Investigating Moderation Challenges to Combating Hate and Harassment: The Case of Mod-Admin Power Dynamics and Feature Misuse on Reddit

Madiha Tabassum[1], Alana Mackey[2], Ashley Schuett[3], Ada Lerner[1]
[1]*Northeastern University,* [2]*Wellesley College,* [3]*George Washington University*

Abstract

Social media platforms often rely on volunteer moderators to combat hate and harassment and create safe online environments. In the face of challenges combating hate and harassment, moderators engage in mutual support with one another. We conducted a qualitative content analysis of 115 hate and harassment-related threads from r/ModSupport and r/modhelp, two major subreddit forums for this type of mutual support. We analyze the challenges moderators face; complex tradeoffs related to privacy, utility, and harassment; and major challenges in the relationship between moderators and platform admins. We also present the first systematization of how platform features (including especially security, privacy, and safety features) are misused for online abuse, and drawing on this systematization we articulate design themes for platforms that want to resist such misuse.

1 Introduction

Online platforms like Reddit provide social spaces to build diverse communities around many topics and interests. However, widespread use of these platforms has also given rise to online hate and harassment, such as hateful speech, sexual harassment, violence, trolling, and doxing. As of 2021, 41% of U.S. adults have personally experienced some harassment online [13]. Members of marginalized or minoritized identities were more likely to be harassed because of their physical appearance, gender, race or ethnicity, religion, sexual orientation, gender identity, or disability [8].

Social media platforms often rely on volunteer moderators to reduce online abuse and harassment and maintain community safety. Moderators are responsible for monitoring the behavior of community members and enforcing community rules. Multiple studies have looked at moderators' experiences and practices [14, 16, 22, 32, 49], moderators' activities and labors [21, 37], and their processes for identifying rule violators as they work toward sustaining online communities [16]. These studies have discovered that moderators face complex challenges, including being personally targeted by harassers on the internet [14, 46], emotional burnout [21, 52], and lack of support from the platform [22].

Though past work has identified moderation challenges in online communities, limited research has looked at moderation specifically through the lens of preventing hate and harassment. Moreover, these studies were mostly conducted via interviews and surveys. Our work complements these works by studying organic, unfiltered, and spontaneous conversations in situ from two online communities for moderators, Reddit's r/ModSupport, and r/modhelp subreddits, avoiding social desirability bias that can occur in more controlled settings like interviews. Moderators post moderation-related topics or queries to these communities to get answers and suggestions from fellow moderators and Reddit administrators.

We have conducted a qualitative content analysis of 115 threads with 2,740 comments mentioning hate and harassment attacks and abuse from these subreddits. We systematically analyzed moderators' discussions in these subreddits to understand the challenges and unwanted trade-offs they face in managing community safety and the needs that emerge from these challenges. We found these conversations emphasize a uniquely challenging and adversarial subset of moderation situations, which appear naturally in r/ModSupport and r/modhelp by their nature as places for moderators to seek support with issues that challenge the limits of their skills, tools, and resources. This emphasis on challenging scenarios enables a major novel theme of our results and discussion - adversarial misuse of platform features - since adversarial misuse causes many such challenging scenarios. Overall, Our contributions include:

- Insight into the breadth of challenges moderators face, including the misuse of platform features for harassment, inadequate and unclear moderation tools, lack of support for moderators' safety, vague platform policy and actions, etc., drawing on a found dataset that uniquely highlights moderation challenges.
- The first systematization of how platform features are

misused for hate and harassment, mapping how different features can be misused to facilitate attacks,, a novel contribution enabled by the unique nature of our in situ, challenge-dominated dataset. We generalized these misuses beyond Reddit, and articulated design themes for creating platform features that are more robust to adversarial misuse by toxic users.
- An analysis of the of the critical but rarely studied relationship between moderators and platform administrators and its role in preventing hate and harassment.

2 Background and Related Works

2.1 Harassment in Online Communities

Online hate and harassment have been steadily increasing in recent years, posing a grave challenge to the digital world's well-being and inclusivity [13]. This rise is influenced by the anonymity provided by the internet and the polarization of online spaces, where like-minded individuals gravitate towards echo chambers, reinforcing extremist beliefs and normalizing abusive behavior [19, 23]. Online hate and harassment manifest in different ways, ranging from dissemination of toxic content (e.g., hate speech, threats of violence, etc.) to surveillance and impersonation [55]. Users employ diverse protective strategies to limit such harassment, including reporting, blocking, and disengaging from online platforms [3, 13, 56].

