"Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior

CHARLOTTE LAMBERT, University of Illinois Urbana-Champaign, USA
FREDERICK CHOI, University of Illinois Urbana-Champaign, USA
ESHWAR CHANDRASEKHARAN, University of Illinois Urbana-Champaign, USA

The role of a moderator is often characterized as solely punitive, however, moderators have the power to not only execute reactive and punitive actions but also create norms and support the values they want to see within their communities. One way moderators can proactively foster healthy communities is through positive reinforcement, but we do not currently know whether moderators on Reddit enforce their norms by providing positive feedback to desired contributions. To fill this gap in our knowledge, we surveyed 115 Reddit moderators to build two taxonomies: one for the content and behavior that actual moderators want to encourage and another taxonomy of actions moderators take to encourage desirable contributions. We found that prosocial behavior, engaging with other users, and staying within the topic and norms of the subreddit are the most frequent behaviors that moderators want to encourage. We also found that moderators are taking actions to encourage desirable contributions, specifically through built-in Reddit mechanisms (e.g., upvoting), replying to the contribution, and explicitly approving the contribution in the moderation queue. Furthermore, moderators reported taking these actions specifically to reinforce desirable behavior to the original poster and other community members, even though many of the actions are anonymous, so the recipients are unaware that they are receiving feedback from moderators. Importantly, some moderators who do not currently provide feedback do not object to the practice. Instead, they are discouraged by the lack of explicit tools for positive reinforcement and the fact that their fellow moderators are not currently engaging in methods for encouragement. We consider the taxonomy of actions moderators take, the reasons moderators are deterred from providing encouragement, and suggestions from the moderators themselves to discuss implications for designing tools to provide positive feedback.

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

Additional Key Words and Phrases: Social Computing; Positive Reinforcement; Moderation; Survey

ACM Reference Format:
Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. "Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 390 (November 2024), 33 pages. https://doi.org/10.1145/3686929

1 Introduction

Moderation is a necessary burden that most online platforms take on [29, 69]. Current social media platforms employ a combination of centralized content moderation [69] that relies heavily on large amounts of manual labor from paid [29] and volunteer [45, 47] moderators, alongside distributed approaches [54, 58, 72] like user reports [49, 50], and automated techniques like word-list-based filters [41, 44] and AI-assisted triaging [9, 12] of undesirable content for manual review. Studies

Authors' Contact Information: Charlotte Lambert, cjl8@illinois.edu, University of Illinois Urbana-Champaign, Urbana, Illinois, USA; Frederick Choi, fc20@illinois.edu, University of Illinois Urbana-Champaign, Urbana, Illinois, USA; Eshwar Chandrasekharan, eshwar@illinois.edu, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.

ACM 2573-0142/2024/11-ART390
https://doi.org/10.1145/3686929

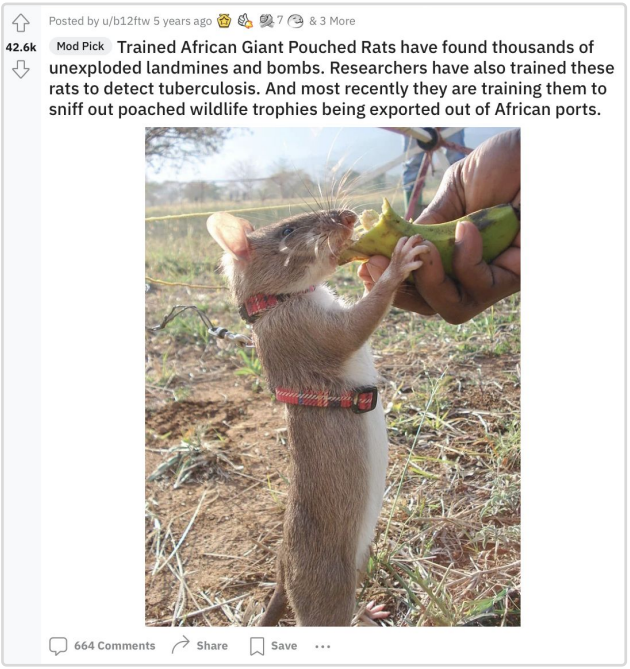

Fig. 1. An example post from the Reddit community r/Awwducational that illustrates several different forms of positive feedback, both from the moderation team and the community. The post was given the "Mod Pick" flair, gilded Reddit coins, given community awards, upvoted highly, and received many replies, all examples of actions our surveyed moderators reported taking to encourage desirable content.

related to measuring the efficacy of reactive methods [13, 14, 42, 68, 79, 82] have shown their effectiveness at curbing the spread of bad behavior.

However, reactive moderation does not prevent the harm already done by allowing the abusive behavior to occur in the first place [71], nor does it inherently educate users about community norms. Moderators using reactive methods are also constantly exposed to offensive behavior, often overloaded with work, and provided no compensation [56, 57], all while sifting through content damaging to their mental health [20, 80, 87]. Furthermore, removing bad content does not necessarily imply that a community is *healthy* or full of desirable behavior [46]. Thus, moderators need strategies to not only abate the spread of norm-violating behavior, but also encourage the presence of desirable behavior.

Additionally, moderation research often focuses on detecting and removing abusive behavior, which is only one of several moderator responsibilities. Grimmelmann [32] details three other types of disruptions which include congestion and cacophony, both of which result in good content getting lost in a larger pool of bad content. The combination of large volumes of content and algorithmically-curated feeds results in users having little control over the content they see. These feeds are often sorted using metrics opaque to users making it difficult for moderators and communities to control what content gets highlighted for users. Introducing ways for moderators to locate and encourage desirable behavior, as well as control what gets showcased to community members, will not only help address several forms of harm, but also improve the curation in algorithmic feeds.

1.1 Benefits of Positive Reinforcement in Offline and Online Contexts

In this paper, we aim to explore a holistic approach to moderation which supplements punitive techniques, such as content removals and bans, with the positive reinforcement of behavior that moderators and communities want to encourage. Many moderators themselves view their role as involving much more than just punishing users, some even see themselves as nurturers of healthy communities [74]. This proactive method has the potential to prevent unwanted content from ever being posted by fostering a healthy environment in which desired content is rewarded. Additionally, highlighting exemplary content has a much higher capacity to educate the community of the norms and desired behavior when compared to punitive methods.

Behavioral psychologists [77, 78] have demonstrated the benefits of positive reinforcement in offline contexts (e.g., classrooms [3, 65], workplaces [5, 7]) for many years. Within online contexts, researchers have explored how platforms such as Slashdot [53], Wikipedia [1, 36], and StackOverflow [2, 64] can encourage high-quality contributions. However, findings related to news-sharing and knowledge-creation sites such as these may not translate to other communities. The wide range of topics covered by communities on platforms like Reddit requires us to broaden current definitions of desirable behavior and explore community-centered definitions. Our research intends to do just that, specifically in the context of moderation, while also building an understanding of how community-specific desirable behavior is reinforced by moderators on Reddit.

1.2 A Case Study of Approaches to Encourage Desirable Behavior on Reddit

In July 2023, Reddit announced a shift towards a new system for awarding contributions, signifying a new emphasis being placed on the ways in which users on the platform can provide positive feedback. The announcement concluded with the following message:

> "In the coming months, we'll be sharing more about a new direction for awarding that allows redditors to empower one another and create more meaningful ways to reward high-quality contributions on Reddit."[1]

We see this as an opportunity to look more deeply into how positive feedback is employed in online communities, specifically with the focus of understanding positive reinforcement, using Reddit as a case study. Reddit's announcement does not specifically focus on the power of rewarding contributions to support community norms, however, moderators do play a large role in shaping their communities and are essential to understanding community dynamics. Thus, we center our research on how moderators employ positive feedback mechanisms to reinforce behavior, building off Seering et al. [75] and the design guidelines presented by Kiesler et al. [48]. Figure 1 provides an example Reddit post that has received the "Mod Pick" flair, a form positive feedback from a member of the moderation team of the community r/Awwducational, potentially signifying the presence of positive reinforcement.

1.3 Research Questions

In this paper, we examine current practices employed by moderators on Reddit to reinforce desirable behavior in their communities and uncover how moderators use positive feedback to encourage desired behavior. Specifically, we aim to answer the following research questions:

RQ1: What content and behavior do Reddit moderators want to encourage?
RQ2: How often do moderators provide positive feedback to encourage desirable behavior? What methods (and tools) do moderators currently employ to encourage desirable behavior?

[1] https://www.reddit.com/r/reddit/comments/14ytp7s/reworking_awarding_changes_to_awards_coins_and/

RQ3: If moderators do encourage desirable behavior, why do they provide feedback? If they do not encourage desirable behavior, why do they not provide feedback?
RQ4: How can we make it easier for moderators to provide positive feedback through design interventions?

1.4 Summary of Contributions

1.4.1 Methods: To answer these research questions, we sent out our survey to the moderators of 1.4K subreddits and received 115 complete responses. The survey responses were analyzed using open-coding techniques and used to produce two taxonomies.

1.4.2 Findings: We find that more than 95% of respondents currently take action to encourage desirable behavior at least some of the time, with more than 50% of them doing so frequently. Analyzing the survey responses revealed that prosocial behavior is the most common attribute that moderators want to encourage (46.1%). Nearly 25% of moderators cited civility, politeness, and/or respectfulness as qualities that should be encouraged. Additionally, 36% of moderators wanted to encourage high-quality content, and provided some specific attributes such as originality, utility, and creativity, that make a contribution high-quality. Roughly 50% of moderators upvote content they want to encourage. Although most responses are concrete actions available through the Reddit interface, certain moderators (3%) made extra efforts to compile exemplary contributions for users to browse (e.g., through a pinned post or the subreddit's sidebar).

We also uncovered motivations behind moderators' actions, finding that roughly 30% of moderators hope to encourage similar behavior from the authors who receive positive feedback. Similarly, 17% of moderators hoped to highlight exemplary behavior through their actions as a way to inform their communities about what the moderators want to encourage. A huge barrier to entry for many moderators was that they had to manually inspect large volumes of content to identify instances of highly desirable behavior. Another challenge when providing positive feedback is the lack of tools that can help moderators easily reward instances of desirable behavior.

