Chillbot: Content Moderation in the Backchannel

JOSEPH SEERING, KAIST, Republic of Korea
MANAS KHADKA, Stanford University, USA
NAVA HAGHIGHI, Stanford University, USA
TANYA YANG, Stanford University, USA
ZACHARY XI, Stanford University, USA
MICHAEL BERNSTEIN, Stanford University, USA

Moderating online spaces effectively is not a matter of simply taking down content: moderators also provide private feedback and defuse situations before they cross the line into harm. However, moderators have little tool support for these activities, which often occur in the backchannel rather than in front of the entire community. In this paper, we introduce Chillbot, a moderation tool for Discord designed to facilitate backchanneling from moderators to users. With Chillbot, moderators gain the ability to send rapid anonymous feedback responses to situations where removal or formal punishment is too heavy-handed to be appropriate, helping educate users about how to improve their behavior while avoiding direct confrontations that can put moderators at risk. We evaluated Chillbot through a two week field deployment on eleven Discord servers ranging in size from 25 to over 240,000 members. Moderators in these communities used Chillbot more than four hundred times during the study, and moderators from six of the eleven servers continued using the tool past the end of the formal study period. Based on this deployment, we describe implications for the design of a broader variety of means by which moderation tools can help shape communities' norms and behavior.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: chatbot, interaction design, community interaction, moderation, Discord

ACM Reference Format:
Joseph Seering, Manas Khadka, Nava Haghighi, Tanya Yang, Zachary Xi, and Michael Bernstein. 2024. Chillbot: Content Moderation in the Backchannel. In . ACM, New York, NY, USA, 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Community moderators are often viewed effectively as hall monitors, taking routine punitive action on content as it passes through the public, frontstage view of the community. As a result, many moderation tools are myopically focused on behaviors such as content removal that take action on this front stage, but effective moderators also operate in the *backchannel* — behind the scenes, in small group or one-on-one conversations — they welcome newcomers, write rules and establish norms for behavior, explain punishments, help offenders understand how they might reform, and more [44, pp. 12–18]. However, tool support for these often-backchanneled behaviors is much sparser. It is not surprising that moderation tools are typically designed around visible actions such as removal, as these actions are a tempting first step as moderators seek to prevent others from seeing harmful content. Indeed, some moderators spend much of their time removing

CSCW '24, November 9–13, 2024, San José, Costa Rica

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

violations [28]. However, such tools are poorly suited for many of the broader social situations that community moderators face. For example, such tools are often intent-agnostic, in the sense that their design may implicitly convey a judgment that the community member intended to cause a problem, causing many recipients to (incorrectly) assume moderator ill will [5, 18, pp. 8–12], and potentially initiate a spiral into more negative behaviors [10]. Similarly, removals without accompanying explanations may drive away well-intentioned users who simply made an honest mistake. Could tools support alternative, often backchanneled, modes of shaping pro-social community behavior?

In this paper, we explore how tools can support moderator feedback in the backchannel as an instance of this broader range of goals inherent in community building. We introduce one example of such a tool: Chillbot, a Discord bot that gives moderators the ability to quickly send private, backchannel feedback to users who are close to crossing a line or may have accidentally broken a rule, giving them a chance to adjust their behavior before any formal action is taken. Chillbot's core user flow, shown in Figure 1, involves a moderator sending an anonymous, gentle nudge to a user. The user who receives this nudge sees it in a private thread, allowing them a separate space to react to the nudge outside the public conversation. While the moderator who sent the nudge is not shown to the user — protecting the moderator from retaliatory harassment — the moderator team is able to see any messages the user sends in response to the nudge. With this approach, we meet calls for incorporating user education as part of the governance process [48, 52], building on prior work that has shown the value of giving users an opportunity to adjust their behavior [20, 27].

In designing and iterating on Chillbot, we aimed to incorporate feedback from moderators at every stage of the process. We first performed formative interviews with volunteer moderators on Discord about their current strategies for setting and conveying rules and norms, both to confirm that prior work on moderator practices on other platforms extends to Discord and to highlight moderation practices that could benefit from greater tool support. Next, we created image and video storyboards showing tool concepts and how they might be used in hypothetical scenarios (Figure 2) and gathered feedback from moderator interviewees about whether they felt that the scenarios could plausibly happen as depicted in these storyboards. During deployment in real communities, we gathered feedback from moderator testers in order to identify the most commonly requested adjustments and updates. Minor updates, e.g., adjusting message formatting, were made throughout the study period, while major updates were made between the first and second recruitment wave as detailed below in Section 5.

In an evaluation, we deployed Chillbot for two weeks in eleven Discord servers, each with membership ranging from 25 to over 240,000 people, for a total summed membership of nearly a half million people. Chillbot was used more than four hundred times by moderators during the study period, most often to gently prompt community members to reconsider whether their content was appropriate for the topic of a channel or to remind a community member of norms or rules. Moderators felt that Chillbot had a positive behavioral impact that was aligned with their intentions in using it: users understood the purpose of the nudge, and generally did not escalate or repeat the behavior again. We found that moderators on mid-sized servers with active moderator teams naturally integrated the tool into their workflows, using it multiple times per day, while moderators in large servers and servers with less socially-engaged moderation teams were less likely to find the tool useful. Moderators in more than half of the servers continued to use Chillbot after the end of the study period, with uses of it on one server occurring more than six months after the study without any further prompting or communication from the study team.

With this work, we demonstrate the value of designing moderation tools that support more socially-engaged moderation practices. We begin by reviewing prior work and then we detail our formative interviews, the Chillbot tool itself, and our evaluation. We conclude by discussing

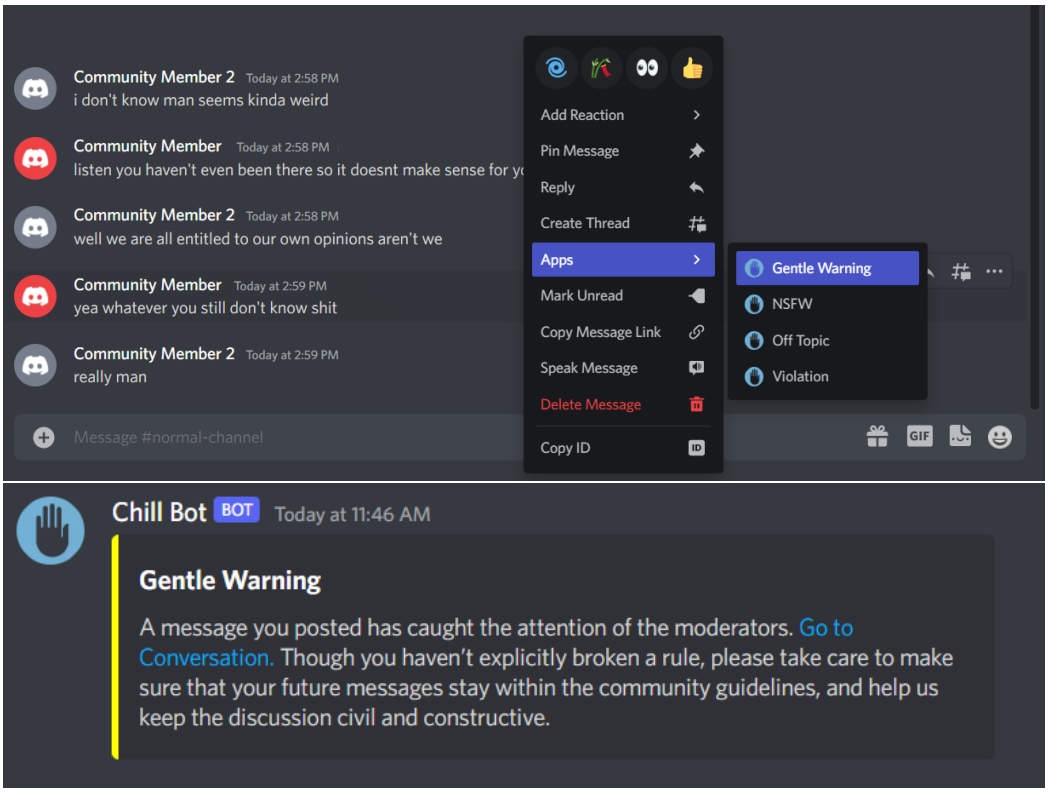

Fig. 1. **Top:** Moderators trigger Chillbot through a context menu for a message, then choose the nudge they want to send. **Bottom:** the user who posted the message receives a message in a private thread with the "Gentle Warning" nudge.

implications for the design space of non-removal moderation tools and how they might impact moderator workflows.

2 PRIOR WORK

In this section, we review challenges in community moderation, the strategies moderators employ in response, and the tools that they rely on.