In addition to user reporting, many major social media platforms such as Facebook, Reddit, and Twitch depend on volunteer moderators to enforce rules, monitor content, and identify and action offenders while fostering an engaging and respectful community [35, 42, 49]. According to the Reddit 2023 Transparency Report, 167.2 million pieces of content were removed from Reddit in the first half of 2023, 49.5% by volunteer moderators [6]. Our paper provides an in-depth analysis of salient challenges to this moderation work.

2.2 Moderation Challenges

User-driven volunteer moderation has been studied on various platforms from varying perspectives in recent years and has demonstrated the challenges faced by the moderators. One of the primary challenges moderators face is the sheer volume and scale of user-generated content they need to moderate and the labor required to do that [18, 37, 60]. In addition to that, moderators need to balance free speech with maintaining a respectful environment, which poses a persistent dilemma when defining the line between acceptable and inappropriate content [33, 44]. There have been some automated moderation tools, such as auto-moderator in Reddit, which can be employed to reduce the workload of the moderators. While these automated moderation tools are aimed at lightening the load, previous studies have revealed that such tools are limited in their comprehension of context, linguistic nuances, and the intricacies of social interaction [14, 26, 30, 36], making manual intervention by moderators necessary for nuanced, case-specific decisions, even with automated assistance [49].

Previous work has also looked at the psychological impacts of reviewing toxic content and handling harassing behavior for extended periods of time [21, 52]. Dosono et al. showcased the substantial emotional labor moderators experience as they engage in moderation tasks, meeting platform, and community expectations that ultimately lead to burnout [21]. Steiger et al. highlighted the arduous exercise of moderators in managing and maintaining personal boundaries, such as avoiding burnout and navigating complex interpersonal conflicts within the community [52]. Furthermore, Almerekhi et al. pointed out the targeted hate and harassment towards moderators by users actioned for violating community rules [11].

Several studies investigated transparency in moderation and highlighted the importance of providing explanations of moderation actions in building trust and accountability [15, 29]. However, moderators encounter challenges in delivering transparency as it requires more moderation effort and time to manually provide explanations [31]. More transparency could also escalate conflict and harm against moderators, leading to emotional burnout [25, 49]. Moderation gets even more complex with the diversity of platform contexts and communication mediums. For instance, Jiang et al. discussed the mechanisms of abuse within voice-based communities and the moderators' struggles in identifying and addressing violations during real-time, ephemeral interactions [32]. Wohn et al. underscored the difficulty of combating harassment within large, fast-paced live chat exchanges on platforms like Twitch, necessitating immediate moderation decisions [59].

Most of the previous works have explored challenges in specific moderation issues, i.e., content governance, emotional labor, transparency, etc. We extend the current literature by providing a comprehensive overview of the challenges moderators encounter in their moderation work by leveraging online discussion data (Reddit) among moderators. Our paper draws from forums specifically dedicated to moderator support, offering a particularly close view of the most challenging issues encountered by moderators.

2.3 Moderation on Reddit

Moderation on Reddit involves a combination of automated tools, volunteer subreddit moderators, and professional platform administrators. Each subreddit has volunteer moderators who enforce community rules. Moderators can remove posts or comments, issue warnings, or ban users who violate those rules. Reddit also has a small team of paid employees, called admins, who manage sitewide policies and legal matters and can issue sitewide bans. Reddit's sitewide automated tools include "automoderator", which assists mods by automatically identifying and removing abusive content [1]. Users can report rule violations and communicate with moderators

through modmail, a shared per-subreddit inbox that moderators use to communicate with each other and handle requests from community members [7]. The "modqueue" is a central, per subreddit listing of "all the pieces of content in the community that need to be reviewed by [moderators]—including user reports, filtered posts, and comments" [2].

3 Methods

In this section, we detail our approach to data collection, explaining how we selected the subreddits and threads to analyze. We then outline our analysis process.

3.1 Forum Selection and Data Collection

We looked for subreddits designed to assist moderators on Reddit's official moderator support page [5]. We have considered the following inclusion criteria for subreddits:

- Moderators can post in the subreddit (i.e., r/modnews was excluded because only Reddit admins can post).
- Subreddits specifically for moderators to discuss moderation-related topics (i.e., r/help was excluded because any Redditors can post any Reddit-related topic).
- Subreddit is not limited to a particular moderation topic or geographical area (i.e., r/AutoModerators was excluded because it only discusses automod, and r/FrMods was excluded because it only includes moderators from French-speaking communities).