1.4.3 Implications: From this research, it is clear that moderators on Reddit are already engaging in positive reinforcement, among other proactive strategies, and would largely benefit from support in the form of tools and systems explicitly for providing positive feedback to users. In particular, many current methods of reinforcement are anonymous and available to non-moderator users, thus obscuring the moderator's effort and influence by making it invisible to the recipient and broader community. These proactive methods have the potential to fill in the gaps left by reactive moderation methods, primarily by rewarding desirable content and encouraging more of it, while also educating newcomers and established community-members on the norms to increase retention. These proactive strategies may create healthier communities by improving moderator and user well-being. While our work is done in the context of Reddit, the findings and implications apply to content moderation more broadly. Reddit has shown an interest in embracing the idea of new ways to provide positive feedback to high-quality contributions, potentially indicating a shift in the importance of positive versus negative forms of feedback from the perspective of platform designers. We encourage moderators and designers of other platforms to reflect on whether strategies for reinforcement are currently utilized and to evaluate the feasibility of incorporating new moderation practices based on our results.

1.5 Ethical Considerations

In designing this study, we worked closely with the Institutional Review Board (IRB) at the first author's university to ensure that our participants were protected throughout recruitment, the survey, and data analysis. Participants consented to participation and all incomplete survey responses

were discarded to allow participants to implicitly revoke consent partway through. Additionally, we are not releasing any of our data to preserve participant privacy.

2 Background & Related Work

The research questions addressed in this work are largely motivated by principles in psychology. B. F. Skinner introduced the idea of positive reinforcement in his work studying reinforcement theory [25, 76–78], along with three other key concepts: negative reinforcement, positive punishment, and negative punishment. These concepts provide four different ways one can theoretically shape someone's behavior. Positive reinforcement specifically is the introduction of a stimulus to increase the likelihood that some behavior will continue in the future. Positive reinforcement strategies have been studied in several offline contexts. Education [3], workplaces [5, 7, 65], and parenting [4, 23, 81] are three areas in which positive reinforcement has been shown to motivate desirable behaviors. To determine if reinforcement is also effective in online settings, our research will *explore if and how moderators are currently employing techniques for reinforcement.*

Using these principles of behavioral psychology as background, this section details relevant prior work in human-computer interaction, specifically related to reactive and proactive forms of moderation, positive feedback, and community values and norms.

2.1 Drawbacks of Reactive Approaches to Moderation

Chandrasekharan et al. [13] explored how quarantining a Reddit community impacts the presence of toxic qualities such as misogyny and racism. While they found that quarantining a community makes recruiting new members more difficult, the existing community members did not adjust the behaviors that incited the quarantine in the first place. We see that this particular reactive method of moderation does not seem to have the capacity to encourage healthy behaviors in a community.

While quarantines focus on community-level moderation, content removals are an extremely popular reactive moderation action that focus on individual users. Their widespread nature has prompted research into the effect of content removals on compliance and non-compliance in communities. Srinivasan et al. [79] found that content removals reduce the rate of subsequent rule violations, but do not increase positive outcomes, such as engagement and community approval.

Jhaver et al. [43] investigated the effects of Reddit removal explanations on an author's odds of being removed in the future. Without removal explanations, content removal implicitly informs a poster of the community's norms by teaching them what they should not post. When employed, removal explanations provide a justification for the preceding punishment that the entire community can see, helping make the moderator action more explicit and informative.

While the findings of this research show the power of providing removal explanations, it is important to remember that this signal requires additional effort on the part of moderators, something many of them do not have the time for. As a result, Jhaver et al. [43] found that only 0.6% of the subreddits in their dataset provided removal explanations, despite the fact that removals themselves are often not self-explanatory, evidenced by the finding that a large percentage of users who had a post removed did not know why [40]. The effort required to effectively communicate norms through removal explanations is one major drawback to this form of reactive moderation. This drawback, along with the inefficacy of other reactive methods at promoting positive behavior, motivate our desire to *shift focus towards more proactive approaches, which have the potential to more easily communicate norms* to the original poster along with any other community member, all through positive action.

2.2 Proactive Forms of Moderation

The idea of proactively moderating a community is not new. Many researchers interpret this concept as preemptively detecting problematic behavior to prevent a conversation from devolving. As a result, researchers have previously investigated strategies for identifying conversations likely to be derailed to preemptively moderate at-risk conversations [73, 89]. Choi et al. [18] focused on community norms from a design perspective by creating a human-AI system to be integrated into Discord to assist moderators. The authors specifically provided conversational metrics to help moderators quickly digest large amounts of content and proactively identify potential problems. Similarly, Yen et al. [88] focused on the context of live-streaming and showed that providing users with a visualization of the negativity in a conversation may actually encourage prosocial contributions and other desirable outcomes. There has also been work examining ways to proactively inform users of other historically toxic users through publicly visible signals [39]. This research highlights how communities can improve user experience and safety through transparency and the use of interface signals.

At the community level, Habib et al. [34] proactively identified communities with the potential of being banned based on their predicted actions. These approaches moderators to be aware of areas that may need more attention. However, moderators then must either preemptively restrict posting in at-risk conversations, or devote time to increased monitoring so norm-violating content can be removed as quickly as it is posted. This could lead to harmless conversations being cut off early, and increased moderator workload with no guarantee that users will not be exposed to harmful content.

Matias [59] examined the effect of publicly announcing the rules in a large science-discussion Reddit community and found that proactively displaying rules was helpful for both newcomer norm acquisition and newcomer participation. Other researchers have given some more agency to the users by developing a tool to warn users if their drafted contribution is predicted to increase tensions in the conversation [16]. This is a method of moderation that puts the onus of making good decisions onto the users and gives them extra information to inform those decisions. In contrast to other proactive moderation methods, this user-focused tool may be able to accomplish the task without additional moderator intervention. However, there is still no understanding of what type of content is encouraged since this tool will only predict potentially inflammatory content.

In an effort to understand the positive attributes of conversations, Bao et al. [6] worked toward identifying signals of successful conversations by developing metrics for measuring prosocial behavior. Lambert et al. [52] added to this work by identifying potential factors that correlate with prosocial conversational outcomes following content removals on Reddit. Both of these contributions help inform our understanding of prosocial behavior on Reddit, but it is not clear whether prosocial behavior is actually a sign of success from the perspective of real users and moderators.

Based on this prior research, we see a gap in knowledge regarding what type of content and behavior is considered desirable in Reddit communities. *We begin to address this missing piece by constructing a taxonomy of desirable attributes according to moderators.*

2.3 Reinforcement Through Feedback

There is prior research into specific forms of feedback that may serve as reinforcement in online contexts. Badges, for example, are a common mechanism platforms implement to motivate their users. Wang and Diakopoulos [83] found that being exposed to a New York Times Pick badge, both as a first-time recipient and bystander, was positively correlated with the quality of a user's future posts. Similarly, Anderson et al. [2] found that providing badges to users can be utilized strategically

to motivate certain behaviors on Stack Overflow, while Papoutsoglou et al. [64] propose additional ways to model Stack Overflow's use of badges as a gamification mechanism. Hamari et al. [37] and Sailer et al. [70] studied the effect of gamification in several contexts and found that gamifying online platforms by adding certain motivating mechanisms tied to user actions may be able to provide positive effects, highly dependent on the context and the game design elements.

Reputation systems [66, 67] and game theory [27, 28] are two more concepts that have been applied to research related to encouraging specific outcomes from contributors in online communities. For example, Adler and De Alfaro [1] developed a reputation system for Wikipedia aiming to motivate users to contribute more accurate content with the promise of additional access and power to contribute.

On a community level, Cunha et al. [21] found that users on a weight loss subreddit reported higher decreases in their weight when they received more comments on their posts from the community, showing the importance of receiving online feedback for real-world outcomes. Gurjar et al. [33] explored the effects of increased popularity and revealed that users experiencing increased attention will match their future content to what experienced the popularity. This is another example of community feedback reinforcing certain behavior.

These works provide supporting evidence that some reinforcing practices are effective in online settings, but demonstrates the gap in research related to if and how moderators specifically utilize reinforcement practices. *We supplement this work by expanding our understanding of what specific techniques for positive feedback are employed by moderators who want to encourage desired behaviors.*

2.4 Learning Community Values and Norms

Previously, researchers have investigated moderator and user perspectives of what their communities value [85, 86]. This prior research involved surveying Reddit moderators and users to understand community values across the platform. With their survey responses, they built and explored a taxonomy capturing the idea of Reddit community values, a concept similar to what we ask of our moderator participants in our survey. However, the values found in prior research are primarily outcomes for communities, such as diversity, trust, and size. Our research takes a bottom-up approach, asking not only what attributes moderators find desirable about contributions, but also identifying specific actions taken by moderators to encourage desired behavior and potentially help their communities reach the outcomes identified by Weld et al. [86].

Halfaker and Geiger [35] utilized human labels to create models of quality related to Wikipedia edits that can be applied by users in several ways. This is an example of work focusing the definition of "high-quality" on the specific context and giving users more agency to define desirability themselves. We also aim to understand what makes content or behavior desirable in a context-sensitive way through our construction of a taxonomy.

From a more quantitative perspective, Chandrasekharan et al. [15] uncovered macro, meso, and micro norms within Reddit communities by studying removed comments to understand what communities do not want, leaving further questions about what communities do want. *Our research helps fill this gap in our understanding by learning about the specific content and behavior moderators actually want to encourage on Reddit.*

3 Methods

Next, we describe our survey instrument, recruitment process (visualized in Figure 2), and approach for qualitative coding of survey responses. All described recruitment and data collection procedures were approved by the IRB at the first author's institution.

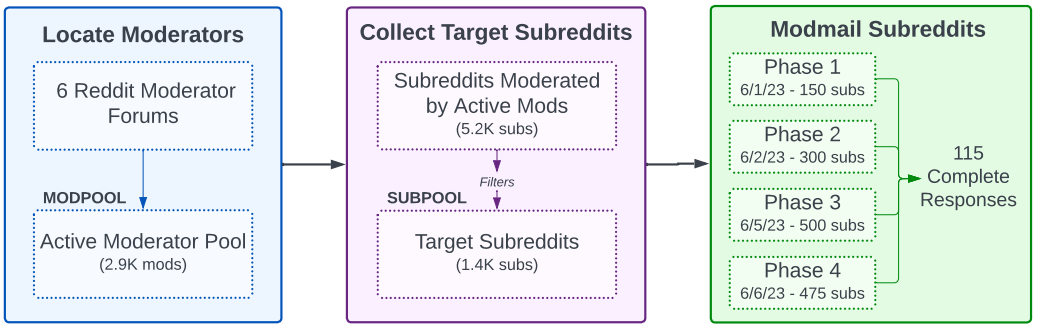

Fig. 2. Visualization of the survey recruitment pipeline from forming our MODPOOL and SUBPOOL to sending out the survey links. The rightmost box reports the number of complete responses received. Note that "moderator" is shortened to "mod" and "subreddit" to "sub".

