)archiving conversations, responding to users, and sharing private notes with moderators. During the experiment period, the bot ignored messages from banned users assigned to the control group.

We stored information about a banned user's current state of appeal and their experimental group in a MongoDB database. The questionnaire was hosted via a Qualtrics webform. We used the Qualtrics API to dynamically control access to the webform, so that it is only accessible to banned users in the treatment group. A user could submit the form only once and could not change their responses afterward. We also used the Qualtrics API to periodically check whether a banned user had completed the AppealMod process, and if so, retrieved their responses. This way, we did not have to rely on any explicit confirmation from the user about completing their process.

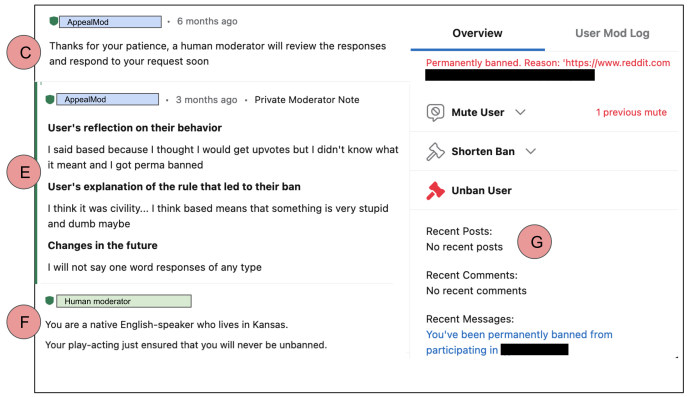

Fig. 4. An example demonstrating how a user's responses are formatted and shared with human moderators (E). The information is readily available to moderators along with the user's past history already provided by Reddit (G). Moderators make the final decision by directly interacting with the user (F).

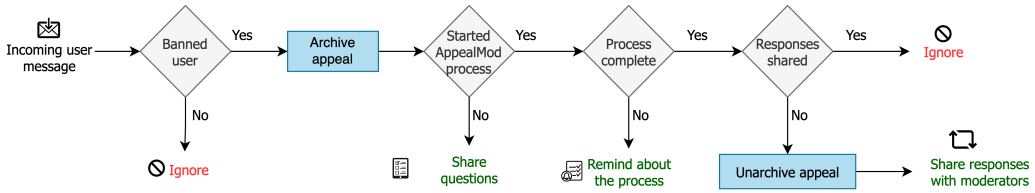

Fig. 5. AppealMod Bot Dialogue Flow

5.3 Changes After the Pilot Deployment

The pilot deployment of the system in r/pics (described in Section 4.1.4) also led to some crucial changes in the design of our system. For instance, we found in a small number of cases (4 out of 88) the ban reason was missing from the initial message sent to the user. Moderators attributed this to a bug in one of the moderation tools they use to ban users from their phone app instead of the website. Since the AppealMod process is designed to collect additional information concerning a user's ban, it would be meaningless to put users through the AppealMod process if they are unaware of their ban reason. Therefore, we updated our design to exclude those users from the AppealMod process so that their appeals are directly visible to human moderators who can notify them of the reason for being banned.

Even when the ban reason was present in the initial message sent to the user, 11 out of 34 users who completed the AppealMod process responded to our questions saying they were unaware of their ban reason. In order to nudge users to read the ban message carefully, the first question in our webform asked them to copy and paste the ban reason provided in the initial ban message.

We also made some changes based on moderators' feedback following the pilot deployment. Moderators mostly provided suggestions on improving the presentation of user responses collected during the AppealMod process. To incorporate their feedback, we used Markdown[4] style formatting

[4]https://daringfireball.net/projects/markdown/

to present the users' responses within the modmail interface (see Figure 4-E). This improved the readability of the responses while making sure they were easily incorporated into the moderators' existing workflow.

6 EVALUATION

We evaluated AppealMod by conducting a field experiment in r/pics. The experiment lasted 4 months, from April - August 2022. We and the moderators agreed it was necessary to deploy our system in a real setting to gauge the extent of effort users are willing to expend toward appealing against moderators' actions. The field experiment's main goals were closely tied to the moderators' needs described in Section 4.2. In line with the moderators' needs, we wanted to examine whether the system is effective at reducing moderator workload (N1) and protecting them from toxic appeals (N4). Additionally, we wanted to examine whether our system would adversely impact any appeal outcomes, such as the number of appeals granted by the moderators. From these goals, we identified six specific hypotheses tested in our experiment. We articulate those hypotheses below and then describe the experiment's design and outline our analysis plan.

6.1 Hypotheses

6.1.1 Impact on Moderator Workload. We measure "moderator workload" as the number of appeals moderators receive and subsequently review. We expect that AppealMod will reduce the number of appeals that are visible to moderators because some users will be discouraged from putting additional effort and abandon their appeal rather than complete the process. Therefore, our first hypothesis is:

H1a: AppealMod will reduce the number of appeals that are visible to human moderators.

Since fewer appeals are visible to the moderators, the raw number of appeals they respond to will also reduce. Therefore, we predict hypotheses 1b:

H1b: AppealMod will reduce the number of appeals that human moderators respond to.

6.1.2 Impact on Appeal Toxicity. Users often criticize moderators for taking actions against them. As we found in our formative study, some users, instead of submitting sincere appeals, may send toxic messages in their appeals. We expect that users whose main intention is to attack or harass human moderators will be discouraged by the AppealMod process. As noted earlier, when users abandon the process, their messages remain hidden from human moderators. Therefore, our next hypothesis is:

H2a: AppealMod will reduce the proportion of appeals that are visible to human moderators that contain a toxic message.

In contrast, users who complete the AppealMod process get the opportunity to directly engage with moderators. The additional effort AppealMod requires may frustrate these users and lead to more toxicity in their follow-up conversations with moderators. It is also possible that some users added toxic content to their form responses which are shared with the moderators.[5] Ideally, we want to ensure that AppealMod does not subject human moderators to additional toxicity from appealing users. To confirm, we propose the following hypothesis:

H2b: AppealMod will not increase toxicity in users' subsequent engagement with human moderators.

[5]Due to Perspective API's limitation, we did not automatically filter out user's form responses; see Section 6.3.1 for more details.

6.1.3 Impact on Appeal Outcomes. Our hypotheses so far have focused on evaluating the effectiveness of our system, AppealMod, at fulfilling moderators' needs, i.e., reducing their workload and protecting them from subsequent toxicity. It is also important to ensure that AppealMod doesn't adversely affect the desired outcomes on users' appeals. One measure to track is the number of appeals granted by moderators. Ideally, AppealMod will not impact the raw number of successful appeals; instead, it will reduce only the number of unsuccessful appeals that have to be reviewed by the moderators.