Based on these criteria, we have decided to study r/ModSupport and r/modhelp.[1] r/ModSupport and r/modhelp subreddits are specifically designed for moderators to discuss diverse topics related to moderation (i.e., moderation issues, moderation tools, online abuse in community, etc.) and seek support and advice from admins and other moderators. These two forums have a large user base: r/modhelp, established in 2009, has 133k members (the largest community of moderators in Reddit), and r/ModSupport, established in 2015, has 92k members as of April 2024. Both are very active with over ten daily threads, providing a rich dataset for our study.

We downloaded a publicly available datset of all available threads[2] from these subreddits from inception to December 2022. We used pushshift.io, a platform that is used to maintain an up-to-date public archive for Reddit to download the threads[3]. We omitted threads where the posts were empty, deleted by the poster, or removed by moderators. The resulting corpus contains 41,256 threads: 12,345 threads from r/Modsupport with an average of 12.58 (SD: 47.46) comments per thread, and 28,911 threads from r/modhelp with an average of 5.82 (SD: 6.52) comments per thread.

[1]Table depicting how these subreddits were selected: https://osf.io/3hk8r

[2]A thread consists of the initial post and subsequent comments and replies.

[3]https://pushshift.io/. We collected our data in January 2023, prior to Reddit revoking Pushshift.io's API access in May 2023

3.2 Sampling a Set of Threads for Analysis

We focused on threads where moderators posted about hate, harassment, and abuse-related attacks towards their community or themselves, asking questions/suggestions about those and sharing challenges in keeping their community safe against those attacks. We used a broad definition of hate and harassment taken from Pew Research [13] and Thomas et al. [55] while sampling the threads: *"Hate, harassment, and abuse occur when an aggressor (either an individual or group) specifically targets another person (including moderators) or group to inflict harm: emotional, financial, or physical. In its milder forms, it creates a layer of negativity that people must shift through as they navigate their daily routines online. At its most severe, it can compromise users' privacy, force them to choose when and where to participate online, or even pose a threat to their physical safety, e.g., doxing and swatting."*

This section will describe the process of generating our dataset for the main analysis, which contains such threads.

Keyword Generation Process: We developed a set of keywords/key phrases to identify posts relevant to hate, harassment, and online abuse. Our initial list of 47 keywords/key phrases drew on Thomas et al.'s taxonomy [55] combined with the set of reasons related to hate and harassment (i.e., harassment, hate, spam, etc) that Reddit's report form offers users for describing content that breaks site rules. We included all hate and harassment-related attacks from Thomas et al.'s taxonomy in our keyword list, except for removing those irrelevant to our topic, such as "IoT manipulation" and "zoom bombing," via discussion among the authors.

We searched for threads in our corpus using those keywords/key phrases such as bullying, trolling, etc. We removed a term from our keyword list if there were no threads containing that term or some word form of that term. For example, we include "bully" in our keyword list if we find a thread with the terms "bully," "bullying," "bullied," or "bullies." After this process, we had a list of 29 keywords/key phrases.

We sought to keep threads where the post contained at least one of the 29 keywords/key phrases or some word form of those terms. The matching process was case-insensitive and was conducted at the word level rather than the character level to avoid mismatches like matching "troll" to "stroll." 6,550 threads contained at least one keyword/key phrase.

We drew a random sample of 1,000 threads from these 6,550. Reading through this sample, we realized that the posts with keyword "spam" were frequent but rarely relevant: 603 out of 1,000 posts contained the word "spam", but only 53 of those were relevant to our analysis. Of those 53, 83% also contained at least one other keyword from our list. Therefore, we removed the keyword "spam" from our keyword list as we did not want our final dataset to be dominated by threads containing such posts. We changed the word "hate" to "hateful" and "hatred," as the word "hate" triggered many false positives and the words "hateful" and "hatred" matched the discussion

on online attacks. We added seven more keywords, "racism," "bigot," "transphobic," "homophobic," "scam," "AEO," and "abuse," because we found relevant posts containing these keywords that were not triggered by any other keywords from our list. Finally, we removed the key phrase "report abuse" from our list to remove duplicates, as we included "abuse" in our keyword list. In the end, we had a list of 35 keywords and key phrases, provided in https://osf.io/3hk8r. Searching these keywords left us with 3,321 threads in our final dataset.

Final Sample: We randomly sampled and coded threads from our final dataset until we reached thematic saturation, following the guidelines in prior research [47]. At each stage of random sampling, two researchers manually reviewed and discussed each sampled thread. If it was a false positive (i.e., unrelated to online hate, harassment, and community safety), we replaced it with a new randomly sampled thread, ensuring that all coded threads were relevant to the study. In total, we coded and reached saturation with 115 relevant threads[4] (2740 comments) sampled from all 3321 threads.