3.1 Survey Instrument

We constructed a survey to answer our four research questions. This section summarizes the survey instrument which is included in Appendix A. To ensure that our survey was easy to understand and complete, we piloted the survey with 7 students at the first author's institution who were not involved with the research. The verbal and written feedback from these students was used to iteratively improve the phrasing and overall structure of the survey.

The survey was structured into three parts. First, we asked participants multiple choice questions to collect basic information about their activity levels on Reddit generally to give us an idea of the participant's familiarity with the platform. Second, we asked participants open-ended questions related to how they provide positive feedback in their community. This involved asking participants what behavior they actually want to encourage in their community (RQ 1).

We then followed up with another set of open-ended questions to reveal how our participants react when they see content they want to encourage (RQ 2). This section also included questions to learn the moderators' motivation for taking those actions (RQ 3).

The third section of the survey consisted of open-ended questions asking participants to explain how they currently locate posts they want to encourage and to provide any suggestions or ideas for ways to make locating or reinforcing such posts easier (RQ 4).

3.1.1 Avoiding biases. Müller et al. [60] define five types of questionnaire biases that we sought to avoid in our survey design. To avoid the satisficing bias, we made sure our multiple choice questions were simple and did not require too much cognitive exertion to answer. We avoided including options for participants to express "no opinion," "don't know," and other similar responses which may present participants with an easy default response. Additionally, we presented the survey to our participants in the context of improving their communities to increase their motivation to complete the survey thoroughly.

Our survey design also intended to reduce the likelihood that participants exhibited acquiescence bias by avoiding questions asking participants to agree or disagree with a statement.

To discourage the social desirability bias, we ensured that participants knew they were completely anonymous and did not have to provide us with their Reddit username.

Additionally, we randomized multiple-choice response options when possible to address response order bias. Early questions in the survey were easier to answer, while later questions were more

Table 1. Descriptive information about the identified moderator forums which were used for recruitment. The description was manually retrieved from each subreddit's page. The subscriber counts and descriptions were accurate as of 5/10/23.

Subreddit	# Subscribers	Description
r/ModSupport	76,299	An official community to provide a point of contact for moderators to discuss issues with Reddit admins, mostly related to mod tools.
r/modhelp	124,078	Have questions about moderating your subreddit? We might be able to help!
r/moderator	512	An unofficial community for moderators, by moderators!
r/modnews	223,324	An official community for announcements from Reddit, Inc. pertaining to moderation.
r/AskModerators	6,298	A place for users to ask moderators questions and have them answered.
r/modclub	5,819	A subreddit for moderators to discuss moderation things.

difficult and required more thought. This organization, along with grouping together related questions, helped us avoid question-order bias.

The organization of our survey sought to address the five aforementioned questionnaire biases. As advised by Müller et al. [60], we also avoided leading and double-barrelled questions.

3.2 Recruitment

Moderators on Reddit define their communities' rules and enforce them, along with providing implicit feedback through punitive actions like content removals. However, taking actions to encourage behavior is not as frequently discussed, nor is it as easy to spot across communities the way content removal is. Because of their position of authority within subreddits and their inside knowledge of whether their teams provide encouragement, we focused this work on a survey of Reddit moderators.

3.2.1 Locating moderators. Our goal in recruitment was to find active moderators from communities of varying sizes across Reddit. However, there is no publicly-accessible way to assess the activity and engagement levels of moderators. Additionally, while related work often focuses analyses on the most popular communities on Reddit [15, 17, 52, 79], we hypothesize that communities with less content to moderate may have more time and capacity to engage with methods of positive reinforcement. Thus, we did not focus recruitment on any predetermined set of communities. Instead, we used Reddit moderator forums to identify particularly active moderators.

The specific moderator forums we explored were located first through a manual Google search, which identified two subreddits intended for moderators: r/AskModerators and r/modclub. We then looked through the communities and located all other public communities linked in the sidebar which explicitly targeted moderators in their descriptions. This process was repeated until there were no more linked communities. Subreddits targeting new moderators (e.g., r/needamod) were excluded to avoid getting only new moderators. A summary table of these moderator-specific subreddits can be found in Table 1.

We assert that participation in moderator forums is an indicator of how active moderators are within their respective communities. We use the Reddit API, PRAW, [2] to identify all the moderators who posted in at least one of the six moderator-specific subreddits between January 1, 2023 and

[2] https://praw.readthedocs.io/en/stable/

May 30, 2023. This collection is limited by the API and only contains at most 1,000 posts from each subreddit. We refer to this collection of active moderators as the MODPOOL.

For each moderator in MODPOOL, we traced back the moderation teams that the moderator was part of. Although the members of these moderation teams may not all be as active as the members of the MODPOOL, we wanted to maximize the chances that we received multiple responses from the same moderation team which would allow for additional analyses of agreement.

We used PRAW to monitor the 21,103 subreddits moderated by at least one member of MODPOOL for one week, May 23, 2023 to May 30, 2023. Each subreddit that saw at least 7 posts during our monitoring was added to our list of moderation teams to recruit, called SUBPOOL, which contained 5,212 subreddits before we filtered based on the following inclusion criteria:

(1) **Safe for work:** we excluded all communities marked NSFW.
(2) **One month old**: all communities less than one month old were excluded.
(3) **At least 5 moderators:** communities with fewer than 5 moderators were excluded to increase the chances that we would receive multiple responses from each community.
(4) **At least 10 comments per post**: we required communities to have an average of at least 10 comments per post in the month leading up to recruitment to ensure that participants had sufficient experience moderating comments.
(5) **At least 1,000 subscribers:** we pruned extremely small communities by requiring all subreddits in SUBPOOL to have at least 1,000 subscribers.
(6) **English-speaking**: we manually confirmed that the description of each subreddit and the title of its last post are both in English.

3.2.2 Sending the survey. Upon filtering, we were left with 1,425 subreddits in SUBPOOL. We utilized Reddit's modmail functionality to send a unique survey link to the moderation teams of each filtered subreddit to reduce the chances that moderators would take our survey more than once. Additionally, it was clear to moderators which community we were interested in hearing about since they received their survey link directly through a subreddit's modmail system. This allowed us to reach a large group of potential participants and avoid publicizing survey links to non-moderators. Through modmails to 1,425 subreddits, we distributed the survey to 14,594 moderators.

As shown in Figure 2, we sent out the survey in four different phases. For each phase, we used the PRAW API to send out one link per 30 seconds to randomly-selected subreddits from SUBPOOL to stay within the rate limit for sending modmails. Users who responded to our modmail with questions about privacy or IRB approval were sent links to the consent form and our IRB approval from the university and were advised to contact our university's IRB department with further questions. Other users who asked whether we wanted multiple moderators to fill out the survey or just one per subreddit were sent a message clarifying that we would like to hear from as many moderators as possible.

Compensation for the survey was done through a raffle, similar to past surveys of moderators [49, 86], to increase the number of participants we were able to recruit. Participants were given the option to enter into a lottery in which we raffled off one $20 Amazon gift card for every 50 participants who entered.

3.3 Qualitative Coding

After all survey responses were collected, the first and second author independently went through all responses for every survey question and assigned labels using an inductive, open coding approach. This involved reading through the responses, building up tentative codes, and then going back through the responses and labeling each one, iteratively adding new codes when needed. Each response was not limited to a single code per question. Instead, we assigned codes for each distinct

Fig. 3. These three plots visualize distributions related to the moderators who participated in our survey. Plot (a) shows a histogram of our participants' age on the platform (i.e., the number of years since they created their Reddit account). Plot (b) visualizes the number of years our participants have been moderators of any subreddit. Finally, plot (c) shows a distribution of posting or commenting frequency for the surveyed moderators in their moderated subreddit.

thought mentioned by a participant, similar to an approach taken in prior work [86]. Thus, the results of our qualitative coding analysis consist of more assigned codes than there are participants. After both authors completed the open coding process, they came together and generated a codebook by combining their independent labels. Conflicts were discussed until both authors agreed on the final assigned code. The authors then grouped all codes into logical categories based on similarity. This process resulted in hierarchical codebooks for each survey question.

To validate the codebooks we use to construct our two taxonomies, we drew upon approaches used by Fiesler et al. [26] and Weld et al. [86]. Specifically, we recruited two additional labelers to label one third of our survey responses, selected at random ($n = 38$). After this step, all labelers came together to make sure any disagreements in the labels were because of subjective judgments and not because of misunderstandings. Inter-rater reliability was measured using Krippendorff's α, which allows for multiple labelers and multiple labels for each response.

Our first taxonomy categorizing what moderators want to encourage obtained a Krippendorff's α of 0.710, meaning our taxonomy can be used to draw "tentative" conclusions when applied by other researchers [51]. Through a manual inspection, we observe that the majority of the disagreements are the result of some labelers adding additional codes to a response that other labelers did not include. However, measuring agreement on the higher-level categories of assigned labels yields a Krippendorff's α of 0.801, indicating reliability in our labels. Additionally, in all but three responses, there was at least one overlapping code across all labelers. In those three instances, the distinct labels assigned by each labeler fall under the same overarching category, indicating some agreement.

The second taxonomy explaining what actions moderators take when they want to encourage content or behavior obtained a Krippendorff's α of 0.945, demonstrating reliability in our labeling [51]. We believe this disparity in α values between the two taxonomies is because the second taxonomy consists of actions used across the platform, while the first taxonomy requires more subjective interpretation because of the more open-ended, community-specific nature of the responses. We believe both of our taxonomies can be utilized by other researchers, though we recognize that labeling responses using the first taxonomy is a more difficult task.

The remainder of the survey responses were labeled by the first and second author using the finalized codebooks.

Table 2. Summary statistics about the 117 subreddits our participants reported moderating. For each statistic, we report the minimum and maximum values, along with the median and the first and third quartiles.