Specifically, we expect that users who have a strong case for their appeal are more likely to self-select themselves into completing the process. In that case, a reduction in the number of appeals that are visible to human moderators would not lead to a decrease in the number of appeals that are granted. For our next hypothesis, we analyze the presence of self-selection among appealing users, and make the following claim:

H3a: AppealMod will not reduce the number of appeals granted by human moderators.

Another important outcome concerns with the human moderators' response rate on appeals and their follow up engagement with appealing users. Recent work on content moderation found that users of some social media platforms described the appeals process as "speaking into a void", due to its lack of human interaction [43]. Prior work also found that users considered communication, especially direct communication with human moderators, as one of the top three avenues for improving their abilities to contest content moderation decisions [56].

While AppealMod increases the cost of directly interacting with human moderators, it does not prohibit it. Users who complete the AppealMod process have a chance to interact with human moderators. Ideally, we want to ensure that AppealMod does not make it less likely that appealing users would receive a response from human moderators. In fact, we might even see an increase in moderators' responses given the extra effort initially put in by appealing users. We propose a conservative hypothesis to ensure appealing users still receive a human response when using AppealMod:

H3b: AppealMod will not reduce human moderators' engagement on appeals.

6.2 Experiment Design

We designed the experiment process in close collaboration with moderators of r/pics. Subreddit users were not made aware of the experiment. We took the decision as moderators believed that we will not be able to capture the users' true behavior if they were informed of the experiment. Furthermore, our system presented no more than minimal risk to the subjects. The experiment design, including a waiver of informed consent requirements, was approved by our Institutional Review Board. We describe our ethical considerations in more detail below.

During the experiment period, AppealMod managed appeals (and any subsequent messages) from banned users who were requesting reinstatement. AppealMod considered a banned user's first message to moderators via modmail as an appeal[6] and randomly assigned the appealing user to either the *control* or *treatment* group. We maintained the user assignment throughout the experiment so that users who were banned more than once during the period would remain in the same group. However, none of the users whose appeals were granted were banned a second time.

Users assigned to the *control* group experienced normal interactions as per Reddit's current ban appeals process. Their messages were immediately visible to moderators in modmail, and the AppealMod bot did not interact with these users. Figure 2 shows an example interaction between an appealing user and a human moderator under control condition.

[6]Reddit uses the same criteria when displaying ban appeals as part of the moderators' modmail interface.

	Hypothesis	Measure	Statistical test
Moderator workload	H1a	No. of appeals that are visible to moderators	Chi-square test
	H1b	No. of appeals that moderators respond to	Chi-square test
Appeal Toxicity	H2a	Proportion of visible appeals that are toxic	Chi-square test
	H2b	User toxicity in subsequent messages to the moderators	Independent t-test
Appeal outcomes	H3a	No. of appeals granted	Chi-square test
		AppealMod completion ~ predicted probability of success	Regression Analysis*
	H3b	Ratio of appeals responded to and appeals visible	Chi-square test
		No. of messages exchanged	Independent t-test
		No. of characters exchanged	Independent t-test

Table 2. Overview of our hypotheses and their associated measures. *All statistical tests but the regression analysis compared between the control and treatment groups. Regression analysis was only applied to the treatment group.

In the *treatment* condition, users received automated messages from a bot that guided them through the AppealMod process. As we described in our system design (Section 5), the AppealMod process required users to complete a webform. Until they completed the form, any messages they sent to the moderators were automatically archived and hidden from moderators. Once users completed the form, the bot added their responses to their modmail conversation, which was unarchived and made visible to human moderators. From this point onward, users experienced normal interactions with moderators analogous to the control group.

Ethical considerations. While we were not able to ask for the consent of users in r/pics due to the study design, we asked for and obtained moderators' consent as they represent the community and are knowledgeable about users' preferences. This is an approach used by prior work that involved deploying interventions in subreddits, such as Zong and Matias [61]. The authors considered asking moderators' consent as asking for *proxy consent* of the users in the subreddit, which is often used in social experiments when researchers cannot directly ask for participants' consent due to preserving the validity of the research [20]. Instead, researchers disclose full information about the study to someone who can decide whether to give consent on the participant's behalf [20].

6.3 Analysis

Table 2 summarizes the analyses we carried out for the hypotheses described earlier. Below we explain the additional variables computed and the statistical tests used to verify the hypotheses.

6.3.1 Variables. In addition to the outcome variables that can be directly captured from the experiment (e.g., number of messages sent, number of appeals visible, and whether an appeal was granted or not), we constructed two variables from the text of user appeal and messages: toxicity and predicted probability of success.

Toxicity: We measure human moderators' exposure to toxicity in the messages sent by users, and their responses as part of the AppealMod process. We use the Perspective API[7] to assign a toxicity score to user appeals and any subsequent messages they sent. Perspective API has been used in prior work for measuring toxicity of texts [21, 37]. For any input text, Perspective API returns a score between 0 to 1, which denotes the likelihood of the message being toxic. Among the scores provided by the API,[8] we used the TOXICITY score—which is described as measuring how likely the text is a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." We experimented with thresholds of 0.5, 0.7, and 0.9 to classify appeals and

[7]https://perspectiveapi.com/

[8]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

messages as toxic or non-toxic. For our main results, we set the threshold at 0.7. Results for other thresholds are reported in the Appendix.

We recognize that Perspective API is not robust to quoted sentences. For instance, consider the following text which received a toxicity score of 0.8 (i.e., 80% likely to be toxic) – *"I said go fuck yourself because he was being disrespectful to women. Although, I accept that I should not have acted in anger."* User appeals (and follow up conversations) often include similar quotations, and Perspective API classifies those messages as toxic. However, Perspective's scores in these cases should equally impact both control and treatment groups, and their comparison should be robust to this classification.

Next, we also look for toxicity in the user responses provided during the AppealMod process. We manually analyzed these responses to only count instances of additional toxicity from users and excluded instances that had quotations to prior toxic behavior. We expected such responses from users since one of the questions asked them to reflect on their past behavior. For the manual analysis, one of the authors and one additional rater (a graduate student) independently rated a sample of 30 form responses. The rating guidelines defined toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion" (as defined by the Perspective API). The raters were particularly instructed to look for instances of increased toxicity from users given their past behavior. The two raters met to discuss their disagreements and arrive at a consensus label. Each of the remaining form responses was then rated by one of the raters.