3.3 Data Analysis

We used an inductive coding process to analyze the threads. Our analysis considered both the post and all the comments in each thread. First, three researchers went through 50 threads in multiple rounds to reveal initial codes. The research team met multiple times in this process to discuss the codes, clarify definitions, resolve disagreements, and establish an initial codebook with fifteen codes. Two researchers then coded sets of 20-25 codes at a time, meeting between sets to compare codes, resolve disagreements, and revise the codebook until no new code emerged. Seven new codes were added to the codebook in the process. After the 95th thread, no new codes appeared from the additional 20 threads. Therefore, it was deemed that the data collection had reached a saturation point at the 115th thread. In the end, there were 22 codes in our final codebook. The final codebook is provided in the Appendix A.1. Both coders coded and discussed the same set of threads and agreed on the codes, so we do not report inter-coder agreement. Then, we conducted a thematic analysis by revisiting all the threads and codes multiple times to identify common themes. The research team held regular meetings to review and discuss the analysis results and the generated themes .

3.4 Ethical Considerations

Our institution's Institutional Review Board determined that this study was out of scope for their oversight. Nevertheless, this work has significant ethical implications for the moderators whose words we studied and the communities they protect. The data we analyzed are direct quotes from Reddit moderators, many of whom are from marginalized communities. Though this content is publicly available to anyone, our aggregating it as a dataset and highlighting aspects of it in this manuscript could induce unwanted or dangerous attention (including hate and harassment) towards moderators and their communities. We took several steps to mitigate these dangers. We chose not to release the aggregated dataset publicly. We redact any usernames, specific subreddits (other than `/r/ModSupport` and `/r/modhelp`), or specific communities (e.g., when discussing subreddits associated with a physical city). Additionally, to increase the difficulty of re-identifying specific comments and commenters for targeted harassment, we have paraphrased all quotations that appear in the paper (one researcher paraphrased and another reviewed each paraphrase for fidelity to the original meaning).

3.5 Limitations

This study only focused on publicly available threads from Reddit moderator communities. Our future work will address this limitation directly by recruiting Reddit moderators as interviewees to discuss the in-depth and complex aspects of questions inspired by our findings. Another limitation is that we only analyzed communities primarily using English. The nature of our research left many variables unknown. We did not have access to any data on the demographics of the moderators we studied, so their gender, education level, occupation, age, and location remain unknown to us.

4 Results

In this section, we present the results of our analysis of the Reddit data from two moderator support communities, r/ModSupport and r/modhelp. Though we analyzed additional major themes (e.g., the types of support exchanged among moderators), we chose to narrow our presentation of results to those focused on moderation challenges. We made this choice because the forums we studied provide a particularly rich source of data on challenges. Since by their nature as venues for moderation support, moderators often post in them when facing tricky issues they were unable to handle on their own. Thus compared to other work, our findings likely represent more difficult, frustrating, or unusual challenges, while downplaying more common ones that are typically handled by standard procedures, solo problem solving, or moderation tools. We then present a set of complex tradeoffs that moderators identified as underlying many of these challenges.

4.1 Moderation Challenges

Moderators described various challenges they face as they fight hate, harassment, & abuse to maintain community safety.

[4]The sample is used in another paper of our authorship [54] to understand mutual support among moderators in mod-support communities.

4.1.1 Misuse of Platform Features

As moderators discussed challenges with harassment, one major theme that emerged was the frequent and creative misuse of platform features in ways that enabled toxic behavior, amplified its harms, or made it more difficult to moderate. While the misuse of platform features for hate and harassment has been noted before (e.g., on Discord [32] or Facebook [27]), we believe that we are the first to systematically describe and characterize it as a broader phenomenon. Table 1 summarizes all of the platform features that were mentioned as being misused in our dataset and how it affects moderation. We categorized these misused features into four categories:

- **Community engagement:** Ability to create, engage or interact with community (e.g., posting or creating subs).
- **Mod/admin engagement:** Ability to interact with Mods or admins about different issues including abuse (e.g., mailing mods or reporting abuse to mod/admin).
- **Privacy control:** Ability to control one's privacy, including controlling the persistence of as well as others' experience of one's content and account (e.g., blocking, the ability to create multiple throwaway accounts to keep anonymity and avoid association with the primary account or privacy aspects of general platform features such as a lack of edit history on posts and comments).
- **Tracking:** Ability to track and learn about someone else's activity (e.g., the fact that Reddit accounts have unconfigurably public post comment histories).