	Min.	25%	50%	75%	Max.
# Subscribers	5.6K	67.1K	191.1K	456.5K	49.6M
# Moderators	5	8	11	16	53
Subreddit Age (years)	1.0	7.8	11.8	13.7	15.4

4 Results

Next, we detail findings from our qualitative analysis. Results are summarized in Tables 3-6.

4.1 Survey Responses

During our survey recruitment phase, we contacted the moderation teams of 1,425 subreddits which consist of a total of 14.6K moderators. 286 moderators from 178 subreddits started the survey and 115 moderators from 107 unique subreddits were eligible, completed the survey, and gave valid responses. We note that while we aimed to collect enough responses from each moderation team to understand the dynamics of entire teams, we did not receive enough responses to draw conclusions about moderation teams as a whole. Instead, we focus on the behavior of individual moderators.

We visualize some summary statistics about the participants in Figure 3. We see from the statistics that our participants have a range of experience both on the platform generally and as moderators of their reported subreddits. Most of our moderators have at least 3 years of moderation experience, though we still hear from a significant number of newer moderators. The vast majority of our respondents have had their Reddit accounts for at least 5 years, indicating an understanding of the platform from a moderator and non-moderator perspective. Additionally, most of our participants frequently post in their subreddits. These plots show that our responses are not an exhaustive sample of all the communities on Reddit, however, we can use this data to understand existing trends and motivate future research that can generalize more broadly.

Table 2 reports statistics about the subreddits moderated by our participants. These statistics demonstrate that our data focuses on moderators of well-established subreddits with smaller moderation teams, but is still able to capture some data from moderators of larger subreddits.

4.2 What do Reddit moderators find desirable?

To answer our first research question, we asked our participants the following:

> *"What kinds of content and behavior would you want to encourage in [subreddit] as a moderator?"*

Note that we use "[subreddit]" to indicate the place in a survey question where a participant's moderated subreddit was dynamically inserted using Qualtrics. This question aimed to understand what content our surveyed moderators want to see more of in their community. Our taxonomy of attributes of desirable content and behavior is reported in Table 3.

4.2.1 Prosocial behavior. The attribute mentioned by the most moderators (46%) related to behavior was being prosocial. Some responses mentioned being prosocial generally, while 24% specified that civility, politeness, and/or respect were a priority. For example, participant P33 said they want to encourage "*Kindness, helpfulness, information sharing.*"

Especially positive or supportive contributions were also mentioned frequently as encouraged by our respondents. This overall pattern of valuing prosocial behavior indicates its importance in many communities.

Table 3. Taxonomy of attributes that surveyed moderators consider desirable, based on open-coding the survey responses. Categories are ordered based on frequency in responses.

Attribute	# Mods	Excerpt From Moderator Response
Prosociality		
Civility, Politeness, Respect	28	"we encourage folks to be civil and respectful."
Prosocial Behavior	15	"Kindness, helpfulness, information sharing"
Positivity	9	"leaving positive comments and feedback"
Support	5	"Being supportive of other users..."
Creating a Safe Environment	3	"safety and anonymity."
Quality		
Quality	13	"Submitting more high quality posts"
Originality, Novelty	12	"Original content submission"
Utility, Helpfulness	12	"Helpful content..."
Humor or Fun	6	"The best content is usually funny memes"
Creativity	5	"Creativity, positive engagement and interactions"
Factual Correctness, Reliability	4	"Scientifically-backed opinions"
Participation		
Engaging with Other Users	15	"Making thoughtful comments, friendly interactions"
Inciting Discussion	10	"Content that inspires discussion..."
Supporting Newcomers	4	"[Answering] the beginner questions we get..."
General Participation	2	"More submissions"
Community-Specific Content		
Relevance to Subreddit	25	"Good posts that are on topic."
Personal Experiences	6	"Progress updates/success stories/follow ups"
Norm-Abidance		
Norm-Abiding	11	"Anything that follows the rules..."
Specific Content Formats		
Format or Type of Contribution	7	"Content like fanart"
Asking Questions	1	"Anything about chihuahuas including questions..."
External Participation		
Actions Outside Subreddit	4	"real world community action and involvement..."

4.2.2 General measures of quality. The next most common attribute mentioned in our responses related generally to the quality of content. Roughly 37% of moderators pointed at quality as a major deciding factor of whether they would want to encourage a contribution. For instance, participant P19 wanted to encourage their community members "*submitting more high quality posts rather than low effort memes or screenshots with little context.*" More specific facets of quality include originality, usefulness/helpfulness, fun, creativity, and factual accuracy.

4.2.3 Participation. Roughly 23% of moderators explained the importance of participation and discussion in their responses. Some moderators were not very picky about the type of content or behavior, but instead emphasized that participation generally was important. Others were interested in whether a contribution was able to provoke discussion or generate genuine engagement from the community as a way to decide if it should be encouraged. Participant P4, for example, specifically wanted to encourage "*content that inspires discussion.*" Several moderators also mentioned a specific interest in posts that welcome newcomers to the online (or offline) community, such as participant P16 who valued "*resources/posts that answer many of the beginner questions we get.*" Moderators who were interested in these attributes seemed to value community growth. This means encouraging

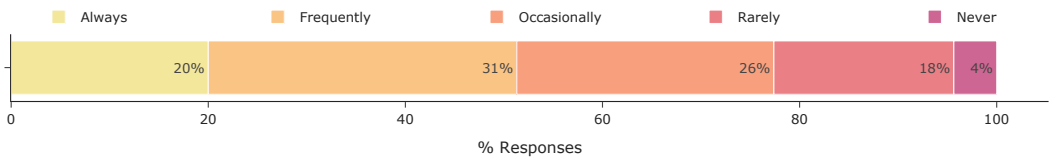

Fig. 4. This figure visualizes a breakdown of participant Likert responses to the question *"How often do you take actions to encourage such behavior in [subreddit]?"*

posts that inspire more activity, and those that help create a welcoming environment for newcomers so that they will return to the community.

4.2.4 Community-specific content. More specific than participation generally, some moderators wanted to encourage on-topic content. About 22% of surveyed moderators emphasized the importance of relevance, and another 5% favored posts that discussed personal experiences related to the subreddit topic. These attributes vary the most between communities since topic varies greatly between subreddits.

4.2.5 Adherence to norms. Other moderators reported a more hands-off approach to moderation and did not explain the type of content that should be encouraged. Nearly 10% of moderators explained that the only feature of content that should be encouraged is whether it adheres to the rules and norms of the community. Participant P110 explained that they encourage "*anything that follows the rules, we don't discriminate on any grounds.*" Additionally, more than 6% of the surveyed moderators responded with a focus on punitive moderation techniques. These moderators rejected the idea of encouraging good behavior from a moderation standpoint and instead explained the type of content they remove or discourage.

4.3 How do moderators encourage desirable behavior?

With this understanding of what participants are interested in encouraging, we sought to learn whether moderators actually take actions to encourage those behaviors. Figure 4 visualizes the frequency with which participants reported taking action to encourage behavior. This graph shows that more than 95% of participants reported encouraging behavior in their subreddit at least some of the time, while only 4% of participants reported never taking such actions. This demonstrates that positive feedback and the effort to reinforce desirable behavior is something already employed by many moderators.

For participants who reported encouraging behavior at least "Rarely," we followed up with questions to learn what actions moderators take to do that encouragement. We asked participants:

> *"What actions do you take as a moderator when you see comments or posts you want to encourage in [subreddit]?"*

The codes assigned to the responses to this question comprise our second taxonomy, and are reported in Table 4.

4.3.1 Feedback through the interface. We found that most of the actions taken by moderators when they see something they want to encourage are concrete actions in the interface. More than 70% of participants reported taking action through one of these concrete features, with nearly half reporting using the upvote feature to encourage. As shown in Table 4, we notice that the most popular methods for reinforcement fall into the category of anonymous interface signals. This means 56% of our surveyed moderators were providing feedback that is indistinguishable from feedback other community-members are providing. As a result, the effort made by a large proportion

Table 4. Taxonomy of actions moderators take to encourage behavior and content they find desirable. Categories are ordered based on frequency in responses.

Action	# Mods	Excerpt From Moderator Response
Anonymous Interface Signals		
Upvote	57	"I usually just upvote them and move on"
Award	17	"I give them awards..."
Coins	1	"maybe even gild them."
Respond		
Comment	31	"I may comment on the post"
Respond	17	"Respond with my own thoughts"
Respond (As Mod)	5	"I comment back to them distinguished as a mod."
Respond (Not as Mod)	3	"praise as a user, nothing as a moderator"
Direct Message	1	"Perhaps some thanks in a direct message..."
Modmail	1	"direct contact via modmail"
Moderator-Only Interface Signals		
Sticky/Pin	12	"We sticky good posts..."
Flair	7	"Flair them having Mod Favorite"
Approve in Mod Queue	6	"Approve the post/comment if it is flagged..."
Consolidate	4	"I might link them on the sidebar"
Punish		
Warning	4	"Giving a warning to be civil..."
Deletion	1	"delete posts that encourage flame wars"
Bans	1	"bans if non-compliant"

of moderators is not visible as moderator feedback by the recipients nor by the community generally. Only about 21% of surveyed moderators choose to give positive feedback through interface signals that are only available to moderators, such as stickying posts and adding flairs.

Nearly 50% of the moderators also reported that they respond to the post or comment in some way. Some respondents specified that they reply with a comment, send a direct message, or send a message through modmail. One important distinction made by some participants was whether they responded as a moderator (i.e., distinguished) or as a user. Some just want to contribute to the conversation, while others want to explicitly give positive feedback from a position of influence.

4.3.2 Actions through the moderator queue. About 5% of participants mentioned taking action from within the moderator queue (modqueue), Reddit's interface for moderators to review contributions. Participant P61 wrote:

> "*Hit the 'approve' and 'ignore reports' button to notify the other moderators that the post is what we would like to see and it does not break any rules.*"

This indicates that some moderators use approvals in the modqueue to communicate implicit approval for a post or comment, meaning other moderators would not have to review it. This action is less visible from the user side, so it is not exactly considered feedback. However, it is still an interesting method of supporting a good contribution. Five of the six moderators who reported encouraging behavior in this way came from communities with more than the median number of subscribers in our sample, and one response was from a moderator of our largest community with nearly 50M subscribers. We believe this indicates that these moderators do not have the time to encourage behavior in ways outside of their expected moderator responsibilities.