Predicted Probability of Success: Because AppealMod requires more effort, we expected some users to abandon their appeals. However, we didn't expect users to randomly abandon their appeals. Rather, we expected users whose appeals were likely to be granted to be more likely to complete the AppealMod process. To investigate this link, we wanted to predict an appeal's initial probability of success (PPS), which would be independent of whether the AppealMod process was completed or not. We modeled this likelihood of success using the initial appeal message, as the message was sent before the user's assignment to one of the experimental groups. To construct the PPS model, we collected (1) the initial message of all appeals made by users who were banned from r/pics between April 2021 and December 2021 (i.e., 6 months before the experiment began) and (2) the final decision on their appeal. The dataset consists of 6543 appeals, 846 (12.9%) of which were granted. We use a simple Logistic Regression-based classifier. The classifier predicts the outcome of the appeal (granted or not) using unigram and bigram features from the initial appeal message. Using 5-fold cross validation, we found the model to be fairly accurate (F-score(macro)=0.83).

For all appeals under treatment during the experiment period, the output of this classifier was used as their PPS score. We then performed a logistic regression analysis to estimate if there was any relationship between PPS and whether users completed the AppealMod process. Based on our hypothesis, we expected users with higher PPS to be more likely to complete the AppealMod process.

6.3.2 Statistical analyses. We used statistical tests to compare outcomes between control and treatment groups (see Table 2 for a summary). Specifically, we use a Chi-squared test to compare frequency or count variables (e.g., number of appeals granted) and a Mann–Whitney–Wilcoxon test [12] to compare the distributions (e.g., number of messages exchanged between users and moderators per conversation). Mann-Whitney-Wilcoxon test is a non-parametric test that does not require the samples to be drawn from a normal distribution [12]. We also used regression analysis to characterize the relationship between the users' predicted probability of success (based on messages sent before they are assigned to an experiment group) and whether they completed the AppealMod process.

	Control	Treatment
Total appeals submitted	438	442
Appeals visible	438	131
Appeals responded to	263	105
Proportion of visible appeals that are toxic	13.24% (58 out of 438)	3.8% (5 out of 131)

Table 3. Comparison between appeals visible to human moderators under control and treatment

6.4 Post-Experiment Session with Moderation Team

Following the experiment, we had a debriefing session with one of the moderators of r/pics who participated on behalf of the entire team. In addition to quantitatively studying AppealMod's impact via the main study—a field experiment, we wanted to find out about the moderation team's perceptions of AppealMod. The first author conducted the session via Zoom which lasted for approximately 1 hour 30 minutes. We began by asking the moderator to describe their experience of addressing appeals via AppealMod. As the session progressed, we picked a random set of appeals from the treatment group and asked them to walk us through their process for addressing those appeals. The complete protocol that guided the post-experiment session is provided in Appendix A.6.

7 RESULTS

We conducted the experiment for four months, from April to August 2022. During this period, 880 users appealed their bans. Users were roughly equally divided into control (438) and treatment groups (442). Out of the 442 users assigned to the treatment group, 131 completed the AppealMod process. The median time for completing the process (after clicking on the form link) was 4 minutes and 50 seconds. 90 out of the 880 appeals were granted by moderators. In the following subsections, we describe the results specific to our individual hypotheses.

7.1 Moderator Workload was Lower

H1a: AppealMod will reduce the number of appeals that are visible to human moderators.
H1b: AppealMod will reduce the number of appeals that human moderators respond to.

Table 3 shows a comparison between number of appeals that were visible to human moderators and that they responded to under control and treatment. In the control condition, all appeals submitted (n = 438) were immediately visible to moderators. In the treatment condition, only 30% of appeals (n = 131) were visible to moderators. Users abandoned the other 70% of appeals (n = 311) in the treatment condition, and those incomplete appeals were not visible to moderators. Therefore, AppealMod reduced the human moderators' review queue by 70%. Human moderators did not respond to all appeals in their queue under either condition. In the control condition, moderators responded to 263 appeals, and in the treatment condition, they responded to 105.

Chi-squared tests confirmed that fewer appeals were visible to the moderators in the treatment condition ($\chi^2(1, N = 880) = 437.55, p = 5.39\text{e}{-105}$) and that they responded to fewer appeals in the treatment condition ($\chi^2(1, N = 880) = 117.59, p = 2.13\text{e}{-27}$).

7.2 Fewer Toxic Appeals were Visible to Moderators

H2a: AppealMod will reduce the proportion of appeals that are visible to human moderators that contain a toxic message.

Roughly 13% (n = 115) of all initial user appeals were classified as toxic; with nearly equal numbers of toxic appeals in control (58) and treatment (57). Under control, all toxic appeals were directly

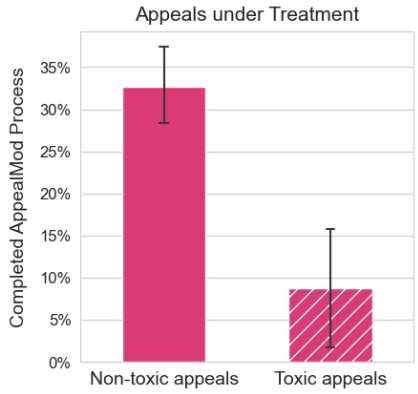

Fig. 6. Comparing the proportion of toxic and non-toxic appeals with AppealMod process completed. The 95% confidence intervals for the two groups are non-overlapping, which indicates that these differences are large enough to be reliable.

visible to human moderators. However, under treatment, only 3.8% of the appeals visible to human moderators were toxic (see Table 3); the other toxic appeals were hidden because users abandoned those appeals. A Chi-squared test verified that the difference in the proportion of appeals visible to human moderators that were toxic under control and treatment was statistically significant, $\chi^2(1, N = 569) = 8.17, p = 0.004$. We set a threshold of 0.7 on the Perspective API output to classify appeals as toxic or not. Varying the threshold to 0.5 or 0.9 did not qualitatively change the results (see Table 8 in Appendix).

These results indicate that, under treatment, most users who authored toxic appeals did not complete the AppealMod process, and therefore, these toxic appeals were hidden from human moderators. Figure 6 shows that while 32.7% of non-toxic appeals were completed, the completion rate dropped to 8.7% for toxic appeals. Furthermore, the 95% confidence intervals for the two groups are non-overlapping, indicating that these differences are large enough to be reliable.

H2b: AppealMod will not increase toxicity in users' subsequent engagement with human moderators.