2.1 Community moderation as a socially-nuanced practice

Volunteer moderators handle a wide variety of issues in their communities. Some of these relate to content quality — e.g., dealing with spam or content that does not meet community standards — and others relate to more social behaviors like harassment, hate speech, and interpersonal conflict. Prior work has identified a breadth of complex sociotechnical practices that moderators engage in to deal with these issues. At a high level, moderators must write rules and establish norms that can evolve over time as a community grows [44]. On a more individual, moment-to-moment level, they must make decisions about how to handle potentially-problematic behaviors that arise while also acting to encourage more prosocial behaviors [43]. Moderators can consider various factors in determining how to respond to a particular behavior, including the content of a post, but also evidence of past behaviors and indicators of status in or commitment to the community [5].

Throughout this process, moderators also have to communicate with each other to ensure that their goals and actions are in sync.

In addressing problems that arise, volunteer moderators adopt many different roles that vary from moderator to moderator. Seering, Kaufman, and Chancellor [42] catalogued twenty-two different social metaphors with which moderators self-identified. These included metaphors that align with the procedural and punitive visions for the roles of moderators, such as moderators as "police" or "governors," but also included metaphors such as "gardener," "protector," and "teacher" which imply more nurturing approaches to moderation. Similarly, in studying moderators on Twitch, Wohn [53] identified four roles that ranged from the punitive "justice enforcer" and "surveillance unit" to the more nurturing "conversationalist" and "helping hand." This breadth of self-identified social roles define many different approaches for community moderation—yet, as we argue, tool design is over-invested in a small subset of these roles.

The strategies that moderators take can be roughly divided into two categories: proactive, meaning actions that occur before a particular issue arises, and reactive, meaning actions that take place after an issue arises in an attempt to address it. These overlap with Grimmelmann's categories of Ex-ante and Ex-post, though Grimmelmann focuses more on a regulatory model for moderation and less on interpersonal engagement and education [15]. Kiesler et al. [27] identify numerous strategies that moderators might deploy, each of which has been shown to be effective in subsequent work, including writing rules [32], creating filters to screen out certain types of content [19], and highlighting examples of good behavior [43]. Similarly, Kiesler et al. list reactive strategies such as removing inappropriate content or moving it to a space where it is more appropriate, giving explanations for punitive actions taken, giving users face-saving ways to correct their behavior, and banning users when necessary [27]. Subsequent work has identified uses of variations of each of these strategies by volunteer moderators on multiple modern platforms ranging from Facebook [44] to Reddit [13, 19, 20, 24, 44] to Twitch [5, 6, 44, 53] and Discord [23, 26, 40].

More broadly, both platforms and volunteer moderators may take steps to educate or reform users. Literature on restorative justice in online communities [54] emphasizes the value of bringing users together to help shape mutual understanding of the impact of actions and to determine how to repair injuries caused. Similarly, scholars have emphasized the value in transparency in content moderation across platforms, particularly in cases where users are confused about what they have done to be punished [52]. Some platforms have implemented processes for appeals, but these appeals processes frequently seem arbitrary or difficult to navigate [49, 51, 52], and their heavy-handedness and limited flexibility can constrain users' behaviors into pre-defined templates for what is acceptable [12].

2.2 Tools for moderation

Regardless of their level of social engagement, most community moderators will need to use tools to address issues that arise. The primary tools used in online communities allow moderators to either withhold content until violations are corrected, delete content, temporarily ban ("time out") users, or permanently ban users. Some platforms also allow moderators to restrict who can participate in a community using features like Facebook groups' pending member questions[1] and Twitch's chat verification tools.[2]

Research on technical solutions for handling problematic behaviors has focused primarily on identifying and removing problematic content and/or users. Though some research has described the development of tools for supporting different forms of online social behaviors by, e.g., summarizing

[1] https://www.facebook.com/help/200755420421098
[2] https://help.twitch.tv/s/article/chat-verification-settings

conversations [56], removal remains a major theme of the moderation tool literature: for example, Crossmod [7] helps moderators on Reddit identify comments that may violate subreddit norms, flagging them for moderator review, while FilterBuddy [21] and ModSandbox [47] both help craft algorithmic word filters in different contexts. Likewise, in practice, moderation bots on Twitch are designed to automatically remove content that meets certain conditions, and can also help with certain administrative functions [41]. Technical tools on Discord, often modeled after similar tools on Reddit, help with managing content removal, tracking user punishments, and logging other moderation actions [26]. In spaces beyond traditional online communities, tools like shared blocklists allow users to collectively manage a type of crowdsourced moderation capacity [22], an approach to moderation that aligns with the philosophy of Mahar, Karger, and Zhang's Squadbox [30] but on a larger scale.

Likewise, beyond the deployment and testing of specific tools, central emphasis in technical literature on moderation has gone into improving methods for detecting and removing problematic content at scale. Various datasets (e.g., [8, 31]) have been compiled so that researchers can test different algorithms and approaches to detecting different types of problematic content and can compare their results to other published work. Recent research has also begun to explore detecting hate speech in languages other than English, including Spanish [34], Bengali [35], Greek [33], and even code-switched languages [45]. Improvements in the ability to detect problematic content have led to the development of tools that can identify problems before things get out of hand; Schluger et al. developed a prototype tool concept that proactively identified Wikipedia Talk pages that were predicted as having higher likelihood to derail, which could allow moderators to intervene early on [37], and research studying a Twitter intervention prompting users to pause and reconsider a potentially-offensive Tweet before posting significantly reduced offensive behavior among prompted users [25]. On the other hand, Bao et al. [3] were able to identify conversations that would result in prosocial outcomes by examining their first comment, which could allow moderators or platforms to highlight valuable comments in time to inspire others to follow their example.

Existing tools that support proactive and non-removal activities are more sparsely populated in the literature and less-frequently used directly in real time by moderators. Instead, asynchronous approaches are more common. For example, one common tool enables moderators to pin, or sticky, guidelines posts to highly visible places in the community [32]. By adapting term lists for automoderator tools [19], moderators may likewise create a bot that automatically replies with guidance when a post is suspected of violating a community norm. Tools might instead focus on the community member rather than the content, for example showing broad background information about the person and their other online activities [17]. Finally, communities may adopt existing tools to engage in restorative justice activities instead of traditional removal or punitive approaches [54]. However, none of these tools are available for moderators to make use of, or to intervene in real time as a conversation is unfolding: they are restricted to before the behavior or after. Our work expands this design space to consider moderation tools for such feedback in-situ, using a backchannel as a design mechanism.

In the present work, we aim to better characterize this negative space in the literature by focusing on the space of tools that can help moderators shape community behavior without relying primarily on removal. The research covered in the first part of this section has clearly shown that moderators take different approaches to dealing with problems depending on the specific nature of the content, the context in which it is posted, and the inferred intent of the person posting it [5]. However, despite work showing their importance to communities [46], relatively few *tools* have been developed and tested in real communities for supporting moderators in handling many of the situations that moderators may encounter, such as a burgeoning conflict that has not yet crossed a line but appears

likely to in the near future, an honest mistake by a well-intentioned user who wasn't completely clear on the rules, or a regular community member who became a bit too comfortable with their status and pushed a line. In each of these cases, moderators might find success by backchanneling with the person, and thoughtfully designed tools may offer some of these benefits. In this paper, we aim to extend prior work by expanding the design space of non-removal moderation tools by introducing a moderation tool for backchannel-based feedback.

3 DESIGN PROCESS

In creating a moderation tool to be used within Discord communities, we aimed to draw insights both from prior work on community moderation and from feedback from users involved in our design and development process. In this latter regard, we followed the design process previously used to create Squadbox [30], a moderation system to combat email harassment based on a combination of preliminary interviews with users and feedback from users who tested the system. Matching this process, our first step toward developing a tool was to conduct interviews with Discord moderators about strategies they use for addressing problems aside from removal, and to gather feedback on potential tool concepts. Moderator interviewees were recruited from a Discord server dedicated to discussion about moderation, where the majority of members are active moderators in servers of varying sizes and topics on Discord. A post was made in this server with the permission of the server administrators announcing the study and its goals and recruiting participants. In this work we elected to focus on building a system for Discord both because of the community-based social structure of the platform and because of the flexibility in tool design allowed by the Discord API. Preliminary interviews lasted approximately 20-30 minutes, and participants were compensated with an Amazon gift code for $15 or equivalent in local currency. Participants were asked about their strategies for identifying problematic behaviors before they escalate, the factors they take into account when determining what action(s) to take, and the non-punitive actions they incorporate into their moderation work. Following preliminary interviews, we generated concepts for potential tool designs and then returned to these same interviewees to gather feedback. This interview study was approved by the Institutional Review Board at Stanford University.

3.1 Preliminary interviews on non-removal approaches to moderation

We performed preliminary interviews with eight Discord moderators[3] to get a broad sense of potential directions for a tool. Though relatively little work (e.g., [23, 26]) has studied volunteer moderation specifically on Discord, we did not aim with this research to conduct a full additional interview study on volunteer moderators' practices on Discord. Instead, in accordance with a community-centered design process, we aimed to determine whether and how findings from prior studies on moderator practices on other platforms applied to this new context, rather than moving directly to tool development under the assumption that they would. The questions we asked in these formative interviews (see Appendix) were informed primarily by Cai and Wohn's work [5] on Twitch moderators' profiling processes for offenders, as well as Seering, Kaufman, and Chancellor's work [42] on different philosophical approaches to community moderation, as these works provided a starting point for discussing how moderators determine how to apply tools to various types of conflicts.