Table 5. All intentions or motivations participants have for taking action when they see something they want to encourage, revealed through open-coding. Categories are ordered based on frequency in responses.

Intention/Motivation	# Mods	Excerpt From Moderator Response
Reinforce		
Encourage Similar Behavior	35	"To encourage that kind of behavior"
Highlight Exemplary Behavior	19	"Show the community that this is a values post."
Provide Positive Feedback		
Approve, Praise, Be Positive	26	"Feedback that the person is on the right track"
Be Supportive or Welcoming	13	"make the user feel welcome"
Show Gratitude or Appreciation	10	"Let people know that their posts are appreciated"
Express Agreement	3	"Generally post in agreement or say yes, this."
Participate		
Boost Visibility	21	"To make the best content more visible and at the top"
Encourage Participation	11	"encourage engagement on the posts/comments"
Engage with Content	11	"Participate in a successful and healthy community."
Encourage Crossposting	2	"encourage people to share it in other subreddits."
Moderate		
Improve Mod Reputation	4	"Better optics of the mods on the sub"
Coordinate with Mod Team	3	"let other moderators know the comment [is] validated"
Recruit Future Moderators	1	"note [...] future additions to the moderation team"
Reasons for Inaction		
Not Done	2	"I dont really see the other mods doing it"
Lack of Available Tools	2	"We have no options to promote these comments."
Futile, Won't do Anything	1	"It often times feels futile."
Unnecessary	1	"I don't need to. We're doing it automatically."

4.3.3 Compiling valued posts. Another approach consisted of consolidating high-quality posts in some designated location. Four moderators indicated highlighting posts by including links in the sidebar of their subreddit's front page, by compiling a set of links in a pinned post, or some other method for consolidating posts for their community to browse. Participant P59 wrote that to encourage valued posts, they "*might link them on the sidebar.*"

4.3.4 Not Taking Actions to Encourage. Similar to the responses to the previous question, some moderators responded to this question with a negative perspective as opposed to discussing the ways they encourage positive behavior. This included two moderators who felt that encouraging behavior was simply not part of their job, as the community should self-regulate through upvotes and downvotes to highlight or hide certain types of content. Another four respondents responded with the actions they take to discourage bad behavior, such as warning, removing content, and banning users.

4.4 Why do moderators choose to provide positive feedback?

With this taxonomy of actions moderators use to encourage particular posts, we ask further questions to reveal the implicit motivation behind those actions by asking the following question:

"What do you intend to achieve through these actions?"

Table 5 shows the codes generated by analyzing the responses to this question. This table is also supplemented with some responses from the previous section, since some respondents provided not only the action they take, but the reason why. We note that 6 moderators responded to this question

with a focus on mitigating negative behavior in their communities, as opposed to describing how they interact with positive behavior in their communities.

4.4.1 Positive reinforcement. The most popular response to this question related to reinforcement, with 40% of moderators indicating that they take actions in order to encourage similar behavior from a contribution's author and/or their community as a whole.

> "*Positive reinforcement helps breed positive behavior. We want that.*" (P100)
>
> "*Exemplifying to users what kind of content is appreciated, thus reinforcing it.*" (P84)
>
> "*Model good behavior, encourage others to do the same.*" (P87)

With these responses, participants revealed that through the use of positive feedback, they intend to encourage similar behavior out of authors and bystanders, and/or highlight exemplary behavior for all community members to see.

4.4.2 Provide positive feedback. Aside from reinforcement, 37% of moderators are motivated to take action in response to desirable content simply to provide feedback. Some moderators want to show approval, give praise, or generally spread positivity, without the explicit goal of reinforcement. For example, participant P23 wanted to "*help people feel good about acting in good faith and being polite.*" Other moderators mentioned providing feedback to support their users or to make the community more welcoming. These moderators seem less focused on the future behavior of users in the community, and more interested in fostering a welcoming environment in the moment.

4.4.3 Boost visibility and encourage participation. About a third of responses related to encouraging participation. About 18% of moderators thought their actions could help boost the visibility of a post or comment, and potentially encourage more engagement with the contribution. Participant P86, for instance, wrote that "*upvoting and stickying quality posts gets them more visibility and a better chance to engage with other users over the content.*" Moderators saw this increased participation as a way to give exposure to certain content, benefit the author by giving them more attention, and/or improve the overall quality of their community by making high-quality posts more visible. This is an example of moderators attempting to work with the given signals to take control of their community's algorithmically-curated feed.

4.4.4 Moderator reputation. Interestingly, 6% of moderators reported providing positive feedback to improve public perception of the moderation team. Some participants specifically stated that they wanted to show their community that moderators care about interacting with users and value good content to improve the user-moderator relationship within the community. For example, Participant P110 wrote:

> "*Participating in the sub as a mod seems to increase relations with the users. I've seen it work personally.*"

For the participants who reported never taking action in response to content they want to encourage, we asked:

> *"Why do you not take actions to encourage behavior?"*

Though very few participants reported never taking action, their reasoning for not encouraging content revealed interesting relationships between positive feedback and moderation. These responses are also included in Table 5.

4.4.5 Unavailability of tools to take actions. The most frequent reason moderators do not take action was that there were no tools available for identifying desirable behavior and providing encouraging feedback. Obviously, we see from the other survey responses that some moderators get around the lack of moderation tools to provide positive feedback by adapting existing mechanisms,

however, it is important to note that not all moderators will take it upon themselves to do so, and will likely only provide positive feedback if concrete mechanisms and tools exist.

4.4.6 Lack of precedent. Two moderators said that they were relatively new additions to their moderation teams, like participant P98, who says, "*Im also the newest mod on the team, so I dont want to rock the boat.*" This demonstrates that some new moderators follow trends set by the more senior moderators. From responses like this, it seems like it is important to establish a process for encouraging behavior for more moderators to adopt the practice.

4.4.7 Concerns about extra effort and efficacy. Moderators were unsure if the extra effort to reward desirable behavior was necessary. Two moderators explicitly stated that they avoid giving positive feedback because they believe this extra effort was futile or unnecessary. This concern emphasizes the need to explore the effect of reinforcing desirable content so moderators know how it may be able to benefit their communities.

4.5 What design interventions can facilitate proactive moderation strategies?

To determine how we can improve the experience for moderators to find and reinforce content they want to encourage, we asked our participants to reflect on how they currently locate content they want to encourage, and what improvements, if any, they would like to see. First, we asked the following question:

> *"As a moderator, how do you find posts and comments that you want to encourage in [subreddit]?"*

4.5.1 Browsing the subreddit. From the responses, we learned that the majority of moderators (57.5%) find content to encourage through their regular browsing. Some moderators did not indicate taking any action specifically to seek out this type of content, but 19 moderators reported sorting the subreddit by "new," "hot," or "top" to locate quality content.

There were also some responses, typically from moderators of smaller subreddits, which explained that the respondent reads all, or almost all, of the new content in their communities. This reflects lower levels of activity or a coordinated moderation team that splits the task among several moderators such that all content gets reviewed by at least one moderator.

4.5.2 Going through the moderation queue. Interestingly, 21 participants (18%) mentioned that they find posts they want to encourage while going through the moderator queue to approve posts and read reports. Participant P28 said, "*we get so many reports that I find most content through reported threads or comments.*"

Overall, these responses show us that moderators do not typically take specific actions to seek out high-quality content. Instead, they rely on their typical browsing and moderating habits to stumble across desirable contributions.

As a follow-up question, we asked moderators the following:

> *"How can Reddit make it easier for you and other moderators to find posts that you want to encourage?"*

4.5.3 Varying levels of interest in additional features. In response to this question, 52% of moderators did not have any suggestions for new features or thought it was unnecessary, while about 3% of participants explicitly expressed disinterest in changes being made to existing features. Another 3% of participants reiterated that they prefer to allow their communities to self-regulate.

4.5.4 Provide better metrics and tools to discover positive content. The most common problem mentioned by participants was the need to address the lack of adequate metrics and methods to

Table 6. Collection of suggestions our participants had for tools or mechanisms that could help them encourage behavior more easily. Categories are ordered based on frequency in responses.

Suggestion	# Mods	Excerpt From Moderator Response
Extend Existing Feature		
Free Awards for Moderators	13	"[Free] awards only awardable by [...] moderators."
Allow Comment Stickies	6	"Stickying other users comments would help."
Provide Tangible Awards	5	"Perhaps a way to reward users with reddit premium"
Allow More Stickies	3	"pin more than 2 posts at a time"
Extend Flair Mechanism	1	"add flair to posts and comments [using automod]..."
Add Mega Upvote	1	"a 'mega upvote' for exceptional content"
Introduce Positive Features		
Mechanism to Highlight Content	6	"highlight something without stickying it..."
Give Weight to Good Users	3	"Weighted reports by users who [...] report accurately."
Badges to Recognize Users	2	"a badge awarded to [...] top quality and repeat posters."
Disable Downvotes	1	"I would REALLY like to be able to disable downvotes..."
Introduce General Features		
New Ways to Curate/Sort Content	4	"Limiting certain topics..."
More Punitive Actions	2	"[An] automated warning system by Reddit itself"
General Suggestions	4	"post title editing"

discover high-quality comments. Roughly 25% of participants generated new ideas for metrics that can be used to discover good content. Moderators were also interested in being able to view the comments in a subreddit as easily as they can view posts to allow them to locate comments to encourage more easily. Another suggestion four participants mentioned was to provide ways to customize the sorting of content in their communities or allow moderators to develop custom metrics. On a similar note, 4% of participants were interested in AI-assisted discovery, which they believe would give them more power to customize as well.

Roughly 6% of participants presented ideas for tools to isolate positive content, including an automatically-populated queue of posts predicted to be high quality. As stated by one participant:

> "*The mod queue shows us posts that have received reports or have been flagged by automod. We should have something like that for posts and comments people are upvoting.*" (P79)

This queue would be similar to the modqueue, but with a focus on encouragement. These suggestions indicate a need for an integrated tool like ConvEx [18] with a focus on highlighting desirable content instead of toxic content. In addition, 2 moderators suggested a "report"-button-equivalent for good content that users can click to nominate posts and comments. Other moderators wanted more ways to give positive feedback, such as mod-specific awards to show users that their contributions were appreciated by the moderation team.