The categorization was devised by considering the primary purpose for which the users used these features in our dataset. We analyze these features in terms of the types of threats and harms they enable and the ways they make moderation more difficult in Section 5.2, where we also provide general lessons for thinking adversarially in the design of such features. We hope that our analysis will aid designers in more effectively evaluating and threat-modeling social systems against different types of adversarial misuse of platform features.

4.1.2 Complexity of Policy and Rules

Echoing some past work [38,45], moderators often discussed situations where rules or policies aren't well-defined, or where their details or enforcement are subjective. For instance, one moderator said: *"The primary issue in this situation is that nobody, including the administrators, can clearly define or provide evidence of brigading.* [5] *This remains a persistent source of disagreement between the moderators and the administrators."* Reddit's policy on sharing personal information [6] was another source of confusion: *"Are we required to delete a bad review detailing a user's experience with their doctor? We were threatened with a lawsuit from someone claiming to be the doctor's attorney and one of our mods ended up having to find a lawyer to advise us. Handling this situation should not be the job of unpaid moderators!"*

A lack of admin clarification of policies was also a common complaint. For example, one wrote: *"When admins refuse to provide clarifications on the specific rules, it's unreasonable to expect people will adhere to them."* That said, a few moderators acknowledged the complexity of defining concrete policy, especially in the complicated landscape of online misconduct and harassment. For example, one moderator said: *"This is one aspect where I agree with admin regarding the necessity for making contextual judgments. Policies can't encompass every scenario, leaving room for someone to claim the rule is unclear and disagree with its application somewhere."*

Others noted that too much specificity in policy may enable harassers to follow the letter of the policy while subverting its spirit: *"r/<redacted> is a prime example of why admin avoids giving a straightforward definition as they did bare minimum to comply with it. Yet they would still facilitate their community to engage in invading, manipulating, and harassing content/users they disagreed with. When mods were told to forbid linking to content, they complied but direct people on how to find that same content by other means."*

The issue of opaqueness in moderation policy is not exclusive to Reddit. Singhal et al. found a lack of clear guidelines for content moderation policies on all the major social media platforms. Moreover, there is no consensus among social media platforms on what type of content should be moderated for some categories, such as sexual content [50].

4.1.3 Issues with AEO and Admins

Anti-Evil Operations (AEO) and Reddit administrators moderate Reddit by enforcing content policies, addressing rule violations, and taking action against rule-breaking content or users. Though they seek to support mods in maintaining a safe and respectful online environment, mods frequently mentioned challenges when dealing with AEO and admins.

Lack of Response and Action: A major concern of moderators was admin's perceived slow or inconsistent response to reports and queries. They complained of a lack of responses or of generic assurances such as *"we'll investigate"* from Reddit, even in situations demanding urgent attention, such as reports involving suicide, sexual abuse of minors, and organized attacks and harassment. One moderator wrote:

"Has anyone encountered issues with admin not replying to private messages? I mod a mental health subreddit, and attempting to contact admin since last week about a user who is actively encouraging our users to self-harm. One would expect that a serious matter such as this would receive a quick response, but it's now been over 72 hours with no response."

[5] Coordinated activity of a group of users on social media or online communities to manipulate or disrupt discussions, forums, or platforms, usually by flooding content with a specific agenda or engaging in harassment.

[6] The policy reads: *"Respect the privacy of others. Instigating harassment, for example by revealing someone's personal or confidential information, is not allowed. Never post or threaten to post intimate or sexually explicit media of someone without their consent [4]."*