Interview transcripts were jointly coded by two authors using a grounded theory based approach to identify themes. As the goal of these interviews was to understand how moderators employed

[3]Quotes from formative interview participants are labeled with *F*, while quotes from exit interview participants are labeled with *E*, their server number, and an additional identifier (A, B, C, or D) if they were one moderator of several from that server.

both removal based and non removal based approaches to moderation, coding focused on identifying themes in these areas [11]. Following a first round of coding, the authors met to discuss the findings and the resulting categories. A second round of coding was then performed by the same two authors to refine the themes. The resulting themes are discussed below.

All moderators interviewed reported having previously used removal tools in moderation — removing both messages and users from their communities. This was typically done through the use of one or more dedicated moderation bots, which could be triggered by the use of various preset commands (see [26] for evaluation of the technological frames for bots in this category). While the specific bots varied, a common flow emerged: moderators either noticed a problem or were pointed toward it via a user report, and then they assessed the situation to determine what type of removal was necessary. Removals were of varying severity (i.e., content removal, temporary bans, permanent bans) and were considered as options depending on the context. This matches findings from prior work on user profiling [5].

Moderators also reported a variety of non-removal approaches to handling problematic content and/or users, which took place both before and after problems occurred:

Displaying rules. All of the moderators we interviewed reported displaying rules to new arrivals in some form, and in some cases they required members to formally agree to the rules before they could participate in the community at large. Moderators would list a set of rules and require newcomers to react to the rules with a specified emoji in order to signal agreement before they could participate in the community, after which they would automatically be granted access to the server. Moderators were mixed on whether requiring users to agree to a set of rules significantly affected behavior, though prior research has suggested that prominent display of rules can impact user behavior [32]. However, all agreed that it was valuable to write a set of rules both in order to consider what behaviors were appropriate and to have a reference to point to when moderation actions were taken.

Public presence. Some moderators believed that their visible presence on its own would impact how users behaved. They felt that reminding users that moderators were watching could encourage them to think twice before posting potentially-problematic content, an idea supported by prior work on moderators' social influence [43]. These moderators also mentioned that taking a slightly more active role—participating directly in conversations that appeared to be escalating in a problematic direction in order to steer them in a more positive direction—was an effective strategy. This approach, however, required significant investment of time both to monitor conversations and to determine how best to engage publicly with users in a way that wouldn't be counterproductive.

Personalized private communication. Another common social process for dealing with problematic behavior after it had occurred was to privately engage with an offender in order to explain why the moderator intervened or to warn them that their behavior was problematic. Prior work has found explanations of removals to be effective at reducing future problematic behavior [20]. However, most moderators noted that their use of this strategy was limited by two factors: the time it took to write out an explanation that fit the situation, and concerns about potential reprisal. As noted in prior work [40], many Discord moderators are targeted for harassment by community members, and in some cases such harassment emerged from users who were upset about being called out, criticized, or punished. This phenomenon served as the primary inspiration for the anonymous nature of the interaction supported by Chillbot, as we discuss below.

Formal warning system. Moderators also reported using a formal warning system. When a user broke a rule in a way that moderators determined was not egregious enough to warrant a temporary or permanent ban, the offending content was removed and the user was given a written warning.

Users who received a predetermined number of written warnings were then punished by temporary or permanent removal from the community. Moderators noted that even though a written warning was not always considered as full punishment, users often reacted as if it were.

Broadly, a common theme that emerged from these interviews was that, in line with the literature discussed above, moderators in these communities engaged in a variety of social practices to shape behavior, but the moderation tools that these moderators used were primarily designed to support removal as the primary approach to moderation. This was the case both for platform-provided tools and user-created tools. While moderators did spend time engaging with users about the rules and their behavior, this process was time-consuming, very difficult in larger spaces, and — critically — lacked tool support. Lastly, the systems in place in each of the communities vary heavily and demonstrate how when conducting fieldwork in these environments, the various systems in place need to be accounted for. Therefore when designing non-removal based tools to be tested in active communities, they are limited in that they need to be designed to fit into the existing, mostly punitive moderation ecosystem.

3.2 Iterating on Tool Concepts

Following our initial interviews with moderators, we regrouped to consider possible design concepts for tools that might support moderators in addressing problematic behaviors aside from removal. We elected to focus on the emergent theme of challenges in responding to problematic behaviors where offenders' intentions were unclear. As moderator interviewees noted, it is likely that many of the problematic behaviors that occur in online social spaces are the result of either a) users who are unfamiliar with the rules and unintentionally make a mistake, or b) users who forget about the rules due to contextual factors (e.g., participating in a heated debate) and then accidentally cross a line [9]. These two cases share a common attribute that we highlighted as an opportunity for design: in both cases, the problematic behavior could have been prevented if the user had been reminded about the rules at the right moment, and recurrences of the behavior could likely also be prevented by pointing out the mistake.

Following a round of ideation, we returned to our interviewees and presented them with the two storyboard concepts shown in Figure 2. The figure shows two scenarios that could occur in an online community. In the first, two users are having a heated conversation that may escalate, so a moderator gives an anonymous nudge to let them know that they should calm down. In the second, a new community member is making mistakes with their behavior that don't warrant punishment but need to be addressed, so a moderator gives an anonymous nudge to inform them about how they're expected to behave. Moderators were asked to give feedback on whether they identified with the scenario being described and whether they could see it occurring in their communities. Both of these concepts received generally positive feedback; moderators accepted both of the stated problems as real issues — it wasn't uncommon for newcomers to the server to misunderstand how to behave, and heated conversations sometimes did lead to problems — and agreed that both situations could benefit from some sort of intervention. Feedback about the imagined form of this intervention prompted three questions about where such a tool might be most useful.

Should users be engaged with personally, or will an automated message suffice? Though prior work has shown that personalized messages from moderators and pre-written bot messages can both be effective [20], moderators identified different cases where each type of message might be preferable. In cases where moderators felt that a user might not immediately understand why what they did was problematic, moderators preferred to be able to have a conversation rather than to rely on an automated message. On the other hand, in cases where the issue was fairly clear, e.g., a reminder to

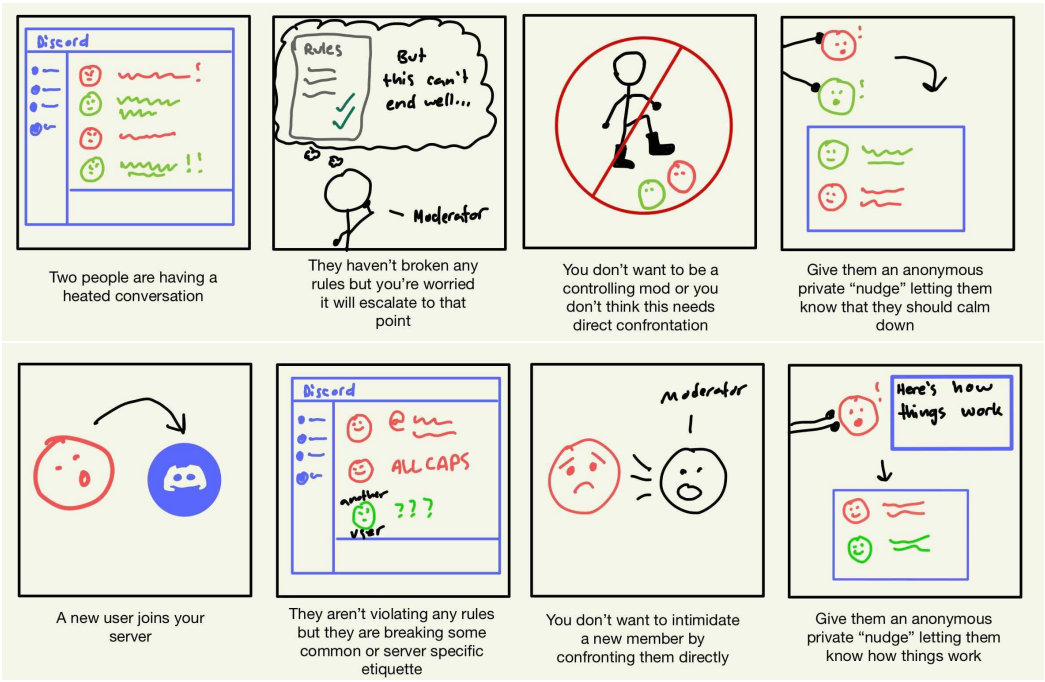

Fig. 2. Two storyboard concepts presented to moderator interviewees.

post NSFW content only in the designated channels, moderators didn't feel that a conversation would be productive.