To further explore suggested improvements to the moderation experience, we asked moderators one final question:

> *"Are there any actions you wish you could take as a moderator to encourage contributions in [subreddit]?"*

We found that 40% of our participants had no suggestions, however, we were able to collect 62 labels from the remaining responses which contain several interesting suggestions. All participant suggestions are reported in Table 6.

4.5.5 More flexible awards. The majority of suggestions, made by 24% of participants, related to extending existing mechanisms and tools to allow them to better encourage content. Roughly

11% of moderators wish they had free awards to give to their communities. Many cited finances as restricting their ability to give encouraging feedback to participants who were contributing meaningfully to the community. Thus, moderators suggested giving moderation teams some number of free, moderator-specific awards to give out as rewards. Many moderators were also interested in some sort of tangible award system, possibly involving giveaways of physical items, or even memberships to Reddit premium.

4.5.6 More flexible sticky-ing. It was mentioned by 6 participants that they wanted the ability to sticky comments so they could not only promote posts, but also comments. Participants also found the 2-sticky limit restrictive. Three moderators explained that they utilize at least one sticky to store the rules or other helpful announcements to their communities, leaving them at most one spot to sticky a post they want to highlight. As said by Participant P27, "*it would be great if I could pin more than 2 posts at a time, so both regularly scheduled posts and quality community posts can be highlighted.*" As mentioned previously, some moderators got around this by linking good posts in a single stickied post, allowing them to effectively highlight more posts at once. However, moderators still believed it would be beneficial to relax this restriction.

4.5.7 Different highlighting features. Beyond extending existing features, participants had ideas for new features and tools. The most common response, given by 6 participants, was that they wanted new ways to highlight content:

> "*The ability to highlight something without stickying it. It may be valuable in the moment and unimportant 2 days from now.*" (P85)

> "*Temporarily starred posts or publicly mod-recommended posts*" (P117)

Related to highlighting, 3 moderators wanted to weigh certain users' content more heavily, and 2 other participants wanted ways to recognize specific users, possibly through badges, to reward on the user level. Specifically, one participant wrote:

> "*how about a badge awarded to users who are top quality and repeat posters. The badge would indicate that the moderators have 'curated' this user as one worth paying attention to.*" (P96)

It is clear that many Reddit moderators are interested in being able to highlight content in new ways that can support their efforts at proactively building healthy communities through feedback.

5 Discussion

Through this paper, we have made four main contributions. We contribute two novel taxonomies describing desirable behavior in Reddit communities from moderator perspectives and positive feedback strategies currently employed by moderators. We present clear evidence that many moderators employ these actions with the explicit purpose of positive reinforcement. Finally, we provide design recommendations to allow moderators to engage more effectively with reinforcement. This section discusses the implications of these contributions and raises questions for future work.

5.1 Taxonomies Contributed

For CSCW theory, this paper presents a novel compilation of content and behavior Reddit moderators want to encourage. This taxonomy (Table 3) is imperative to researchers wanting to quantitatively detect desirable behavior. Specifically, these findings indicate that we need to have effective measures of prosocial behavior to quantitatively detect contributions that embody this attribute. Recent work has explored automatic detection of prosocial behavior [6, 8, 22, 62, 63, 84], however, this taxonomy provides evidence that we need to put more emphasis on detecting a

wider range of desirable behavior, especially considering the abundance of research on detecting undesirable behavior.

Another contribution of this work is the second taxonomy we created (Table 4) based on survey responses describing the different actions moderators take when they see content or behavior they want to encourage. Prior work shows that moderators on some other platforms already use positive reinforcement in moderation [83], but our taxonomy of actions is the first to provide concrete evidence that moderators on Reddit actually engage with positive feedback when presented with content they want to encourage. Additionally, we see that many of the actions moderators report taking are public-facing, and thus can be used in future work as a way to computationally identify examples of behavior that moderators want to encourage.

5.2 Implications for Online Moderation

Our two taxonomies provide important implications for moderation on Reddit and beyond. Across different communities, moderators had different perspectives about what their roles were. We saw that some participants viewed their moderator role as solely punitive or thought that communities should be completely self-regulated, while others were in full support of encouraging desired content and behavior. This means that it is important for moderation teams, on Reddit and other platforms more generally, to reach some consensus about their role as moderators. Considering the moderator roles identified by Seering et al. [74], moderators who see their role as "Nurturing and Supporting Communities" may be more open to utilizing positive feedback for reinforcement.

Therefore, moderation teams need to decide whether encouragement of desired content should be employed, weighing the potential benefits with the perceived loss of democracy. For moderation teams interested in employing reinforcing feedback methods, they must also discuss what types of behavior to encourage and how they should go about doing so. In Reddit's case, even moderation teams that wish to keep reinforcement separate from their role as moderators can look to the second taxonomy for suggestions for anonymous feedback, such as upvoting or commenting without being distinguished as moderators. This can help moderators encourage desirable content and behavior without invoking their moderator status. We also suggest that new and emerging Reddit communities in particular think about the role they want positive reinforcement to play in their moderation practices.

Beyond Reddit, moderators generally can take these same considerations into account when deciding the role of moderation in their communities. For example, moderators on Twitch should understand the norms of their community as determined by the creator, and be on the same page about how to reinforce those behaviors. This may involve small actions such as sending encouraging replies in the chat, or something more tangible like a gifted subscription to the channel. Moderation teams on any platform can decide how to encourage behavior given their specific norms and available signals for reinforcement.

5.3 Incorporating Positive Feedback Mechanisms into the User Interface

From our taxonomy of actions, we know that many moderators are putting in the extra effort to provide positive feedback to members of their communities. However, Reddit's current design does not always make it clear to the recipients that the feedback was provided by moderators. This is potentially a waste of limited moderator resources and suggests that Reddit should provide moderators with the option to make these forms of feedback distinct from when they are provided by a non-moderator user. For example, Reddit may consider the design of YouTube creator hearts [19], a form of endorsement from the creator that is visually highlighted in the interface.

Additionally, we believe incorporating more explicit tools for providing positive feedback into the interface can both reconcile this invisible moderator effort and also address many reasons

moderators cited to explain why they do not provide positive feedback. This applies to existing platforms like Reddit, but also to new platforms looking to build healthy communities from the start. On Reddit, for example, moderators can currently see punitive actions, such as content removals and bans, explicitly through the interface. However, to provide positive feedback to content they want to encourage, moderators needed to work with tools and mechanisms which were not explicitly intended for providing moderator feedback. Thus, to give moderators the opportunity to use positive reinforcement to shape their communities, we must give them the tools to do so.

We described several key design suggestions made by moderators during our survey, including the idea of a *positive* queue that would automatically populate with content predicted to be considered desirable by the moderators. This would allow moderators to easily parse through posts and comments that are potentially high-quality. This helps make content more easily discoverable—a well-known struggle faced by moderators of large-scale online communities [18, 45].

To provide moderators with a mechanism explicitly intended for positive feedback, we can turn to another suggestion made by our participants. Several moderators suggested free mod-specific awards, or other ways to specifically highlight content without sticky-ing, so users can easily see which contributions were especially appreciated by moderators. This new highlighting mechanism can be incorporated into the positive queue with a button, helping moderators feel like providing positive feedback is feasible and a part of their role [74].

Incorporating a highlighting mechanism such as this can also help communities abide by principles from prior work. Kiesler et al. [48] assert that communities can encourage adherence to norms by showing examples of norm-adhering behavior. Currently, Reddit does not provide ways for moderators to abide by this design claim. We know from our survey responses that some moderators are adapting pins and the subreddit sidebar to show examples of high-quality content, but Reddit can give communities the tools to engage with example-setting to increase the likelihood that community-members will adopt norms.

5.4 Designing Tools with Moderator Roles in Mind

While we have evidence from moderators that many of these design changes would be welcome, we note that it is necessary to consider who we are designing for. As mentioned in Sections 4.3.4 and 4.5.3, there were several moderators who did not believe encouraging behavior was part of their role as moderator, some who did not want existing features to change, and others who did not want new features at all. This is particularly important since prior work has shown that introducing new automated tools for moderation may change the type of work moderators are responsible for [41]. Additionally, it has been shown that users want to be involved and in control when incorporating automation into their moderation practices [44]. Thus, we need to design tools that account for the fact that not all moderators view their role as including encouragement and that give users the agency to decide if and how they want to engage with our tools.

Additionally, different communities may benefit from different tools. Moderators from larger communities, for example, may require more automation or community-led efforts to successfully locate and reward desirable behavior. Some smaller communities are already capable of sifting through all new content, and instead may benefit from tools to promote content they manually find. Moreover, independent of community size, any tools that are introduced should be flexible enough to adapt to community needs. For example, prior work argues that we should develop community-specific tools that are capable of working well with moderators' existing norms and workflows, regardless of what type of community is using the tool [12, 15]. This speaks to the value of configurability when developing tools to improve the moderator experience.

5.5 Giving Online Communities Control Over Algorithmically-Curated Feeds

In addition to giving moderators the power to decide what tools to use, we also suggest that platforms give communities more agency over their feeds. Most prior research in the area of curation relates to the ways in which users can curate their individual feeds [55, 61] or investigates the effects of appearing on platform-wide algorithmically-curated feeds [10, 11]. However, there is also some existing research exploring scalable ways to utilize curator feedback to curate content on feeds at a community level. He et al. [38] develop a system called *Cura* which leverages the opinions of trusted curators to dictate the content that should be made visible to their communities. Moderators themselves are a pre-defined group of users that may be considered trusted curators for their communities. Additionally, in our survey responses, we noticed examples of moderators currently trying to take control over the sorting of the content in their communities using existing features. For example, moderators linked high-quality posts in a pinned post or their subreddit's sidebar, while others explained that they upvoted, commented on, and flaired posts to boost their visibility in the community. These responses show us that moderators want more control over the way content is displayed in their communities' feeds. Right now, sorting algorithms are most often "black boxes" that do not explain how they operate [24]. As a result, moderators must use some mental model of those curation algorithms to try and promote desirable content. However, it is unlikely that one moderator upvote on a large platform like Reddit will have much impact on where a post appears in the feed. We suggest that social platforms empower their users and moderators to curate their communities' feeds more explicitly.

Alternatively, communities can employ a similar approach to Cura [38] by giving moderator feedback more weight to take advantage of their perspectives. The platform Slashdot gave limited moderator power to users based on a reputation system [54]. Other platforms like Reddit can similarly emphasize their voting mechanism, replies, and other interface signals of dedicated users, either through their own reputation system, or by drawing from the current moderation teams.