Feature	How Misused	Example Quotation
Community Engagement Features		
Create account (Username & profile pic)	Create account with offensive username or profile pic	*"One account is following people with a slightly obscured racial slur for a username & a profile pic showing a black man being executed."*
Post & comment	Post and comment harassing content, misinformation, scam, etc	*"We are getting overwhelmed at r/<redacted> because of the World Series and at least one troll is actively creating new reddit accounts to post racial messages. It is getting difficult to manage."*
Flair (post categories & user tags)	Use harassing text as flair	*"A subscriber decided it would be a great idea to fill their flair with offensive, racist and homophobic language."*
Crosspost (sharing posts on subreddit)	Crosspost posts to trolling subreddits; Crosspost harassing posts to a subreddit	*"He is crossposting many of our sub's threads to troll sites. He doesn't even try to hide it because surprisingly, all of it is legal. There is no way to ban offensive cross posting. He crossposts then invites his troll buddies to harass me or other community members."*
Create subreddit	Create subreddits with the sole purpose of harassing	*"Individuals in my area who run hate groups on facebook are creating subs that direct users to sites spreading hate and misinformation."*
Direct message (DM)	Send harassing content	*"A user in our sub pointed out some hateful language on a post and now getting DMs filled with hateful/incendiary/harassing words."*
Gilding (purchase and give award)	Award people with account containing offensive username	*"A user with an offensive username now gilding others so that I remove the comment they gilded as it shows their offensive name. I doubt Reddit is going to take action against accounts that buy gold."*
Vote (upvote & downvote)	Mass downvoting	*"After investigating, we believe 2-3 former members of our sub are chronically downvoting. Now the majority of the posts on our subreddit's main page have no upvotes or negative karma, and someone pointed out that that it portrays the subreddit in a negative light."*
Mod/Admin Engagement Features		
Modmail	Send harassing content, spam Modmail	*"We were recently bombarded by a single user who sent us the same harassing Modmail 30 times within a minute."*
Report	False reporting posts/comments; Fill the 'text field' on the report feature with harassing text	*"I moderate a controversial sub & lot of the reports have anti-semitic, violent, threatening, homophobic and transphobic language in it. "* *"Currently, any user (including those banned for harassment) can false report your posts and cause site-wide suspension of your account."*
Privacy Control Features		
Edit	Harass and then edit content to evade penalty	*"I recently deleted an offensive image. The poster began commenting that I am a bully and should not be a mod and when I went to alert other mods, they changed their texts to a more rational response."*
Delete	Delete post or comment and/or delete account after harassment to avoid penalty	*"A user waits 3 days after everyone has seen his abusive posts, then deletes any trace of doxxing from his accounts. Admins do nothing after I submit a report because there is no evidence of harassment."*
Alt/throwaway account (extra or temporary account)	Create account to harass and evade ban	*"A user created a new account and spent 6 hours urging others in the sub to commit suicide. We used automod to pseudo-shadow-ban them, but they simply created a new account and did it again. "*
Block	Harass and block someone so they can no longer view or respond to discussions	*" Someone post an insulting comment and then immediately blocks the accused. This gives the impression that the accused user is wrong and does not want to respond to the insults made against them, but in reality, they are unable to even see the comment."*
Tracking Features		
User post & comment history	Follow someone throughout Reddit to harass	*"Someone's stalking me, they check my post history and inundate all of my posts with offensive comments, even resorting to doxxing by using personal info they've gathered about my real life."*
Follow	Follow with harassing usernames, profile pics; follow to track targets posts to harass	*"New users in our teen subreddit are being bombarded by followbots with names related to pedophilia. 13 year olds are being notified on their cellphones that convicted pedophiles are following them."*

Table 1: Summary of platform features abused for harassment from our dataset

Moderators also expressed their frustration with Reddit's failure to enforce its terms of service and policies when it comes to content and individuals who blatantly violate these rules. This encompasses trolling-focused subreddits, subreddits associated with organized attacks, offensive content, and comments targeting specific communities, among other violations. For example, one moderator shared:

"I manage multiple large subreddits and frequently see blatantly violent anti-Semitic or Islamophobic comments. Even though I handle these at the subreddit level, I find it perplexing that when reported to admins, they constantly say that the content does not violate Reddit's policies on hate speech."

One moderator also shared an example of how this lack of action escalated to hate and harassment offline:

"Look at r/<redacted> for an example of a subreddit illustrating admin's reluctance to ban a community that exhibited blatant hatefulness, destructiveness, and disruption on Reddit and contributed to real-world violence by right wing extremists. Admins only banned the sub after it moved off-site to continue their extremism elsewhere."

Mods also discussed a lack of effectiveness of admin actions. For example, one popular admin sanction is the shadowban, which allows a banned user to post and comment, but their activity is invisible to others, with the goal of not revealing the punishment to the banned user so they cannot evade it. Nonetheless, mods observed that shadowbanned users can still engage in harassment:

" Shadowbanning a user that utilizes private messaging to run scams or sell drugs is ineffective. Shadowbanning a user who inundates mods with hundreds of messages is also ineffective. I know first-hand that getting admin to address explicit violations of content policy is not possible when the account has already been shadowbanned. It's incredibly frustrating."