For this hypothesis, we are interested in (1) comparing toxic messages sent by users under control and treatment when directly interacting with moderators, and (2) toxic content provided by treatment users in their form responses. For (1) we are only interested in appeals that had at least one back-and-forth interaction between the user and the moderators. In these cases, both the user and the moderator(s) sent at least one message each after the user's initial appeal. There were 162 conversations under control and 84 under treatment. Under control, users sent a total of 326 messages out of which 12 were classified as toxic. Under treatment, users sent a total of 207 messages out of which 10 were classified as toxic. A Chi-squared test verified that the difference in the proportion of messages sent by users that were toxic under control and treatment was not statistically significant, $\chi^2(1, N = 533) = 1.57, p = 0.209$.

Next, we look at the number of form responses from treatment users that were toxic. As we noted in Section 6.3.1, we analyzed the form responses manually to count instances of additional toxicity from users and exclude instances that had quotations to prior toxic behavior. Out of 131 form responses, 3 were flagged as toxic. However, the rate of toxicity in form responses (2.3%) is still lower than the base rate of toxicity among initial appeal messages (13%).

	Control	Treatment
Appeals visible	438	131
Moderator response rate on visible appeals	60.04%	80.15%
Median number of messages exchanged per appeal	5	5
Median number of messages sent by human moderators per appeal	2	2
Median number of characters contributed by human moderators per appeal	210	208

Table 4. Human moderators' engagement on appeals that were visible under control and treatment

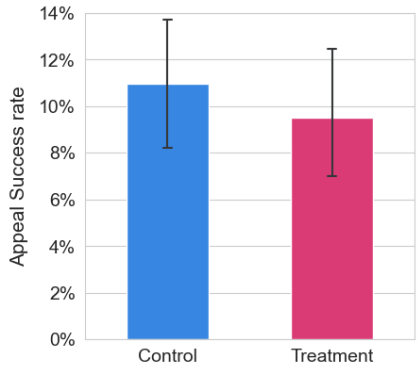

Fig. 7. Comparing success rate of appeals under control and treatment. The confidence intervals largely overlap, and a Chi-squared test further showed that the difference between the two are not statistically significant ($\chi^2(1, N = 880) = 0.362, p = 0.547$).

7.3 Moderators Granted Similar Number of Appeals and had a Higher Response Rate

	Dep Variable: AppealMod process complete
Intercept	-1.051***
	(0.117)
PPS	3.408***
	(0.939)
AIC	525.632
Observations	442
Residual Std. Error	1.000(df = 440)
F Statistic	(df = 1.0; 440.0)
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 5. Regression analysis on the relationship between predicted probability of success (PPS) and whether the AppealMod process was completed

H3a: AppealMod will not reduce the number of appeals granted by human moderators.
During our experiment, the overall success rate of appeals was roughly 10% (90 out of 880). Figure 7 shows the success rate of appeals and their 95% confidence intervals under control and treatment. While 48 appeals were granted under control and 42 were granted under treatment, the

large overlap between the confidence intervals suggests that success rates with and without the AppealMod process are similar. A Chi-squared test further verified that differences in these two results are not statistically significant ($\chi^2(1, N = 880) = 0.362, p = 0.547$).

To further investigate the selection effect among users under treatment, we used a logistic regression model to analyze the relationship between users' predicted probability of success (PPS) and whether they completed the AppealMod process. As noted in Section 6.3, a user's PPS can be reliably computed from the text of their initial appeal message (posted before they are asked to complete the AppealMod process). Table 5 summarizes the regression model and shows a statistically significant $p < 0.001$ relationship between a user's PPS and whether they completed the AppealMod process. Computing the odds ratio, we found that a 10% increase in the users' PPS increased their odds of completing the AppealMod process by 40%.

These results indicate that the AppealMod process did not uniformly discourage users from continuing their appeals. Instead, users whom we predicted to have a high probability of success, based on their initial message, were more likely to complete the AppealMod process. Users with a low likelihood of success were less likely to complete the process.

H3b: AppealMod will not reduce human moderators' engagement on appeals.

Table 4 shows a comparison between the various outcomes on appeals that are visible to human moderators under control and treatment. First, we found that moderators' response rate (a desired outcome) on visible appeals under treatment (80.15%) was higher than their response rate on visible appeals under control (60.04%). A Chi-square test verified that the difference was statistically significant, $\chi^2(1, N = 569) = 16.97, p = 3.78\text{e}{-5}$.

Next, we analyzed follow up conversations on appeals that moderators responded to. Using a Mann–Whitney–Wilcoxon test, we found no statistically significant difference between control and treatment groups in terms of the total number of messages exchanged on a given appeal, the number of messages moderators sent, or the length of moderators' messages. Under both control and treatment, a median of 5 messages were exchanged between the appealing user and moderators (Mann–Whitney U = 6701.5, n=246, p=0.422, two-tailed), and moderators sent a median of 2 messages per conversation (Mann–Whitney U = 6108.5, n=246, p=0.085, two-tailed). In terms of characters, moderators' contributed a median of 210 characters per conversation in control and 208 characters per conversation in treatment (Mann–Whitney U = 6732.5, n=246, p=0.447, two-tailed).

7.4 Notes from the Post-Experiment Session with Moderators

We briefly note the findings from the post-experiment session with a moderator representing the team. Overall, the team was satisfied with AppealMod. The moderators confirmed that with the additional information provided during the AppealMod process, they found it easier to respond to users' appeals. Specifically, moderators reported that users' responses gathered via the AppealMod process provided valuable information and context that is hard to get from the platform's existing signals. For example, the moderator said:

> *"And it gives us the context because what we have to work with is very little. If they're talking a lot about their personal life in their post history, great, but the majority of Reddit accounts are not posting personal information."*

In some cases, the users' responses also helped in gauging their awareness of the community norms. As the moderator put it:

> *"So obviously, if they went back to doing exactly the same thing, they would get banned again. But the form tells me that they are aware of what our guidelines are, and they're stating that they're willing to abide by them. And that, to me, is enough for someone to be unbanned."*

According to the moderators, having an automated bot also streamlined the appeals process, making conversations with the user faster and more efficient. For instance, one moderator said:

> *"When we didn't have the bot and in cases when it doesn't fire, we do all the investigation that we have to do. And when we have a conversation, we have to wait for the user to respond, and go back and forth several times. So that process often slows it down for the user – if they miss a message and drop off, obviously we wouldn't press it. So having that [AppealMod] form both enables and speeds along the process."*

While the moderators found users' responses collected during the AppealMod process to be helpful, these responses alone did not determine the final outcome. Moderators followed a nuanced decision making process that, in addition to utilizing the AppealMod process, also involved additional factors, such as, the original reason behind a ban, who issued the ban, how old is the ban, and how the user behaving in other communities after their ban, to name a few. A moderator provided a relevant example:

> *"Another contributing factor of why I banned them is when you look at their profile, they have negative comment karma, which is very difficult to do just because of the way that Reddit adds karma, which says to me that they are explicitly trolling and that's probably the goal of the account; they are not engaging in good faith."*

.