Interviewees were split on how they thought feedback from a bot would be received by users. Some thought that users would be off-put by automated feedback, finding it either low-effort or impersonal. Other moderator interviewees felt that being contacted by a bot rather than being confronted by a human moderator would seem less confrontational. Moderators also noted varying opinions about the general roles that bots should play in communities on Discord, which were based on personal experience from the moderation pipelines already in use in their communities and the roles the current tools in those pipelines already played. For most of the moderators, existing bots in their servers were primarily reactive tools that were used after problems had already occurred.

Should the feedback be provided anonymously or by a named community member? Moderators identified two cases where being able to give feedback anonymously would be preferable — when a user's behavior indicated that they might react in an especially aggressive way if confronted, or when users were new to a space and might be driven away by a direct confrontation from a moderator.

> *"I never really though about giving anonymous nudges before. It does definitely help with intimidation and de-escalation though, those are very interesting approaches to the situations described. I think if I was a new user, I'd feel better with an anonymous nudge rather than someone I knew to be a mod coming and discussing the situation with me." – F6*

All moderator interviewees recounted cases where they had been insulted or harassed by community members who were upset about a time out or having their content removed. Moderators

thought that being able to issue a reminder or warning anonymously might lead to a lower likelihood of harassment because of the lack of a clear target.

When should feedback occur in public, and when should it occur in private? Moderator interviewees felt that public feedback would be more valuable if there was potential for having a useful conversation about the issue in question. If the people involved could come to a better understanding through a public discussion, that method was preferred. Similarly, if a reminder could benefit multiple users, as in the case of a simple reminder that a certain type of content should be posted in a different channel, public feedback might be preferred. On the other hand, some interviewees felt that public feedback could be viewed as more confrontational and might be counterproductive in some cases because users might feel embarrassed to be called out in front of their peers.

> *"People respond negatively to an admin popping into the middle of a conversation and asking you to move from channel A to channel B. I would say the vast majority of the time, it kills the conversation. I [would hope] that this tool would be able to help with that organically." – F1*

These interviews provided a variety of different questions to consider during our design process for a tool that supports moderators in taking less punitive, less removal-focused approaches to moderation. We drew from these findings that the best target cases for anonymous, private, automated messages were cases where the involved users might react negatively to direct confrontation, violations were fairly simple, and where there was not clear value to be gained from having a more time-intensive public conversation about the problem. In the following sections, we describe the result of this design process, how it addresses many of the issues raised above, and how we tested it in real communities.

4 CHILLBOT: A TOOL FOR PRIVATE BACKCHANNELING IN MODERATION

In this section, we introduce Chillbot, a moderation tool developed to aid moderators in shaping community behavior through private backchanneling. There exists a gray line between appropriate and improper norms within any community: in this setting, it becomes hard for automated moderation bots to operate effectively and for human moderators to take a nuanced approach without dedicating too much effort or time in large communities. Chillbot focuses on a private backchannel instead of visible public removal in order to expand the expressive vocabulary that moderators have at hand. As moderators noted in the formative interviews, pulic call-outs can be a moment of embarrassment for a user and can cause defensiveness and retrenchment, especially in cases where a user hasn't technically broken any rules. It can also be intimidating for new users, potentially driving away well-intentioned newcomers [16].

Chillbot allows moderators in these communities to set up customizable *nudges*, which can then be sent in response to a particular message (Figure 3). Rather than forcibly and visibly removing content, backchannel moderation via Chillbot sets up a private message thread with the user and lets them know that a moderator has taken notice. A nudge, in this context, is a predefined message that is sent to a user, warning them that they are close to violating rules in a given community. Moderators can create a set of such messages to capture different common situations. Table 1 shows the default nudges that were provided to moderators. Chillbot's nudges are anonymous and private, which protects the moderator from retaliation that they might have received from a defensive user if the nudge was made with the moderator's name revealed.

We implemented Chillbot on Discord, as noted above, in part because of its social structure and in part because of the flexibility of the API. We created a set of four sample "nudges", with each matching a common use case identified in the formative interviews. For example, a specific moderation "event" might be addressing concerns caused by a user posting media that might be

Nudge name	Nudge text
Gentle warning	A message you posted has caught the attention of the moderators. [Link to Conversation]. Though you haven't explicitly broken a rule, please take care to make sure that your future messages stay within the community guidelines, and help us keep the discussion civil and constructive.
Violation	A message you posted has been flagged because it violates a norm or convention for posting in this server. [Link to Conversation]. Please consider pausing to get to know the expectations better before continuing.
Off-topic	A message you posted has been flagged as off-topic for the channel where it was posted. [Link to Conversation]. The message does not explicitly violate any rules so it will not be removed, but please take care in the future to find the best channel to put this type of message.
NSFW	A message you posted has been flagged as borderline NSFW. [Link to Conversation]. Please be careful to keep NSFW content only to approved channels.

Table 1. The four default nudges participants were provided, which could be customized or supplemented with additional nudges.

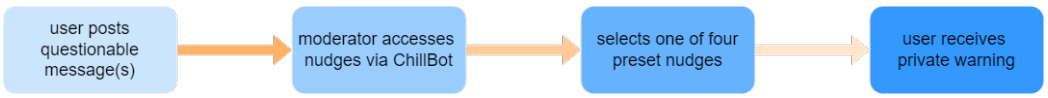

Fig. 3. Moderators use Chillbot to react when removal is not the best response, for example for borderline or accidental rulebreaking.

borderline NSFW (Not Safe For Work), after which a Chillbot nudge could be triggered on the message in question through a context menu interaction on the desktop app (Figure 1) or a slash command on the mobile app (Figure 4). The user would then receive a private message with the predefined nudge message and a link to the problematic message to provide context. When the bot is triggered, it posts a log detailing the case in a channel designated by moderators (See Figure 5). Moderation logs are common on major Discord bots, as teams of moderators rely on logs to track what moderation actions have been taken and what content has been flagged or removed. In this case, in addition to their primary purpose of supporting coordination, these logs also helped with norming [50]: each moderator could see how the others were using Chillbot and could have discussions about these cases in the channel or elsewhere.

4.1 Implementation

Chillbot was built using Javascript and runs using Node.js and MongoDB, and was hosted on Heroku. The chatbot was created and initialized through the Discord API and was given features using the node.js package called discord.js. The bot acts as a client, which can be connected to various discord communities. Discord bots can be used to perform a number of tasks, both invokable and automated. Bots can be added to any server by a moderator of that server, who grants it a number of permissions that decide what tasks it can and cannot perform.

Our bot utilizes two Discord features called Slash Commands and Context Menus, and it is through these features that a customizable nudging system is created. Context Menus allow for commands to be invoked by right clicking a message, which we utilized in order to give the

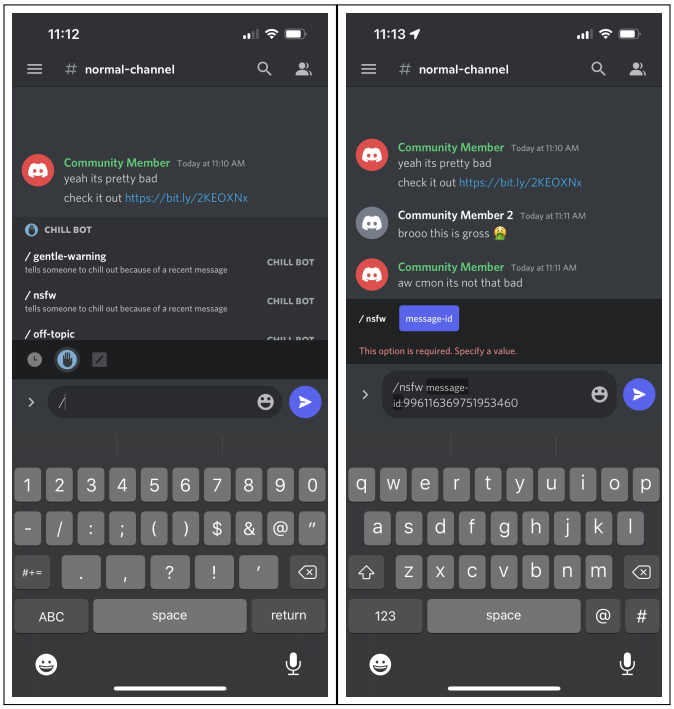

Fig. 4. Mobile Chillbot use flow. Moderators type "/" into the regular chat box, and an autocomplete menu appears with possible bot commands. Moderators can choose which reminder to send by selecting the corresponding command and pasting the ID of the message in question.

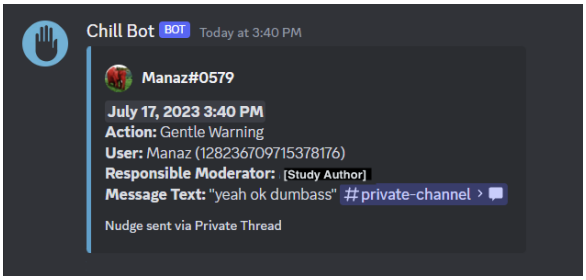

Fig. 5. A sample nudge log including information about when and how the bot was used. This example was created by the study team for demonstration purposes.

moderator the option to invoke any of their server's premade nudges on any message. The lack of this feature on the mobile version of Discord prompted us to utilize Slash Commands. Slash Commands offer the same functionality as Context Menus, except they are triggered via a message beginning with a slash, then a command-line-style input.