Some moderators claimed that when a single user is reported multiple times, Reddit often investigates only the first: *"I flagged two comments from the same user. Reddit decided to delete one of them, but declined to look into the other & it wasn't removed. They said: "We've already investigated this user based on another report about other content. After investigating, we determined that the reported user was in violation of Reddit's Content Policy and took action accordingly."*

Moderators don't seek only admin's practical support; many also valued acknowledgment and appreciation of their work: *"Being heard and recognized means a lot. The fact that admins do not acknowledge these problems is a major issue. Simply making the time to listen and acknowledge the issues with genuine interest and intention to help holds immense value and can go a long way in reducing frustration."*

Lack of Transparency about Admin Actions: Moderators expressed a desire for greater transparency about the actions that admin or AEO take or the reasons for those actions. Some wanted transparency in order to form a better model of unclear or subjective rules (see Section 4.1.2) and better align their moderation with it. For instance, after AEO removed a post that a moderator had approved, they said: *"If there was a legitimate reason for the removal, I would like to be informed of it. This way I know what we did wrong in our approval process and make better choices moving forward."* Others expressed surprise when receiving feedback that users they reported were banned, noting that typically they received only boilerplate with no details about what action was taken: *"I am fed up with constantly receiving this message: "After investigating, we've found that the reported content violates Reddit's Content Policy and have taken action."* Such frustrations were also observed in the prior research on moderation transparency in different platforms [38, 53].

Finally, some were frustrated by the way a lack of transparency made it difficult to contest decisions made by AEO or admin, as the appeal form requires an explanation of why it is an error and why the instance actioned was not a violation of Reddit policy. This is particularly important in instances where moderators are targeted by false reporting to ban their account. One moderator said: *"I was banned for 'harassment' with no explanation of what I did. I appealed and the decision was revoked, but the process was difficult considering I had no idea what I had supposedly done."*

Prior studies emphasized the significance of making moderators' actions [29, 31] and automated content moderation [15] transparent to users and content creators to foster trust and engagement. Our results expand to also examine the role of transparency between the platform and volunteer moderators.

Inconsistent and Inaccurate Actions: Moderators frequently complained that AEO's actions appeared arbitrary and inconsistent, both in failing to address violations and in taking action without cause. Such inconsistencies in action from algorithmic moderation systems have also been observed in other platforms such as Facebook [26, 57].

One moderator wrote: *"Taking isolated messages out of context and imposing penalties based on them undermines the value of our role as moderators, portraying us as ineffective and dispensable."* This quote touches on two themes. First, moderators prided themselves as experts of their community and moderating contextually. Second, one of the problems with inconsistent platform actions is that they see it as undermining their ability to moderate effectively.

Inconsistent platform actions also frustrated moderators who saw them as increasing their workload through the need to appeal erroneous suspensions or request that posts be reinstated. Some found the process for doing so particularly cumbersome: *"I should not have to put so much effort (numerous modmails) just to review an AEO removal."*

Unwarranted actions against moderators were seen as particularly harmful, both because a banned moderator can't moderate (leading to more toxicity in their community), and because some moderators were afraid to take moderation actions after seeing others be banned for doing so:

"A mod in my community was banned, ironically, for reporting abuse of the report button. Our users abuse the report

button regularly, and honestly I'm nervous to report them now, since apparently we can be banned for doing our jobs."

Moderators articulated various theories of why AEO exhibits these behaviors, such as believing it consists of third-party contractors in another country or a poorly designed automated bot. Actions perceived as inaccurate may thus aggravate a lack of transparency around admin actions (by creating & reinforcing potentially incorrect theories). Conversely, a lack of transparency around how AEO works may contribute to perceiving its actions as inconsistent or inaccurate.

4.1.4 Issues with Moderation Tools/Features

Reddit's moderation features include tools such as post/comment removal, user banning, content approval, etc. Two major themes about these tools emerged: misunderstood or unknown tools and inadequate feature sets.

Misunderstood or Unknown Moderation Tools: Moderators were sometimes entirely unaware of the existence of some moderation features, learning about them in response to posts in r/ModSupport or r/modhelp. Others knew about features but misunderstood their behavior. For example, one mod deleted a post containing a link to non-consensual intimate imagery, assuming that removing the text of the post would render the link inaccessible. However, such links remain active on Reddit, which they learned only when told by the victim. In another case, a moderator banned an offender from a subreddit, believing they would be entirely blocked from the subreddit. The mod later discovered that the offender could still view the subreddit, report content, and even give awards, leading to continued misuse of these features. Such misunderstandings may be due to incomplete documentation; Reddit's website outlines what a banned individual cannot do, but doesn't list actions they can still take.