The moderation team also requested one change in the AppealMod design. Currently, the bot is configured to update moderators via a private note when the Qualtrics API is down and users' responses cannot be retrieved. This happened twice during the experiment period. The bot was able to retrieve responses once the API became functional after a few hours. Moderators gave feedback that users should also receive a message from the bot informing them of any technical issues and requesting them to be patient while the issue gets resolved.

The moderation team requested that we configure AppealMod for addressing all user ban appeals moving forward, removing the control condition. We have not yet done so, in part because we do not yet have a maintenance and monitoring plan that would allow us to provide the service reliably enough. With the current setup, where AppealMod is turned on for only some of the appeals, moderators are able to handle downtime of the system without disruption to their normal practices.

8 LIMITATIONS

While the novel system we developed and experimented with shows promise for reducing moderator workload without negatively impacting the subreddit community, our work has several limitations worth noting.

First, our system might lead to other unintended negative outcomes that we have not measured. We note that users who post toxic comments and users with insincere appeals were more likely to be discouraged by the AppealMod process. However, some of the users discouraged from completing the AppealMod process did not display any toxic behavior. It is possible they indulged in other kinds of problematic behavior, for example, spamming bots will not be able to complete the AppealMod process, as it requires human intervention to manually fill out the form. However, since we did not collect any additional information from the users (such as their demographic information), we cannot ascertain whether any other factors played a role in users abandoning the process. For instance, we designed the AppealMod questionnaire in English as contributions to r/pics are limited to English language. It is possible that users whose first language is not English found it more challenging to respond to the questionnaire and were therefore more likely to abandon their appeal. More broadly, prior research has found that content moderation interventions can disproportionately harm marginalized users [16]. We encourage future work to explore whether

our specific design of AppealMod and the more generic approach of inducing friction in moderation processes has differential impact on different groups of users, and whether it causes further harm to marginalized users.

Second, we found that AppealMod reduced the number of appeals reviewed by moderators by 70%; however, we evaluated AppealMod in only one community on Reddit. Furthermore, the AppealMod design is based on interactions with a handful of moderators. While the experience of these moderators is vast and varied (all moderators have years of experience and most have moderated multiple subreddits), we cannot ascertain how the AppealMod design will operate in other communities, and whether its impact will be different for communities with different benefits, rules, and moderators. For instance, r/pics is the one of the largest communities on Reddit, the most popular place for sharing pictures on Reddit. Given its popularity, users who are banned are likely motivated to get back into the community. At the same time, when users are banned from a public Reddit community, they can still see the community's content, but they cannot post anything. This is unlike platforms like Facebook where users who are banned from a group can no longer see the group or any of its content.[9]

One implication of a complete shutout from the community is that more users would be willing to complete the process compared to Reddit users. Therefore, while Reddit moderators only reviewed 30% of the appeals submitted under AppealMod, Facebook moderators, for instance, might have to review a higher fraction of appeals submitted under AppealMod. Given the higher decision stake for users on platforms where banning leads to a complete disconnect from the community, another implication is that users might be willing to put more effort in their appeals process. So the AppealMod process for such platforms can be potentially expanded to collect more details from the users that would support the moderators' review. Overall, the benefits of the community and the implications of its banning functions will impact how willing users are to bear the additional costs associated with the AppealMod process. Future work should explore how interventions on users' effort impact moderators' workload in different contexts.

Third, we used the Perspective API[10] to automatically classify user appeals as toxic or not. As we briefly noted in Section 6.3, Perspective API occassionaly makes erroneous classifications. These errors will equally impact both the control and treatment groups; our comparison between treatment and control is robust to these errors. However, we caution our readers from generalizing the extent to which AppealMod can reduce the proportion of appeals that are toxic or the overall extent of toxicity in user appeals. While we find that approximately 12% of all user appeals are toxic, and only 5% of appeals visible under treatment are toxic, the actual number of toxic appeals may be higher or lower depending on the types of Perspective's errors and their distribution.

9 DISCUSSION

We worked directly with moderators of a large subreddit to design an appeal process that reduced moderators' workloads, honored users' and moderators' need for direct interaction, and minimized negative impacts to the community. We deployed the AppealMod system for 4 months and found that the process effectively met these goals. Moderators needed to review only 30% of appeals; users were able to directly discuss their behavior with moderators who made the final decision; users who abandoned the process were either insincere in their appeals or toxic in their comments. In this section, we discuss the impact of AppealMod's effort-based moderation technique on different groups of users. Then, we address the importance and feasibility of meeting both users

[9]https://www.facebook.com/help/211909018842184
[10]https://perspectiveapi.com/

and moderators' needs when improving moderation systems' contestability, including relevant future work.

9.1 Effort Asymmetry and Friction in Content Moderation

One of the key findings from our design process with moderators was the discovery of *effort asymmetry* between moderators and banned users during the ban appeals process. Moderators spent considerable effort reviewing user appeals of their bans, but users could submit an appeal with minimal effort. To address the asymmetry, AppealMod requires users to provide additional details before their appeals are reviewed by human moderators. Appeals from users who abandon the AppealMod process remain hidden from human moderators. By asking users to put more effort into their appeal, we effectively induced *friction* in appealing, which discouraged many users from completing the process. Our results show that moderators reviewed only 30% of appeals under the AppealMod process. Overall, our system demonstrates that carefully designed friction can reduce moderators' workload.

An important concern with inducing friction in a process is the unintended negative effects on participation. For instance, if the increased effort uniformly discouraged users from completing the process, moderators would eventually grant fewer appeals. Fewer successful appeals likely prevent many deserving users from rejoining the community. However, the AppealMod process induced a selection effect among users. We found users who abandoned the process were either insincere in their appeals or toxic in their comments. This selection effect reduced the number of appeals reviewed by moderators by 70% while moderators still granted roughly the same number of appeals. We hypothesize that the increased effort had a differential impact on users. The AppealMod process likely aided users with a sincere appeal by giving them a space to reflect on their behavior and demonstrate awareness of community norms. Other users, for instance those who were angry at the moderators or did not care about community norms, likely found it more challenging to complete the process and abandoned their effort altogether.

The success of inducing friction in a moderation process will largely depend on the type and degree of effort required. As is the case with AppealMod process, the increased effort must be relevant and useful to deserving users and discouraging to problematic users. Furthermore, one must ask for the right amount of effort. The AppealMod process included three-open ended questions and one comment labeling task; the median time for completing the process was under five minutes. Asking users to put in more effort will result in more users abandoning the process and exacerbate any negative outcomes. Asking for too little will invite everyone to complete the process and undermine the effectiveness of this approach. Our design process with moderators was crucial in coming up the right type and amount of effort when designing the system.