5.6 Limitations and Future Work

5.6.1 Larger-scale moderator study. Our current study surveyed a collection of moderators that are not necessarily representative of moderators on Reddit or online communities more broadly. Future work may include a larger-scale survey of moderators from a wider range of communities to better capture the perspectives held by moderators across the platform.

5.6.2 User perspectives on desirable behavior and positive feedback. Though this research focuses on moderator perspectives, it is important to ask the question: *who gets to decide what is desirable?* Prior work has demonstrated the potential for disagreements [30, 31] and misalignment between Reddit users' and moderators' perspectives on rules and moderation practices [49]. As a result, we recognize that our taxonomy of desirable behavior would benefit from user perspectives. What is desirable to moderators may not be desirable to users, and there may even be disagreement between users from the same community. In this paper, we are solely reporting on moderator perspectives and do not assert that our taxonomy captures the opinions of the entire community.

To capture a more complete understanding of our research questions, future work may include surveying users from these communities to supplement the findings in this paper. Additionally, we cannot currently tell whether users interpret moderator feedback as it is intended, nor whether users themselves engage in practices to reinforce their fellow community-members' behavior. All of these questions can be better informed by a follow-up user study.

5.6.3 Efficacy of actions from our taxonomy. On the quantitative side, our work presents a taxonomy of actions moderators take, however we do not currently evaluate whether any of them are effective

for positive reinforcement of desirable behavior. Many moderators report wanting to reinforce behavior through their actions, opening up avenues for future research examining the efficacy of current methods when applied to real-time interactions online. Additionally, since many methods for reinforcement included in our taxonomy are accessible to all Reddit users and not only moderators, we have the opportunity to explore how receiving positive feedback from both moderators and communities as a whole impacts the recipient's future behavior. This can help inform moderation practices by revealing the actions that can be taken to effectively encourage desired behaviors.

5.6.4 Mechanisms for positive feedback. Since several moderators mentioned the lack of tools to provide explicit positive feedback as an impediment for their employment of strategies to encourage desirable behavior, there is an opportunity to develop computational tools to aid moderators in providing positive feedback to the behaviors they want to encourage. By providing concrete mechanisms for reinforcing desirable contributions, we may be able to help more moderators adopt proactive moderation strategies.

5.6.5 Reinforcement in moderation beyond Reddit. Finally, our research focuses specifically on the context of Reddit, a popular platform which has shown interest in engaging with mechanisms for rewarding desirable behavior. Some of the findings of this work may not apply to contexts beyond Reddit, such as some Reddit-specific actions in our taxonomy of how moderators encourage behavior in their communities. However, we believe that, based on our findings, it is likely that moderators on other platforms are also engaging with methods for reinforcing behavior in their communities. Future work can be done to expand our understanding of how our findings relate to other moderation teams across platforms aside from Reddit.

6 Conclusion

By surveying 115 Reddit moderators, we constructed two taxonomies containing the attributes of behavior moderators find desirable along with the actions they take to encourage those behaviors. We found that more than 95% of surveyed moderators provide positive feedback to encourage desired behavior, and we confirmed that positive reinforcement is an explicit goal of many of those moderators. This paper provides a foundation to further explore positive reinforcement across communities, including non-moderators, both qualitatively and quantitatively. We also contribute design suggestions to make the most of limited moderator resources and to ease the acquisition of positive reinforcement for moderators that do not currently employ it. These contributions together can help researchers and designers support new methods of proactive moderation that may combat the physical and mental challenges associated with traditional reactive moderation, while also promoting healthy communities.

Acknowledgements

We thank the members of the Social Computing Lab (SCUBA) at the University of Illinois Urbana-Champaign for their valuable input that improved our survey design. We also thank Jackie Chan and Yoshee Jain for their help validating our codebooks. Finally, we thank the anonymous reviewers for their very helpful reviews.

References

[1] B. Thomas Adler and Luca De Alfaro. 2007. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*. ACM, Banff Alberta Canada, 261–270. https://doi.org/10.1145/1242572.1242608

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, Rio de Janeiro Brazil, 95–106. https://doi.org/10.1145/2488388.2488398

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. 687–698.
[4] Shaljan Areepattamannil. 2010. Parenting practices, parenting style, and children's school achievement. *Psychological Studies* 55 (2010), 283–289.
[5] Michael B Armstrong and Richard N Landers. 2018. Gamification of employee training and development. *International Journal of Training and Development* 22, 2 (2018), 162–169.
[6] 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*. ACM, Ljubljana Slovenia, 1134–1145. https://doi.org/10.1145/3442381.3450122
[7] Christiane Bradler, Robert Dur, Susanne Neckermann, and Arjan Non. 2016. Employee recognition and performance: A field experiment. *Management Science* 62, 11 (2016), 3085–3099.
[8] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4758–4765. https://doi.org/10.18653/v1/D18-1507
[9] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing Live Streaming Moderation Tools: An Analysis of Twitch. In *International Journal of Interactive Communication Systems and Technologies*, Vol. 9. 36–50. https://doi.org/10.4018/IJICST.2019070103
[10] Jackie Chan, Aditi Atreyasa, Stevie Chancellor, and Eshwar Chandrasekharan. 2022. Community Resilience: Quantifying the Disruptive Effects of Sudden Spikes in Activity within Online Communities. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–8. https://doi.org/10.1145/3491101.3519813
[11] Jackie Chan, Charlotte Lambert, Fred Choi, Stevie Chancellor, and Eshwar Chandrasekharan. 2024. Understanding Community Resilience: Quantifying the Effects of Sudden Popularity via Algorithmic Curation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18.
[12] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–30. https://doi.org/10.1145/3359276
[13] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. In *ACM Transactions on Computer-Human Interaction*, Vol. 29. 1–26. https://doi.org/10.1145/3490499
[14] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1. 1–22. https://doi.org/10.1145/3134666
[15] 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. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. 1–25. https://doi.org/10.1145/3274301
[16] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. 1–37. https://doi.org/10.1145/3555603
[17] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks* (Palo Alto, California, USA) *(COSN '15)*. Association for Computing Machinery, New York, NY, USA, 233–243. https://doi.org/10.1145/2817946.2817959
[18] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 262 (oct 2023), 30 pages. https://doi.org/10.1145/3610053
[19] Frederick Choi, Charlotte Lambert, Vinay Koshy, Sowmya Pratipati, Tue Do, and Eshwar Chandrasekharan. 2024. Creator Hearts: Investigating the Impact Positive Signals from YouTube Creators in Shaping Comment Section Behavior. *arXiv preprint arXiv:2404.03612* (2024).
[20] Christine L. Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–19. https://doi.org/10.1145/3555168
[21] Tiago Cunha, Ingmar Weber, and Gisele Pappa. 2017. A Warm Welcome Matters!: The Link Between Social Feedback and Weight Loss in /r/loseit. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, Perth, Australia, 1063–1072. https://doi.org/10.1145/3041021.3055131
[22] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, 250–259. https://aclanthology.org/P13-1025

[23] Don Dinkmeyer and Gary D McKay. 1989. *The parent's handbook: Systematic training for effective parenting*. ERIC.

[24] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2702123.2702556

[25] Charles Bohris Ferster and Burrhus Frederic Skinner. 1957. *Schedules of reinforcement*. Appleton-Century-Crofts, East Norwalk. https://doi.org/10.1037/10627-000

[26] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12. https://doi.org/10.1609/icwsm.v12i1.15033

[27] Arpita Ghosh and Patrick Hummel. 2012. Implementing optimal outcomes in social computing: A game-theoretic approach. In *Proceedings of the 21st international conference on World Wide Web*. 539–548.

[28] Arpita Ghosh and Preston McAfee. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*. 137–146.

[29] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[30] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3502004

[31] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445423

[32] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[33] Omkar Gurjar, Tanmay Bansal, Hitkul Jangra, Hemank Lamba, and Ponnurangam Kumaraguru. 2022. Effect of Popularity Shocks on User Behaviour. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 253–263. https://doi.org/10.1609/icwsm.v16i1.19289

[34] Hussam Habib, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand. 2019. To Act or React: Investigating Proactive Strategies For Online Community Moderation. (June 2019). http://arxiv.org/abs/1906.11932 arXiv:1906.11932 [cs].

[35] Aaron Halfaker and R. Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *CoRR* abs/1909.05189 (2019). arXiv:1909.05189 http://arxiv.org/abs/1909.05189

[36] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. 1–10.

[37] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. IEEE, 3025–3034.

[38] Wanrong He, Mitchell L. Gordon, Lindsay Popowski, and Michael S. Bernstein. 2023. Cura: Curation at Social Media Scale. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–33. https://doi.org/10.1145/3610186

[39] 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

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

[41] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[42] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5. 1–30. https://doi.org/10.1145/3479525

[43] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–27. https://doi.org/10.1145/3359252

[44] 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 (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. https://doi.org/10.1145/3491102.3517505

[45] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[46] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3658–3666. https://doi.org/10.18653/v1/P19-1357

[47] Charles Kiene, Kate Grandprey-Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun "Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, and Bryan Dosono. 2019. Volunteer Work: Mapping the Future of Moderation Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. ACM, Austin TX USA, 492–497. https://doi.org/10.1145/3311957.3359443

[48] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012).

[49] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7. 286:1–286:36. https://doi.org/10.1145/3610077

[50] Yubo Kou and Xinning Gui. 2021. Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[51] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology* (2 ed.). Sage, Thousand Oaks, California.

[52] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559. https://doi.org/10.1609/icwsm.v16i1.19314

[53] Cliff Lampe. 2012. The role of reputation systems in managing online communities. *H. Masum, M. Tovey, eds* (2012), 77–88.

[54] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[55] Francis L. F. Lee, Michael Che-ming Chan, Hsuan-Ting Chen, Rasmus Nielsen, and Richard Fletcher. 2019. Consumptive News Feed Curation on Social Media as Proactive Personalization: A Study of Six East Asian Markets. *Journalism Studies* 20, 15 (Nov. 2019), 2277–2292. https://doi.org/10.1080/1461670X.2019.1586567 Publisher: Routledge _eprint: https://doi.org/10.1080/1461670X.2019.1586567.

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

[57] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 596–606. https://doi.org/10.1609/icwsm.v16i1.19318

[58] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest Q&A site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2857–2866.