After developing a first version of Chillbot, we deployed it in partnership with moderators of six Discord servers, gathering feedback as we went in order to address issues that arose. Our goal in this iterative deployment was to work through cycles of feedback in response to the communities' needs.

As such, the deployment period involved regular back-and-forth conversation between the research team and moderators. Across all of the servers, moderators and the study team exchanged more than 1,000 messages discussing the tool and potential improvements. In some cases, moderators would send the study team a message about an issue they had discovered or a feature they would like implemented, while in other cases the study team would perform a periodic check-in to see how things were going or to update moderator participants on a new feature that had been added based on their feedback or feedback from other participants.

In order to protect the privacy of community members on the participating servers, the data tracked included only the number and type of nudges invoked, the timestamp of the nudge, and the server in which the nudge occurred. Conversation logs from the participating servers were not collected. Instead, we use examples provided by participating moderators as the basis for our analysis.

5 EVALUATION

Did moderators make use of Chillbot to help shape their communities? In this section, we describe the process for evaluating Chillbot. We find significant usage of the tool across the majority of participating communities, and we summarize the most common use cases and feedback. We conclude by identifying major design challenges that arose from the results.

5.1 Method

Chillbot was deployed in eleven different Discord servers with membership ranging from 25 to 240,000 members for a minimum of two weeks each. These servers, in total, featured a summed membership of nearly a half million members. Recruitment proceeded in two waves – after completing the first six servers, recruitment was paused to allow for time to make adjustments to the bot, and the updated version of the bot was used in the final five servers. The first version of the bot relied heavily on direct messages to send nudges, but moderators in the first wave of servers reported difficulties reaching users due to the various permission settings that could block direct messages. During the second wave, the primary messaging option was instead set to use private threads,[4] which acted as a form of in-server direct messaging, and moderator participants reported much higher rates of success in reaching users after this change was made. The revised version of the bot also output visually cleaner logs to the designated log channel; while the original logs were in plain text, the updated logs were in an embedded graphic that matched the style used by many other popular Discord bots with which participants were already familiar. Aside from these two changes, the core nudge functionality of the bot remained the same through both deployment waves.

For this study, moderator participants were recruited from meta-moderation servers and moderation hubs on Discord. In this type of server, active moderators gather to discuss topics related to moderation, including philosophies, processes, and tools. Because we recruited from these communities, participants in this study were typically moderators with a broader interest in the philosophies and practices of moderation; though we did not specifically ask moderators about their philosophical orientation using, e.g., taxonomies of metaphors [42] or categories [53] of moderation styles, participants likely brought awareness of a diverse array of approaches to moderation that they had gained through discussion with other moderators on these servers.

We posted advertisements for the study with permission from the server administrators and messaged Discord moderators from the servers who had indicated interest. For each server, we

[4] https://support.discord.com/hc/en-us/articles/4403205878423-Threads-FAQ

Wave	Server	Category	Member Count	Total Nudges	Duration of Use
1	1	Gaming	650	3	14 days
1	2	Gaming	20,000	13	14 days
1	3	Gaming	5,300	166	145 days
1	4	Web Development	37,000	43	192 days
1	5	Remote Work	150	1	14 days
1	6	Community Management	25	2	14 days
2	7	Professional Development	25,000	84	36 days
2	8	Gaming	700	12	52 days
2	9	Anime	6,500	64	14 days
2	10	Anime	240,000	59	71 days
2	11	Gaming	150,000	19	29 days

Table 2. Total usage of Chillbot across participating servers. Participants who enrolled in the study were required to try the tool for a minimum of 14 days, after which they completed their exit interview and received compensation. However, some servers continued to use the bot after the study ended without further compensation. The "Duration of Use" column is calculated from the date of first use to date of last recorded use.

onboarded a team of up to four moderators who would be using the tool during this study. Participants were presented with documentation to explain how the bot worked, including a short video and a longer document with instructions for use and answers to common questions. They were then offered the chance to modify the text of the "nudge" messages that the bot could send to fit their server's needs. Moderators from eight of the eleven servers chose to edit the nudges. Of these, six servers kept the four default nudges but edited the wording to be more specific to the context of their servers, while the remaining two each added another nudge. One server chose to add a nudge to remind users not to post memes in a particular space, and the other changed the "Violation" nudge to focus specifically on disturbing content and renamed the "Off Topic" nudge to "Wrong Channel". Of the default nudges, the most frequently edited was the "NSFW" nudge; the default text for this nudge reminds users only to post NSFW content in approved channels, but four of the servers prohibited NSFW content entirely so the text was edited to reflect this.

When customization was complete, moderators from these servers were asked to install the bot in their server and test it to make sure that it functioned correctly, but they were not given any formal requirements for how much to use the tool during the study. Upon completing the two week study period, they participated in an exit interview where they were asked about their experiences, after which they were compensated with an Amazon gift code for $50 (or local currency equivalent). Though some of the servers had more than four moderators, we capped the number of participants to four per server in order to fit within a maximum of $200 per server. This limit was set in order to fit within the budget established for this study. There were no cases in which this limitation caused any issues reported by participants; in all cases, servers were able to designate four or fewer moderators who would participate in the study. Participants in the second wave of tests – servers seven through eleven on Table 2 – were also given the opportunity to catalog specific examples of cases where the tool had been used for $1 per cataloged case, and moderators from servers seven, nine, and eleven chose to do so for a set of 86 total cases. For privacy reasons, our implementation of Chillbot did not store conversation logs from the servers it was installed on, so this cataloging was necessary.

In general, platforms made successful use of Chillbot. On six of the eleven servers, moderators continued to use the tool past the end of the study date. The total duration of use ran from a minimum of 14 days — the required duration for the study — to a maximum of 192 days. Servers that continued using the tool past the end date of the study did so at their own volition; they did not receive any additional compensation or encouragement from the study team to continue using the tool. Some moderators on these servers expressed that they had incorporated Chillbot into their regular moderation practices. However, in smaller servers, the tool was used much less frequently; in one small professional server, it was used only once.

In this section, we discuss findings from the study, focusing primarily on feedback from the exit interviews. One researcher performed the exit interviews by Discord audio call, which lasted between approximately 5 and 40 minutes, with variation in length depending on how much the moderator(s) used Chillbot and how much feedback they wanted to share about their experiences and potential future directions. Text was separated into chunks of varying sizes where each chunk contained one core point; these chunks ranged in size from a few words to a few sentences [4, p. 62]. The researcher then identified chunks that applied to three core categories we chose for the purposes of this analysis: "Primary use cases," "Outcomes," and "Challenges." A second researcher reviewed the coded chunks to determine agreement with their classification. We focused on these three categories to guide our analysis on the ways in which Chillbot was used, but also to consider its strengths and weaknesses; though there are many ways a tool like Chillbot could be evaluated, we rely primarily on moderators' reports about the functionality of the tool because we elected to minimize the data we collected directly from participating servers due to concerns over privacy of server members.

5.2 Most common use: an "intermediate" option

The most common use case reported by moderator interviewees was to gently prompt users to reconsider whether the content they were posting was appropriate for the topic of a channel. In some cases, users simply weren't aware that certain content was expected to be posted only in certain channels. Likewise, in other cases, moderators found that users responded well to a reminder.

> *"I think it was particularly useful for kind of one-off messages. So if it was one specific message where a user said something that wasn't quite appropriate, I think it was really useful in instances like that because we were able to point to the specific message in the thread and the user was able to see exactly which message was violating the rules. So I think that was particularly useful." — E7A*

Of the 86 cases that moderators submitted, the most common nudge was the "Violation" nudge with 31 uses, followed by "Gentle Warning" with 25 uses, "Off Topic" with 18 uses, and "NSFW" with 12 uses. Table 3 shares example posts in the deployment Discord servers and the nudges that moderators sent in response to messages of those types using Chillbot.

Moderators were split on their preference for proactive vs reactive uses of the bot. On some servers, including Server 2, moderators mostly reported using the bot retrospectively, after an incident had occurred. The default process for moderation in these servers was primarily dependent on user reports and moderation queues, and moderators did not regularly monitor conversations in the server in such a way that would allow them to intervene before things got out of hand. For some moderators, the window of time in which they felt they could effectively use the bot was a big factor in their usage:

> *"I think the big issue really is that because the idea is to keep people from breaking the rules if they're close, if you're too late with the message then the situation has already*

> *resolved itself one way or the other and there's no point in going back and telling them to chill if the conversation has moved on. And the conversation can move on in a matter of seconds potentially, depending on how things are going." – E4A*

On other servers, such as Servers 7 and 11, moderators were more socially involved in the community and were more likely to use the bot to de-escalate tense situations, as was imagined during ideation and design of the bot.