9.2 Balancing User and Moderator Needs For Improving the Process of Contesting Moderation Decisions

In this paper, we presented a new system for contesting content moderation decisions, in particular, the decision of banning users from a community. While we designed our system from the moderators' point of view, some elements of our design support users' contestability needs as well. For instance, both users and moderators desire to directly interact with each other and maintain human agency over the final decision. In fact, social media users consider communication, especially direct communication with human moderators, as one of the top three avenues for improving their abilities to contest moderation decisions [56]. While AppealMod increases the effort required to directly engage with a human moderator, it does not prohibit it. Our findings show that moderators' response rate was higher on completed appeals under the AppealMod process (80.15%), compared to their response rate on appeals under control condition (60.04%). The initial effort users put in to

complete the AppealMod process likely prompted human moderators to engage with users more frequently.

The AppealMod process can also guide users in completing their appeals by laying out factors that moderators considered in their decision making. Providing scaffolding can address the cold start problem as many users do not know how to write their appeals [55]. Doing so can also make the appeals process more transparent and reduce perceived inconsistencies in moderation decisions [43]. Research on procedural fairness finds that transparency along with the feeling that one's voice has been heard are important components of users' fairness judgements [33]. Therefore, completing the AppealMod process can improve a user's perceived fairness and overall satisfaction with the system, especially as it leads to more desirable outcomes for the users.

Finally, researchers have argued for designing the appeals process to improve users' learning and make them more likely to adhere to the community norms [43, 55]. AppealMod's current design nudges users to develop a better understanding of the community norms by asking them to differentiate between behaviors permitted and prohibited in the community. Future work could explore extending AppealMod's bot to be an interactive chatbot that can answer users' questions about community norms or how to apply them to their particular situations. Such dialectical exchanges between users and decision systems are identified as an important part of supporting users' contestability needs [48].

However, these benefits for users come at costs. Some users may find it frustrating and cumbersome to interact with an automated bot [42], especially since it is the first step before any interaction with human moderators. Research also argues that adding barriers to a process can be detrimental to users' fairness perceptions [30]. While our work centered on addressing moderators' needs, we encourage future work to further explore users' perceptions about the AppealMod process, how it impacts their behavior after their appeals are decided, and compare it with Reddit's existing ban appeals process.

10 CONCLUSION

AppealMod addressed specific challenges moderators' faced in processing ban appeals. Most importantly, it reduced moderators' workload by creating friction that effectively discouraged insincere appeals. AppealMod also minimized the toxic messages moderators were exposed to and maintained moderators' control over appeal decisions. It achieved these gains for moderators while honoring users' and moderators' preferences for direct human engagement in appeals. The primary difference between the existing appeal process and the process under AppealMod is the level of effort users must put forth to submit an appeal. By introducing productive friction in the process, AppealMod increases the costs of each individual appeal while teaching users about the community's rules and its decision-making process.

The volume of content moderation that is necessary means that some automation will be required to allow moderators to keep up. However, even moderators make mistakes, and their decisions should be contestable. Our design process and experiment illustrate one strategy for co-creating a system for automating aspects of the moderation process and enabling contestibility of moderator decisions.

11 ACKNOWLEDGEMENTS

We want to express our deepest gratitude to the moderators of r/pics for their time and cooperation. Without their input, this project would not have been possible. Special thanks to Tawanna Dillahunt and Shagun Jhaver for their generous feedback that helped scope the initial design process. This material is based upon work supported by the National Science Foundation under Grant No. 099726.

12 CONTRIBUTION STATEMENTS

The authors confirm contribution to the paper as follows:
Shubham: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft (lead)
Jane: Writing - Original Draft (supporting)
Paul: Conceptualization, Methodology, Writing - Review and Editing, Supervision (supporting)
Libby: Conceptualization, Methodology, Resources, Writing - Original Draft (supporting), Supervision (lead), Funding acquisition

REFERENCES

[1] Marco Almada. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 2–11.

[2] Edgar Alvarez. 2019. Instagram will soon let you appeal post takedowns | Engadget — engadget.com. https://www.engadget.com/2019-05-07-instagram-appeals-content-review-taken-down-posts.html. [Accessed 06-Jan-2023].

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

[4] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300234

[5] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing Live Streaming Moderation Tools: An Analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies* 9, 2 (July 2019), 36–50. https://doi.org/10.4018/IJICST.2019070103

[6] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–30. https://doi.org/10.1145/3359276

[7] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[8] Alexander Cheves. 2018. The Dangerous Trend of LGBTQ+ Censorship on the Internet. https://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet. [Accessed 06-Jan-2023].

[9] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300372

[10] Cynthia Dwork and Moni Naor. 1992. Pricing via processing or combatting junk mail. In *Annual international cryptology conference*. Springer, 139–147.

[11] Albrecht Enders, Harald Hungenberg, Hans-Peter Denker, and Sebastian Mauch. 2008. The long tail of social networking.: Revenue models of social networking sites. *European Management Journal* 26, 3 (2008), 199–211.

[12] Michael P Fay and Michael A Proschan. 2010. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys* 4 (2010), 1.

[13] Sarah A. Gilbert. 2020. "I Run the World's Largest Historical Outreach Project and It's on a Cesspool of a Website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. https://doi.org/10.1145/3392822

[14] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, New Haven, UNITED STATES.

[15] James Grimmelmann. 2017. *The Virtues of Moderation.* Preprint. LawArXiv. https://doi.org/10.31228/osf.io/qwxf5

[16] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[17] Jeffrey Heer. 2019. Agency plus Automation: Designing Artificial Intelligence into Interactive Systems. *Proceedings of the National Academy of Sciences* 116, 6 (Feb. 2019), 1844–1850. https://doi.org/10.1073/pnas.1807184115

[18] Amanda Holpuch. 2015. Facebook still suspending Native Americans over "real name" policy. *The Guardian* 16 (2015). https://www.theguardian.com/technology/2015/feb/16/facebook-real-name-policy-suspends-native-americans

[19] Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory* 3 (2007), 265–289.

[20] Macartan Humphreys. 2015. Reflections on the ethics of social experimentation. *Journal of Globalization and Development* 6, 1 (2015), 87–112. https://www.degruyter.com/document/doi/10.1515/jgd-2014-0016/html

[21] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376383

[22] Reddit Inc. 2021. Transparency Report 2021 - Reddit. https://www.redditinc.com/policies/transparency-report-2021-2/. [Accessed 14-Jan-2023].