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

[60] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey Research in HCI. In *Ways of Knowing in HCI*. Springer, New York, NY, 229–266. https://doi.org/10.1007/978-1-4939-0378-8_10

[61] Brigitte Naderer, Raffael Heiss, and Jörg Matthes. 2020. The skilled and the interested: How personal curation skills increase or decrease exposure to political information on social media. *Journal of Information Technology & Politics* 17, 4 (Oct. 2020), 452–460. https://doi.org/10.1080/19331681.2020.1742843 Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2020.1742843.

[62] Courtney Napoles, Aasish Pappu, and Joel Tetreault. 2017. Automatically Identifying Good Conversations Online (Yes, They Do Exist!). *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 628–631. https://doi.org/10.1609/icwsm.v11i1.14959 Number: 1.

[63] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*.

Association for Computational Linguistics, Valencia, Spain, 13–23. https://doi.org/10.18653/v1/W17-0802
[64] Maria Papoutsoglou, Georgia M Kapitsaki, and Lefteris Angelis. 2020. Modeling the effect of the badges gamification mechanism on personality traits of Stack Overflow users. *Simulation Modelling Practice and Theory* 105 (2020), 102157.
[65] Chris Perryer, Nicole Amanda Celestine, Brenda Scott-Ladd, and Catherine Leighton. 2016. Enhancing workplace motivation through gamification: Transferrable lessons from pedagogy. *The International Journal of Management Education* 14, 3 (2016), 327–335.
[66] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.
[67] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9 (2006), 79–101.
[68] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2022. Automated Content Moderation Increases Adherence to Community Guidelines. http://arxiv.org/abs/2210.10454 arXiv:2210.10454 [cs].
[69] Sarah T. Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[70] Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. 2017. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior* 69 (2017), 371–380.
[71] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
[72] Ari Schlesinger, Eshwar Chandrasekharan, Christina A Masden, Amy S Bruckman, W Keith Edwards, and Rebecca E Grinter. 2017. Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6912–6924.
[73] 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. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. 1–27. https://doi.org/10.1145/3555095
[74] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. In *New Media & Society*, Vol. 24. 621–640. https://doi.org/10.1177/1461444820964968
[75] 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*. ACM, Portland Oregon USA, 111–125. https://doi.org/10.1145/2998181.2998277
[76] Burrhus Frederic Skinner. 1963. Operant behavior. *American Psychologist* 18, 8 (1963), 503–515. https://doi.org/10.1037/h0045185 Place: US Publisher: American Psychological Association.
[77] Burrhus Frederic Skinner. 1965. *Science and human behavior*. Number 92904. Simon and Schuster.
[78] Burrhus Frederic Skinner. 1989. *Recent issues in the analysis of behavior*. Prentice Hall.
[79] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–21. https://doi.org/10.1145/3359265
[80] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445092
[81] Laurence Steinberg, Susie D Lamborn, Sanford M Dornbusch, and Nancy Darling. 1992. Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child development* 63, 5 (1992), 1266–1281.
[82] Pamela Bilo Thomas, Daniel Riehm, Maria Glenski, and Tim Weninger. 2021. Behavior Change in Response to Subreddit Bans and External Events. In *IEEE Transactions on Computational Social Systems*, Vol. 8. 809–818. https://doi.org/10.1109/TCSS.2021.3061957
[83] Yixue Wang and Nicholas Diakopoulos. 2021. Highlighting High-quality Content as a Moderation Strategy: The Role of *New York Times* Picks in Comment Quality and Engagement. In *ACM Transactions on Social Computing*, Vol. 4. 1–24. https://doi.org/10.1145/3484245
[84] Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 33–45.
[85] Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 1121–1132. https://doi.org/10.1609/icwsm.v16i1.19363
[86] Galen Weld, Amy X. Zhang, and Tim Althoff. 2024. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 1611–1633. https://doi.org/10.1609/icwsm.v18i1.31413

[87] 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*. ACM, Glasgow, Scotland, UK, 1–13. https://doi.org/10.1145/3290605.3300390

[88] Ryan Yen, Li Feng, Brinda Mehra, Ching Christie Pang, Siying Hu, and Zhicong Lu. 2023. StoryChat: Designing a Narrative-Based Viewer Participation Tool for Live Streaming Chatrooms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 795, 18 pages. https://doi.org/10.1145/3544548.3580912

[89] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1350–1361. https://doi.org/10.18653/v1/P18-1125

A Survey Instrument

A.1 Consent Form

What is the purpose of this study? You are being asked to participate in a voluntary research study. The purpose of this study is to understand how Reddit moderation teams can effectively provide positive feedback to community-members. Participating in this study will involve filling out a survey and your participation will last approximately 20 minutes. There will be no more than minimal risk to the participant beyond those risks that exist in daily life; benefits related to this research include improving moderation systems on Reddit, and, more broadly, improving internet safety and online community health.

What procedures are involved? You will be asked to reflect on your experience as a user or moderator on Reddit and select one subreddit to focus your survey responses on. The survey will then ask a series of questions about the types of content you like to see in your community and how you perceive different types of feedback from other users.

What are the potential risks and discomforts? Participation in this research will not expose you to a level of risk greater than what you would experience in your daily life.

Are there benefits to participating in the research? There are no direct benefits to participants from participating in this survey, though some may appreciate the opportunity to provide feedback to researchers who aim to improve their experience on Reddit in the long term.

Will my study-related information be kept confidential? We will use all reasonable efforts to keep your personal information confidential, but we cannot guarantee absolute confidentiality. When this research is discussed or published, no one will know that you were in the study. Your email may be collected if you opt into the raffle, but it will not be released publicly. Any reports made using your data will only be in conjunction with all data recorded in a summarized manner.

Will I be reimbursed for any expenses or paid for my participation in this research? You will be able to enter a raffle to win a $20 Amazon gift card upon completion of the survey. Entering the raffle requires that you provide your Reddit username and email address. Your email will only be used to activate your Amazon code if you win the raffle.

Can I withdraw or be removed from the study? If you decide to participate, you are free to withdraw your consent and discontinue participation at any time. Your participation in this research is voluntary. Your decision whether or not to participate, or to withdraw after beginning participation, will not affect your current or future dealings with University of Illinois Urbana-Champaign.

Will data collected from me be used for any other research? Your information will not be distributed or used for future research.

Who should I contact if I have questions? If you have questions about this project, you may contact us via eshwar@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois at Urbana-Champaign Office for the Protection of Research Subjects at 217-333-2670 or via email at irb@illinois.edu.

Please print this consent form if you would like to retain a copy for your records.

I currently reside in the United States.

- Yes
- No

I am above the age of 18.

- Yes
- No

I have read and understood the above consent form. I certify that I am 18 years old or older. By clicking the "I consent" button to enter the survey, I indicate my willingness to voluntarily take part in this study. If you click "No, I do not consent", you will be directed to the end of the survey and your data will not be saved.

- Yes, I consent
- No, I do not consent

A.2 Self-reported Reddit Activity

How many years have you had your Reddit account?

How often do you use Reddit?

- Every day
- A few times per week
- Once a week
- A few times per month
- Rarely

Typically, how often do you post or comment on Reddit?

- Frequently (i.e., I frequently submit posts or comments on threads)
- Occasionally (i.e., I post or comment occasionally, but mostly browse what others have submitted)
- Rarely (i.e., I almost always just browse what others have submitted)
- Never (i.e., I only browse what others have submitted)

How often do you use aggregate subreddits (like your front page, r/all, or multi-Reddits)?

- Always (i.e., I never look at individual subreddits)
- Frequently (i.e., I mostly use aggregate subreddits, but sometimes I look at individual subreddits)
- Occasionally (i.e., I look at aggregate subreddits and individual subreddits about evenly)
- Rarely (i.e., I mostly look at individual subreddits)
- Never (i.e., I only look at individual subreddits)

Do you moderate any subreddits?

- Yes
- No

What subreddit do you moderate? If you moderate multiple, select the subreddit that received our modmail with the link to this survey. You will be asked to focus on this subreddit for the rest of your responses in this survey. Format it with the prefix "r/" (for example, r/science).

How many communities do you moderate? (Please enter a number).

*How long have you been a moderator of **any community** on Reddit?*

- Less than a year
- 1-2 years
- 3-5 years
- 5+ years

*How long have you been a moderator of **[subreddit]**?*

- Less than a year
- 1-2 years
- 3-5 years
- 5+ years

*Roughly how often do you submit posts or comments to Roughly how often do you submit posts or comments to **[subreddit]**?*

- Multiple times per day
- Daily
- Weekly
- Monthly
- Less than monthly
- Never

*To the best of your knowledge, how many times have the moderators of **[subreddit]** removed comments you've submitted?*

- Never
- Once
- 2-5 times
- 5+ times

To the best of your knowledge, how many times have your comments been removed anywhere on Reddit?

- Never
- Once
- 2-5 times
- 5+ times

A.3 Providing feedback as a moderator

*Many subreddit moderation teams use punitive measures like content removals and bans to discourage user behaviors that are considered to be violating community norms. On the other hand, some subreddit moderation teams may actively employ strategies to encourage and incentivize certain types of user behavior. How often does the moderation team in [subreddit] take actions to **encourage** user behavior?*

- Always
- Frequently
- Occasionally
- Rarely
- Never

*What kinds of content and behavior would you want to encourage in **[subreddit]** as a moderator?*

*How often do you take actions to encourage such behavior in **[subreddit]**?*

- Always
- Frequently
- Occasionally
- Rarely
- Never

(If "Never" is selected) Why do you not take actions to encourage behavior?

*(If "Never" is not selected) What actions do you take as a moderator when you see comments or posts you want to encourage in **[subreddit]**?*

(If "Never" is not selected) What do you intend to achieve through these actions?

A.4 Locating Posts to Encourage

*As a moderator, how do you find posts and comments that you want to encourage in **[subreddit]**?*

How can Reddit make it easier for you and other moderators to find posts that you want to encourage?

*Are there any actions you wish you could take as a moderator to encourage contributions in **[subreddit]**?*

A.5 Raffle

*If you wish to enter a raffle to win a $20 Amazon gift card, please provide both your **Reddit username** and **email address**. If you win, you will receive an email with a code to activate your gift card. Would you like to enter the raffle?*

- Yes
- No

(If "Yes") Please enter your Reddit username.

(If "Yes") Please enter your email address.

Received July 2023; revised January 2024; accepted March 2024