> *"what I find it most useful is when people are starting to break a rule but they haven't really broken anything and that really helps them to just let them know that they're actually in the process of breaking rule and they should stop." – E11B*

This difference in moderator workflows likely explains some of the difference in frequency of usage of the tool; Servers 2 and 7 were of similar sizes, but the tool was used much more in Server 7, a server where moderators took a more proactive, socially-engaged approach. We discuss this point in more depth later in this section.

Broadly, moderators agreed that the tool was most useful as an "intermediate" option – more formal than a reminder message typed in conversation, but less formal than an official warning that would go on a user's record. Most servers that participated in the study already had moderation bots installed that allowed them to warn users, but these more traditional warning systems were not implemented as gentle nudges but rather as the first steps toward later punitive action:

> *"I felt like it was a better overall experience than using a primary or secondary moderation block because those, there's a stigmatism [sic] to those as far as, 'Oh, this warn is going on my record,' or whatever and, 'How many warns will I get before I am possibly banned or muted?', or something of the sort." – E10B*

> *"I would say it definitely helps give something in the middle in between just a moderator saying with a message like, 'Hey, tone it down,' or something like that and a formal warning. I think there is a gap in between those two things and I think this bot really helped with that and the gentle reminders really helped with that, bridged that gap. So I think users usually reacted less negatively than they would to a full on warn with the gentle nudges." – E7A*

As illustrated in the quotes above, these existing warning systems were designed to be used when a user had clearly violated the rules but more severe punishment was not yet warranted. Warnings were frequently used as "strikes," with a certain threshold of warnings automatically leading to a ban or mute — a sharp contrast to the more gentle, socially-engaged approach that Chillbot aimed to facilitate. Correspondingly, three moderators specifically noted the value of the private thread created by the bot, noting that this opened a space where the user could ask questions about the rules without disrupting other conversations or directly messaging a moderator.

> *"It's definitely nice to be able to talk out a situation with the user in a separate thread as opposed to them being DM'ed by a bot, which they can't really then talk to, and they don't really know who sent that warning and what exactly they did wrong... [they] can ask more about it and I can talk them through exactly what the issue was and how they can go about avoiding that in future." – E11D*

Moderators expressed strong positive responses to the ability to customize the messages used in their servers, while still speaking positively about the value of having a very quick tool that didn't require typing out a full explanation for each issue. This highlights a core value of Chillbot: it allows moderators to send messages that are somewhat more personalized than typical, one-size-fits-all warning messages, without requiring significant time commitment for each case. Some moderators, however, noted that it was time consuming to collectively agree on message text within

Message	Nudge Used	Nudge Text
"Reminder that [name]'s publisher is a communist shill."	**"Gentle Warning"**	Though you haven't explicitly broken a rule, please take care to make sure that your future messages stay within the community guidelines, and help us keep the discussion civil and constructive.
"I want abig japanese [redacted] in my [redacted]"	**"Gentle Warning"**	Though you haven't explicitly broken a rule, please take care to make sure that your future messages stay within the community guidelines, and help us keep the discussion civil and constructive.
"why did you draw [body part] from top view"	**"NSFW"**	A message you posted has been flagged as borderline NSFW. As per rule #4, any NSFW/NSFL, malicious or shocking content will not be tolerated.
"The Ringo case just seems weird"	**"Off Topic"**	A message you posted has been flagged as off-topic for the channel where it was posted. The message does not explicitly violate any rules so it will not be removed, but please take care in the future to find the best channel to put this type of message."
[Image Meme]	**"Memes"**	Hey, you seem to be posting a lot of images/memes in this channel! We have a #spam-and-memes channel accessible as an opt-in role from the #information-center. If a meme doesn't contribute to the conversation, or if its more than 1 - it belongs in that channel. We won't be removing your posts but ask that you follow this guideline going forward.
"*no listen, you need to live under a rock not to know the difference*"	**"Toxic attitudes"**	Hello, this is ChillBot contacting you from [server name] about this message. I wanted to let you know that this comes kind of close to potentially breaking the Toxic attitudes rule, or potentially leading to conversation that will end up breaking that rule. Although things are fine for now, we'd like you to be more careful in the future.

Table 3. Examples of comments on which a moderator triggered Chillbot.

the moderator team at the time of installation, as they felt that all moderators should give feedback on the text. In this study, moderators were given a set of four pre-written messages as a starting point from which to customize, but this suggests that a larger variety of pre-written messages might make it easier to onboard new communities.

Though this version of the tool was designed so that it could only be triggered manually by a direct action from a moderator, some users initially appeared to believe that it relied on some form of automation in determining when to send a nudge.

> *"There was a lot of people who thought it was an AI type of thing, so something that was automatic and people would try to purposely set it off because they thought it was something that was automatic." – E7A*

In the set of cases that they shared, moderators noted four instances where users probed various behaviors to see what would "set off" the bot, assuming that they'd been caught by an automated filter of some sort. Typically in these cases, a moderator stepped in to explain how the bot worked. Two moderators shared thoughts in their interviews about the potential for integrating AI or some other form of automation into the bot, with both being cautiously optimistic about the potential. One moderator, E7D, noted that the consequences for false positives in this type of tool might be

less problematic than in a moderation tool where users actually received a formal punishment or strike.

5.3 Perceived behavior change

Interviewees agreed that, when Chillbot was used, it typically had a positive impact. They reported that users were able to understand the purpose of the nudge, and that users who had been sent a reminder nudge generally did not repeat the behavior in question. Of the 86 specific examples of usage submitted by moderators, 48 resulted in a silent outcome – the user did not cause additional issues, but did not take any further steps to remedy any harm that they had caused:

> *"[It worked well for] spam, little ones that isn't needed of a mute, if that makes sense. So let's say it was very, very mild toxicity or small-level spam, and you give them a gentle warning or a violation, for example. They'll work quite well in that situation because they'll realize what they've done. It's more of, 'Oh, we've caught you. Can you stop?' That sort of thing. They'll normally stop after that." – E11A*

30 additional cases resulted in an visibly positive outcome, where users went back and deleted their message, apologized to other users and moderators, and/or subsequently helpfully explained the rules to other users.

> *"There was this guy that was being toxic in #general, one of the most popular chats on the server, so I used a gentle warning. After that, he came in and apologized and said, "Yeah, we're going to stop." Sort of thing. Then he stopped from that point onwards." – E11A*

Finally, 8 of the 84 cases resulted in negative escalations: three users mocked the bot, four ignored it and continued their problematic behavior, and one left the server.

> *"I didn't see much of negative outcomes to be very honest. There was just two users fighting with each other, but they just kept on fighting. So at the end of the day, giving out a nudge didn't matter to them, I think. […] And the other one was off-topic, the gentle warning. The user was being just being rude, and I gave them a warning and they immediately left the server." – E9A*

Some moderators reported that users were somewhat cautious about the bot, collectively noting what types of behaviors were being flagged and making some effort to avoid those behaviors in the future:

> *"For our server in particular, most of the traffic was in one or two [text] channels out of the hundreds of [text] channels available, so once people became aware of the bot, they got conscious. Not really conscious of, 'Oh, I shouldn't trigger the bot.' But more conscious like, 'Oh, this topic that I'm talking about might [get] a warning from the bot.' So they refrain from [talking about] it." – E3A*

In the above case, a moderator suggested that Chillbot helped shape community norms. We cannot treat this deployment as causal evidence, but it does motivate further work exploring the value of non-removal based moderation tools in shaping communities.

Participants also noted that the design of the tool protected moderators from the harassment that they sometimes face when they engage directly with offending users. Because the nudge message comes from the bot, no specific moderators are linked to the warning and thus users do not have an immediate human target to vent any disagreement toward.[5]

> *"Since the mods are able to issue a warning anonymously, they're much more effective and prevent backlash to the mods by the users." – E9A*

[5]While all moderators can see the nudge and have access to the private thread where the user has been nudged, their usernames are not present for the targeted user to see.

5.4 Design challenges

Over the course of the study, moderators' use patterns of Chillbot and the feedback they provided raised three broad design considerations that apply broadly to tools in this space in considering when they may and may not be useful. As discussed above, the goal of this work was not to develop or advocate for tools that should work in every situation but rather to identify and explore situations where the tools might be more or less useful.

First, some moderators highlighted the general challenge of switching from *reactive* approaches to moderation to *proactive* approaches. In especially large Discord servers, moderators often rely on "mod queue" systems to direct their attention to problems, where users report issues when intervention is needed – an approach that parallels how many social media companies operate. One of the initial driving concepts for Chillbot, as shown previously in Fig 2, was the idea that moderators could use it to privately intervene *before* a conflict escalated to a problematic level, but this concept presumes that moderators would be proactively monitoring spaces rather than reacting to issues that had already occurred and had been formally reported. Possibly because of this, the largest servers in the test group actually used the tool less than the mid-sized servers.