[23] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (Oct. 2019), 1–35. https://doi.org/10.1145/3338243

[24] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–21. https://doi.org/10.1145/3491102.3517505

[25] Aaron Jialun Jiang. 2020. Toward A Multi-stakeholder Perspective For Improving Online Content Moderation. (2020).

[26] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2022. A Trade-off-centered Framework of Content Moderation. *arXiv preprint arXiv:2206.03450* (2022).

[27] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2022. A Trade-off-centered Framework of Content Moderation. https://doi.org/10.1145/3534929 arXiv:2206.03450 [cs]

[28] Hannes Grassegger Julia Angwin. 2017. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms. [Accessed 06-Jan-2023].

[29] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. https://doi.org/10.7551/mitpress/8472.001.0001

[30] Gerald S Leventhal. 1976. What Should be Done Equity Theory? New Approaches to the Study of Fairness in Social Relations. *Washington DC: National Science Foundation* (1976).

[31] Karyne Levy. 2014. Facebook Apologizes for 'Real Name'Policy That Forced Drag Queens To Change Their Profiles. *Business Insider* 1 (2014). https://www.businessinsider.com/facebook-apologizes-for-real-name-policy-2014-10

[32] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 584–595. https://doi.org/10.1609/icwsm.v16i1.19317

[33] E Allan Lind and Tom R Tyler. 1988. *The social psychology of procedural justice*. Springer Science & Business Media.

[34] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. *arXiv:1804.06759 [cs]* (April 2018). arXiv:1804.06759 [cs]

[35] Claudia (Claudia Wai Yu) Lo. 2018. *When All You Have Is a Banhammer : The Social and Communicative Work of Volunteer Moderators*. Thesis. Massachusetts Institute of Technology.

[36] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[37] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174160

[38] Alice E Marwick and Ross Miller. 2014. Online harassment, defamation, and hateful speech: A primer of the legal landscape. *Fordham Center on Law and Information Policy Report* 2 (2014). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2447904

[39] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1138–1151. https://doi.org/10.1145/2858036.2858391

[40] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (April 2019), 205630511983677. https://doi.org/10.1177/2056305119836778

[41] Sharan B Merriam et al. 2002. Introduction to qualitative research. *Qualitative research in practice: Examples for discussion and analysis* 1, 1 (2002), 1–17.

[42] Charlie Mitchell. 2022. Zendesk Research: Customers Are Still Frustrated with Chatbots. https://www.cxtoday.com/speech-analytics/customers-frustrated-with-chatbots/. [Accessed 12-Jan-2023].

[43] Sarah Myers West. 2018. Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms. *New Media & Society* 20, 11 (Nov. 2018), 4366–4383. https://doi.org/10.1177/1461444818773059

[44] Gal Oestreicher-Singer and Lior Zalmanson. 2013. Content or community? A digital business strategy for content providers in the social age. *MIS quarterly* (2013), 591–616.

[45] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31.

[46] Sarah Perez. 2019. Twitter now lets users appeal violations within its app. https://techcrunch.com/2019/04/02/twitter-now-lets-users-appeal-violations-within-its-app/. [Accessed 06-Jan-2023].

[47] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

[48] Claudio Sarra. 2020. Put dialectics into the machine: protection against automatic-decision-making through a deeper understanding of contestability by design. *Global Jurist* 20, 3 (2020), 20200003.

[49] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why Do Volunteer Content Moderators Quit? Burnout, Conflict, and Harmful Behaviors. *New Media & Society* (Dec. 2022), 14614448221138529. https://doi.org/10.1177/14614448221138529

[50] Joseph Seering and Sanjay R Kairam. 2023. Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (2023), 1–18.

[51] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

[52] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[53] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[54] Paula MC Swatman, Cornelia Krueger, and Kornelia Van Der Beek. 2006. The changing digital content landscape: An evaluation of e-business model development in European online news and music. *Internet research* (2006).

[55] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 167:1–167:22. https://doi.org/10.1145/3415238

[56] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 318:1–318:28. https://doi.org/10.1145/3476059

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

[58] Ian Wren. 2018. Facebook Updates Community Standards, Expands Appeals Process — npr.org. https://www.npr.org/2018/04/24/605107093/facebook-updates-community-standards-expands-appeals-process. [Accessed 06-Jan-2023].

[59] Lucas Wright. 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. *Social Media+ Society* 8, 1 (2022), 20563051221077020.

[60] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 365–378. https://doi.org/10.1145/3379337.3415858

[61] Jonathan Zong and J Nathan Matias. 2022. Bartleby: Procedural and Substantive Ethics in the Design of Research Ethics Systems. *Social Media+ Society* 8, 1 (2022), 20563051221077021. https://journals.sagepub.com/doi/pdf/10.1177/20563051221077021

A APPENDIX

A.1 Recruitment Message

Hi, I'm a researcher at the <masked for anonymity> who is curious about exploring new ways to support moderators, potentially with new tools or other kinds of computational assistance. To make my research impactful, I want to hear from mods about what would be truly useful for them.

As somebody who no longer lives in <masked for anonymity>, the community has brought a lot of nostalgia, especially the recent set of posts about people sharing their experiences after moving from the US. Now as a researcher, I'd love to talk to you and understand more about moderating this community and any requirements that you may have. If you are interested, I can schedule a quick call to chat more about it. I do value your time (since most moderators are overworked) and would like to offer a gift card worth $20 as an appreciation for your time and effort. Please let me know if you'd be interested by replying to this message.

You can find more details about the project here: <masked for anonymity>.

A.2 Design Study Materials

A.2.1 Design session protocol.

- From our initial interviews, we found that moderators often receive modmail requests from users appealing against moderation decisions? How often is that the case?
- How many times is it because the user didn't understand the rule? How many times do the users just want to argue with you?
- What kind of responses do you want to get from the users?
- How would the user responses help you in responding to their request?
- Do you think the "good" and "bad" requests would be sufficiently different from each other?
- How often do you think users are likely to complete a process like this?
- Do you see any negative effects on the users while doing this extra work? Follow-up examples: will the users be irritated? do you expect any kind of backlash?
- What if users fake their responses? How would you deal with that?
- Can you share any past data on back-and-forth interactions between users and moderators that can help me design this flow?