> *"There was a very small window of when we can use the Chillbot. Because [a lot of the time] it's already too late, and we need to escalate to the usual meetings and point system, there isn't that much chance for us to actually use the Chillbot in the appropriate way." – E2*

This highlights a broader issue with developing tools in this space — while proactive interventions allow moderators to leverage their understanding of social context in ways that reactive interventions do not, the established workflows for moderators in some spaces may not lend themselves to the sort of proactive monitoring required for proactive interventions to work. It is plausible that building more proactive monitoring into workflows would reduce the need to rely on reactive (and frequently more punitive) approaches, but this change could require significant time investment from some moderators.

Second, moderator interviewees noted the intersection between social ties and use of the tool. In spaces with strong, preexisting social ties, which were more common in smaller communities, anonymous nudges were less necessary or desirable; servers where all members know each other fairly well, especially in cases where they have offline connections, are likely to find anonymous nudges less useful because server members are more comfortable talking to each other directly about issues that arise. In the opposite case, in servers where personal connections are less common, users may have less of an investment in being receptive when approached personally because avoiding a conversation or engaging less has fewer consequences.

These first two results combine to suggest a relationship between server size and adoption of tools such as Chillbot. Small servers are well-served by direct interaction amongst strong ties already, and very large servers are focused on firefighting clear violations. So, approaches such as Chillbot may be best fit in a "Goldilocks zone" of servers that are large enough to have issues, but not so large that the moderators are focused only on mod queues. However, this best fit on Discord servers does not discount the possibility of a tool such as Chillbot having a different best fit on other online spaces. These results are influenced by the methodologies moderators use on Discord communities, shaped by the inherent properties of discussion based, multi-media online platforms as well as the tools that Discord affords moderators. The moderation paradigm on most of these servers are reactive due to moderators not being able to reasonably monitor chat in real-time, which was the bottleneck that lead to the drop in usage in larger servers. Therefore, Chillbot may have potential to work in larger communities that focus on a real-time moderation paradigm, such as communities surrounding live-streaming.

Finally, moderator interviewees noted the intersection between community goals and tool use. In communities with an inherent normative focus — where a core goal of the community was to encourage certain types of prosocial interaction — back-and-forth discourse about how to behave was valued. This type of discourse was best suited for direct interventions from moderators, whether public or private.

6 DISCUSSION

The results of the Chillbot field study show the potential for expansion of the space of non-removal moderation tools through a focus on backchannel feedback. Our evaluation confirms that this approach can observe adoption in a variety of communities with regular moderation needs. In this section, we begin by discussing our findings with regard to the backchanneling approach used by Chillbot. We then consider questions of moderator labor and automation in moderation tools. Finally, we outline the limits of non-removal tools more broadly, identifying boundaries for the set of cases where they might be useful, and conclude with thoughts about metrics for the success of non-removal tools.

6.1 Backchanneling as a moderation strategy

Moderators found Chillbot most useful as an intermediate strategy for moderation — serious, but not yet punitive. The private, back-channel feedback was gentle enough that, in most cases, it did not trigger defensive reactions from users. The bot's use of the private threads feature created an easy space for moderators to discuss users' behavior with them when needed, and moderators reported positive experiences with this in a number of cases. Though we cannot make causal claims about Chillbot's impact on user behavior, moderators who used Chillbot felt that the bot was effective. This evidence lends credence to the value of backchanneling as a moderation strategy, and it also suggests that Chillbot was successful in supporting moderators' efforts to help users learn to be better members of their communities. Paralleling West's work on transparency in platforms' moderation decisions, Chillbot helped moderators see users as "emotionally engaged in participating in the life" of their communities, "invested in learning from mistakes, and confused about where things went wrong," [52], and this framing suggested pathways to reform that moderators might not otherwise have considered.

We do not claim that Chillbot is appropriate for every context; in this study, we focused on spaces where both moderators and the majority of community members have a generally positive goal for their communities. Chillbot would be less useful in cases where moderators permit or even encourage offensive behaviors in their communities, meaning that it could not be relied on to reduce the prevalence of such communities across a platform.

6.2 Moderator labor

As discussed above, many moderators — particularly in larger servers — are used to moderating reactively; whether by working through a moderation queue or by responding to cases where they have been tagged by users, the majority of these moderators' time is spent on moderation actions taken after problematic content is posted rather than before. Shifting to an approach to moderation that requires proactive monitoring of spaces would require a significant investment both in changing procedures and in transitioning between the approaches. Proactive monitoring paired with interventions that defuse situations before problems escalate might end up saving moderators time and effort in the long run because they could have fewer situations to respond to overall, but in the interim it could increase moderators' workload significantly. Accordingly, as we observed, Discord's lack of support for mobile context menus, and the fallback of slash commands taking two to three times the amount of time to use, meant that moderators who largely engaged

while mobile used Chillbot less frequently. As a rough heuristic, if a moderator cannot take action within a few seconds, they may not engage such a tool.

To this point, automation of Chillbot might seem like a logical next step. Recent advances in language processing have made the detection of potentially-problematic conversations a more approachable task [37], though the results are still far from perfect. Two participants (E7A and E7D) discussed the possibility of incorporating some sort of AI-based detection into Chillbot during their exit interviews as potentially-interesting, though neither saw it as a requirement for the tool to be useful. Prior research has studied the consequences of failures in moderation actions likely taken by imperfect or biased algorithms [29], but no research to our knowledge has studied the impact of failures of moderation algorithms when the stakes are lower – e.g., a gentle warning with no associated punitive consequences. However, we chose to focus on a manually-operated bot in this case both because we felt that the likely nontrivial error rate of automating nudges would make it more difficult to evaluate the core design concept and because we aimed to develop a tool that supports moderators' ability to engage with their communities in a socially nuanced way. Also, as we discuss below, a key strength of automation is that it increases the potential frequency of usage, but frequency of usage may not be the best metric by which to assess non-removal based tools.

With this work and with future work in this space, it is important to avoid the mindset that, because tools can expand the space of things that moderators can do, moderators should spend more time and effort moderating if new tools are developed. As prior work has noted, many moderators already put an enormous amount of time and effort into moderating the spaces they care about, even to the point where moderation resembles a second job for them [40], and this labor should not be taken for granted.

6.3 Toward more equitable moderation

Platforms' frequent lack of communication before, during, and after moderation decisions contributes to the perception that they are arbitrary or even malicious, and this silence is in and of itself disempowering to those are disproportionately targeted by punitive actions but rarely invited to be involved in the decisions that govern their spaces [2]. The invisibility of these moderation decisions to all but those affected and the invisibility of the processes used to reach them exacerbate inequalities by laundering social biases through systems that cannot be reasoned with [14, 49]. While Chillbot is far from a comprehensive answer to these problems, its design fundamentally prioritizes communication over punishment in a way that could serve as one example of more visible, tangible moderation infrastructure in the future.

More broadly, a growing body of work shows that different groups have widely varying preferences for how problematic behaviors should be handled. For example, Schoenebeck et al. found that participants from non-US countries were more likely to favor alternative measures taken in response to harassment, including apologies, publicly revealing offenders' identities, or even monetary compensation [38]. Similarly, Xiao, Cheshire, and Salehi identified five major needs among adolescents in addressing online harm: sensemaking, support and validation, safety, retribution, and transformation [54], and other work from Schoenebeck et al. found that a majority of youth reported that they would like an apology under certain circumstances after being bullied or harassed [39]. These studies clearly show a strong need for more designs that explore approaches to content moderation that go beyond the punitive and carceral approaches that remain mainstream.

In order to encourage a broader array of designs for systems that support these alternative approaches to moderation, we must carefully reflect on our relationship with metrics in assessing trust and safety practices. For example, while the increasing prevalence of transparency reports is better than the alternative, these reports are typically centered around the statistics that are most straightforward to gather, analyze, and understand. The approaches to moderation highlighted

above are not necessarily those that are most amenable to maximization through volume; as Xiao, Jhaver, and Salehi note, moderation processes like those involved in restorative justice are labor intensive and can never feasibly occur at the same rates as automated punitive actions [55]. However, with help from features designed to better support practices like these, we argue that communication-rich, socially-engaged moderation practices are well within reach.

6.4 Limitations

Our evaluation focused on eleven servers ranging from dozens of members to more than two hundred thousand. While each server reached many individuals, we cannot fully generalize from the experiences of eleven servers: moderation cultures may vary from server to server, and responses to Chillbot might change if the tool were more broadly adopted. Our evaluation was not a randomized controlled trial, so we refrain from making any causal claims. In particular, we cannot test whether Chillbot caused behavioral changes in the servers or amongst those who were nudged. With moderators' collaboration, it would be possible in future work to craft a blind "noisy channel" experimental design that randomly delivers or does not deliver the nudge to its intended recipient. By analyzing user behavior before and after a nudge is intended to be sent, depending on whether the nudge actually was sent, we could test this outcome in future work, but this would likely require additional buy-in from moderators willing to test a tool that does not always work as expected.

Respecting moderators' and server members' privacy meant that our data was limited to interviews with moderators and analysis of the example use cases some moderators provided. Per moderators' requests, we did not personally join the servers in which the bot was present, so we did not observe the full context in which the bot was used—only the specific cases in which it was triggered. This also meant that feedback was filtered through the lens of the moderators in the servers, limiting the perspectives reported. Moreover, the moderators who used this tool were not given any training beyond what was needed to understand the tool's functionality; moderators with different skill levels in addressing interpersonal conflict might have different experiences with Chillbot. Future work could also recruit server members who received Chillbot messages to add further insight and complexity to our results.

The principles behind Chillbot can directly generalize to other online community platforms that host text conversation, allow bots, and contain a direct message ability. Such platforms include Discord, Slack, Reddit, and Facebook Groups. However, as we observed in our deployment with the lack of support for context menus in Discord's mobile client, small differences in the effort required to utilize the tool in moderators' workflow—channel factors [36]—make for substantial differences in adoption. So, attempts to translate Chillbot to platforms without sufficient API or interface support are likely to fail in practice.

We did not observe any purposefully malicious use of the tool, but all socio-technical systems can be co-opted. In theory, a moderator could use Chillbot to send pseudo-anonymous messages to a server member, since the messages are not credited to a specific moderator, though this use would appear in the logs generated by the tool. Chillbot does not have a meta-moderation governance structure: if a moderator adds nudges that harass, there is not yet a structured process for responding to complaints to the Chillbot team.

7 CONCLUSION

In this study, we sought to explore the potential for moderation tools to more fully cover scenarios where backchanneling is a more appropriate response than public removal. To do so, we iteratively designed and developed Chillbot, a tool for rapidly sending backchannel feedback, in coordination with Discord moderators. We then performed a field study of Chillbot on eleven servers ranging from 25 to roughly 240,000 members and found that the tool was well received and frequently

used on the majority of these servers, including voluntary usage past the end of the study period, indicating that it filled an ecological niche in their moderation practices.

Social interactions are complex and multifaceted, and yet our tools collapse these complex interactions onto a small set of available behaviors [1]. In creating moderation tools, our design imagination has lagged the social scientific understanding of the complex and multifaceted behaviors that moderators engage in. This paper represents an effort to help envision a broader, more socially-engaged future for moderation tools and to execute toward it.

ACKNOWLEDGMENTS

We would like to thank our participants for sharing their time and their communities to help advance this research. We would also like to thank the Computing Research Association and the Computing Community Consortium (CRA/CCC) for funding through the Computing Innovation Fellows program, as well as the Brown Institute for funding through the Magic Grant program and the Office of Naval Research for additional funding.

REFERENCES

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Carolina Are. 2023. An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society* 45, 4 (2023), 822–840. https://doi.org/10.1177/01634437221140531

[3] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1134–1145. https://doi.org/10.1145/3442381.3450122

[4] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[5] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 410 (Oct 2021), 25 pages. https://doi.org/10.1145/3479554

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

[7] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 174 (Nov. 2019), 30 pages. https://doi.org/10.1145/3359276

[8] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (Nov. 2018), 25 pages. https://doi.org/10.1145/3274301

[9] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (2015), 61–70. https://ojs.aaai.org/index.php/ICWSM/article/view/14583

[11] John W Creswell. 2013. *Qualitative Inquiry and Research Design: Choosing Among Five Traditions.* SAGE, Thousand Oaks, CA.

[12] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 40 (may 2020), 28 pages. https://doi.org/10.1145/3392845

[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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 019 (May 2020), 27 pages. https://doi.org/10.1145/3392822

[14] Kishonna L. Gray and Krysten Stein. 2021. "We 'said her name' and got zucked": Black Women Calling-out the Carceral Logics of Digital Platforms. *Gender & Society* 35, 4 (2021), 538–545. https://doi.org/10.1177/08912432211029393

[15] James Grimmelmann. 2015. The Virtues of Moderation. *Yale J.L. & Tech* 17 (2015), 42–109.

[16] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (2013), 664–688. https://doi.org/10.1177/0002764212469365

[17] 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

[18] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[19] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[20] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (Nov. 2019), 27 pages. https://doi.org/10.1145/3359252

[21] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-Led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. https://doi.org/10.1145/3491102.3517505

[22] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[23] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 55 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359157

[24] Prerna Juneja, Deepika Ramasubramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. In *Proceedings of the 21st International Conference on Supporting Group Work*. ACM, New York, NY, USA.

[25] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 477–487. https://doi.org/10.1609/icwsm.v16i1.19308

[26] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 44 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359146

[27] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. In *Building Successful Online Communities: Evidence-Based Social Design*, Robert Kraut and Paul Resnick (Eds.). MIT Press, Cambridge, MA, USA, Chapter 4, 125–177.

[28] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 596–606. https://ojs.aaai.org/index.php/ICWSM/article/view/19318

[29] Renkai Ma and Yubo Kou. 2022. "I'm Not Sure What Difference is between Their Content and Mine, Other than the Person Itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 425 (nov 2022), 28 pages. https://doi.org/10.1145/3555150

[30] 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)*. ACM, New York, NY, USA, Article 586, 13 pages. https://doi.org/10.1145/3173574.3174160

[31] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (may 2021), 14867–14875. https://ojs.aaai.org/index.php/AAAI/article/view/17745

[32] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789. https://doi.org/10.1073/pnas.1813486116

[33] Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technologies and Interaction* 5, 7 (Jun 2021), 34. https://doi.org/10.3390/mti5070034

[34] Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications* 166 (2021), 114120. https://doi.org/10.1016/j.eswa.2020.114120

[35] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. 2021. Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, Mohammad Shorif Uddin and Jagdish Chand Bansal (Eds.). Springer Singapore, Singapore, 457–468.

[36] Lee Ross and Richard E Nisbett. 2011. *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers, London, UK.

[37] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 370 (nov 2022), 27 pages. https://doi.org/10.1145/3555095

[38] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 485, 16 pages. https://doi.org/10.1145/3544548.3581020

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

[40] Joseph Seering, Brianna Dym, Geoff Kaufman, and Michael Bernstein. 2022. Pride and Professionalization in Volunteer Moderation: Lessons for Effective Platform-User Collaboration. *Journal of Online Trust and Safety* 1, 2 (Feb 2022). https://doi.org/10.54501/jots.v1i2.34

[41] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 157 (Nov. 2018), 29 pages. https://doi.org/10.1145/3274426

[42] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2020. Metaphors in moderation. *New Media & Society* 24, 3 (2020), 621–640. https://doi.org/10.1177/1461444820964968

[43] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[44] 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 (2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[45] Arushi Sharma, Anubha Kabra, and Minni Jain. 2022. Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. *Information Processing & Management* 59, 1 (2022), 102760. https://doi.org/10.1016/j.ipm.2021.102760

[46] C. Estelle Smith, Irfanul Alam, Chenhao Tan, Brian C. Keegan, and Anita L. Blanchard. 2022. The Impact of Governance Bots on Sense of Virtual Community: Development and Validation of the GOV-BOTs Scale. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 462 (nov 2022), 30 pages. https://doi.org/10.1145/3555563

[47] Jean Y. Song, Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. 2023. ModSandbox: Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 107, 20 pages. https://doi.org/10.1145/3544548.3581057

[48] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 1526–1543.

[49] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* 0, 0 (2022), 14614448221109804. https://doi.org/10.1177/14614448221109804

[50] Bruce W Tuckman and Mary Ann C Jensen. 1977. Stages of small-group development revisited. *Group & organization studies* 2, 4 (1977), 419–427.

[51] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages. https://doi.org/10.1145/3415238

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

[53] 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* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 160, 13 pages. https://doi.org/10.1145/3290605.3300390

[54] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 146, 15 pages. https://doi.org/10.1145/3491102.3517614

[55] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 83 (sep 2023), 36 pages. https://doi.org/10.1145/3603625

[56] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (nov 2018), 27 pages. https://doi.org/10.1145/3274465

INTERVIEW PROTOCOL FOR BRIEF FORMATIVE INTERVIEWS

(1) Please briefly describe the type or topic focus of the primary server that you moderate.
(2) How many active moderators does this server have?
(3) What are the main behavioral issues you encounter in moderating this server?
(4) When you are monitoring channels, are there common signs you've noticed that indicate that there may soon be a problem that would require moderator intervention? If so, what are these signs?
(5) When taking a moderation action (e.g., a warning, a message removal, a user ban), what are the factors that you take into account when deciding how severe a punishment should be?
(6) Do you consider what an offender's intent was when determining how to respond (e.g., whether they made an honest mistake vs intentionally broke a rule)?
(7) What strategies do you use to respond to (or proactively discourage) problematic behaviors other than traditional punishments? (i.e., "non-punitive" approaches to moderation that aren't timeouts/bans or content removal.)
(8) If, as a moderator, you ever have conversations with users about why something they did was problematic, what factors do you think determine how well that conversation will go?