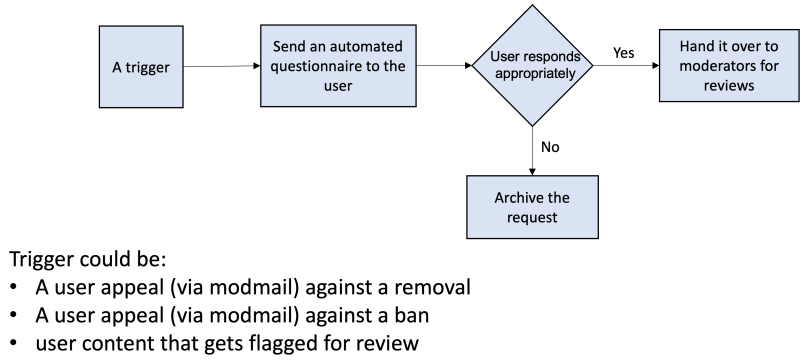

Fig. 8. An overview of our proposed design highlighting that users receive automated messages in response to their requests and that some user requests will be hidden from human moderators. The material was sent to the moderators in advance and used during the design sessions.

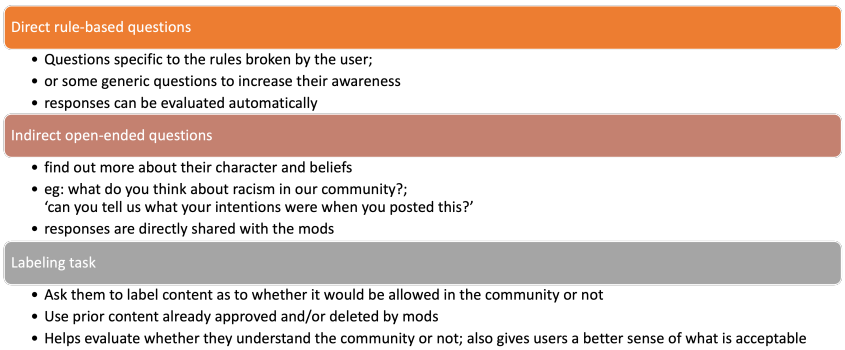

Fig. 9. A Typology of different questions that can be asked during the AppealMod process that was sent to the moderators in advance and used during the design sessions. Our final design mostly used open-ended questions along with a comment labeling task.

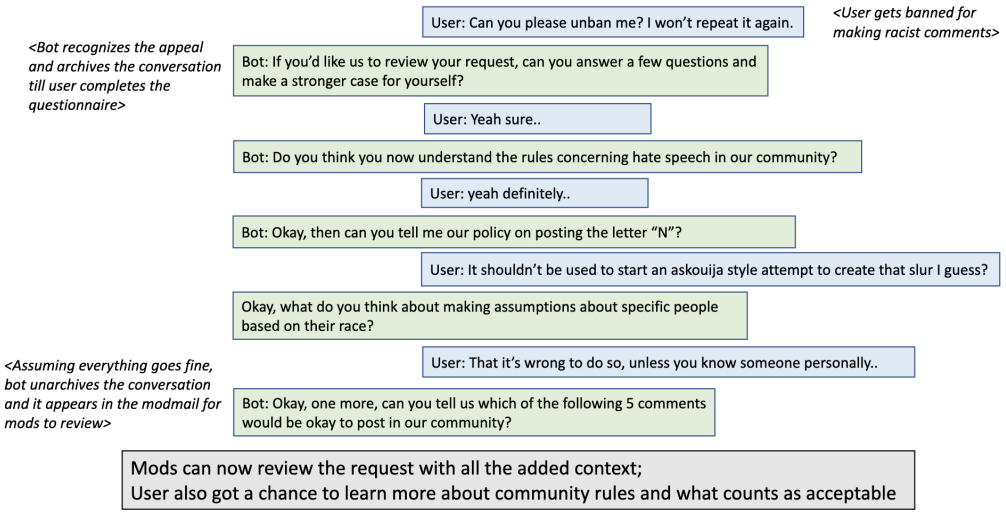

Fig. 10. A mockup conversation between our proposed bot and a banned user which was sent to the moderators in advance and used during the design sessions. Following a round of discussions on this prototype, we decided to share the questions with users in a webform instead of directly asking them in the chat. This was done to keep the overall conversation between users and moderators short and organized.

A.3 Questions asked during the AppealMod process

Design Feature	Question	Question Type
Behavior reflection	Can you describe your actions that led to the ban and the circumstances that made you act that way?	Open-ended
	What steps will you take to ensure that you don't repeat such actions in future?	Open-ended
Awareness of community norms	Can you describe the rule you broke in your own words? You may find more details about the rules here: https://new.reddit.com/r/pics/wiki/index	Open-ended
	Which of the following comment(s) do you think should be allowed in our community?	Multiple choice

Table 6. Questions asked during the AppealMod process.

A.4 Comments used for the comment labeling task during the AppealMod process

Comment (*Content Warning: Below comments include toxic language*)	**Allowed?**
Hahaha, not at all, I love a good discussion, especially with a retard on the internet, but you're too dumb	No
You're still a fat fuck, k..ill yourself	No
I'll hire my hacker friend to doxx you then send a hitman to give you ebola then shoot you in the head and revive you to give you AIDS	No
Even if your mom is a fucking asshole, I believe you that she is, drugging her will NOT do anyone any good	Yes
Fuck him. You dont want or need those assholes in your life. Make a new friend	Yes

Table 7. Comments used for the comment labeling task in which users were asked to select comment(s) that they think should be allowed in the community. These comments were reviewed by the moderators and they selected two of them as permissible.

A.5 Impact of different thresholds on toxicity classification

	Toxic appeals in control	Toxic appeals in treatment	Toxic appeals that used AppealMod	Rate of toxicity in control	Rate of toxicity post-AppealMod
Threshold=0.5	79	81	8	18.03%	6.1%
Threshold=0.7	58	57	5	13.24%	3.81%
Threshold=0.9	32	29	2	7.3%	1.52%

Table 8. Moderators' exposure to toxicity under control and treatment at different toxicity thresholds

A.6 Protocol for the post-experiment session with moderation team

- Have you observed any recent changes in the ban appeals that you're receiving? What about appeals you've been granting?
- When do you usually review ban appeals? Do you have dedicated time slots, or do you review them as they are submitted?
- Now that you have seen our automated bot in action, what are your initial thoughts about it? Does it help with your review process?
- Select a random set of conversations from control and treatment, and ask the following questions for each of them:
  - As you see this ban appeal, would you respond to it? Why/Why not?
  - Would you be willing to grant the appeal? Why/Why not?
  - If treatment conversation: Is there something specific you'd look for in the form responses?
  - If treatment: Did the user's response help you in any way?
  - If treatment: Did it affect your decision? Why/Why not?
  - Are there other factors that impacted your decision? Follow up: what about who issued the initial ban? What about the ban reason?
- Finally, any changes you'd like to make in the design?

Received January 2023; revised July 2023; accepted November 2023