A key action moderators take is to report violations of Reddit's central policies to the platform. Reddit offers several avenues for reporting offensive behavior, such as utilizing the report button under posts or comments, utilizing a dedicated report form, submitting a request for inquiry through Reddit help, and contacting the administrators via Modmail on the r/ModSupport subreddit. Moderators discussed all of these options, but were often unclear on how they differed and when to use them. Further, when reporting content, mods need to choose a category under which to report, such as "This is spam", "This is abusive or harassing", "Minor abuse", and more. Some mods felt there was a lack of clear explanation of the meaning or scope of these categories, which they said led to incorrect use and less effective responses to reports. For example, one moderator pointed out that the category "report abuse" could be misleading, as some individuals might interpret it hastily as *"I want to report this abuse"* rather than *"This user is abusing the report button"* (i.e., by reporting non-problematic content). Such confusion can have major consequences: this moderator attempt to report an abusive user comment, but instead inadvertently reported another mod who had already reported the same comment, resulting in the suspension of that moderator for misuse of the report button.

Another moderator noted that clarity and consistency of report categories could also serve a role of communicating site standards and Reddit's consistent enforcement of them: *"By presenting users with categories of what constitutes unacceptable behavior, Reddit can enhance clarity, demonstrating its commitment to upholding every aspect of its code of conduct and reinforcing these rules with its community members."*

Moderators also discussed uncertainty about how much and what kind of evidence is needed for reporting. For example, a common form of harassment involves mass (false) reporting of threads in a subreddit, which wastes moderator time by filling the modqueue with gratuitous reports. Moderators were unsure whether to report each false report or just a few, and whether or not to clear the modqueue or if they should leave it as evidence for administrators. In this case, admins responded to say that it is fine to report just a few cases of spam abuse and then clear all the abusive reports from the modqueue. Nevertheless, we saw that many other inquiries about misunderstood and under-documented moderation features went unanswered by admin. Past research found similar misunderstandings and confusion about the content moderation and appeal process among users on social media platforms [58].

Inadequate Features: Moderators are often unsatisfied with the lack of adequate features to prevent abuse and action offenders, even though they are held accountable for the safety of their community. They often named specific toxic behaviors, such as ban evasion and misuse of the report and block features, for which features were particularly insufficient: *"How do you think we're going to moderate block abuse if we can't see when people have blocked each other, and just have to take their word for it?"*

Reporting to admin was seen as one of the few options available in such cases, but this was more difficult when nonstandard approaches were used to harass, such as paying to gild a post while using an offensive username: *"The report categories don't include offensive username and/or gild giver name)."* Furthermore, deleted accounts can't be reported. Mods described harassers circumventing bans by creating a new account and deleting the old one. One said:

"We can't report most usernames through the report form due to the error it returns: 'this account does not exist'."

The inability to provide contextual information while reporting offenses frustrated some mods, who considered this particularly important when the harassment was in another language or used relied on tropes or dog whistles. For instance, one moderator said: *"I once encountered a subreddit promoting hate against the queer community and it was near impossible to report because most things posted were subtle. They weren't so obvious like 'trans people go die'.*

At times, limits in place to resist abuse can backfire. For example, Reddit limits users' rate of reporting to mitigate

report abuse, a limit that also applies to mods dealing with a surge of toxic content: *"When you have an uptick in report abuse in the queue, you have to spread out reporting across multiple moderators to quickly clear the queue."* Similarly, Reddit does not accept images as evidence for reports since they could be edited. However, offenders can edit or delete posts and comments at any time to conceal their actions. According to Reddit's 2013 privacy policy, Reddit retains only the most recent versions of posts and comments, making it nearly impossible to detect such offenses. This policy has not been mentioned in subsequent privacy policy updates, but many moderators believe that Reddit still follows this policy.

4.1.5 Non-Reddit Authorities

In extreme cases of harassment, mods discussed contacting authorities such as law enforcement, and internet service providers (ISPs). However many mods believed that law enforcement has a high bar for acting on online abuse, a claim justified by past work on intimate privacy [17]. Others saw the platform as a blocker to such action after contacting police about doxing: *"I was told by police that they've asked for information from Reddit, but it could be two to six months before they're sent the necessary info. This does not appear to be high priority for them."* In other cases, the international nature of the internet was the problem.

"The problem I'm facing is that the user at risk are located in the United States, while I am in Germany. If I reach out to the local authorities in Germany, they would be unable to take any action due to jurisdictional constraints."

Moreover, mods noted that law enforcement and ISP practices vary from country to country. While online harassment is treated seriously in some regions, it may not be elsewhere. Mods felt that without other support, victims often have two choices: leaving the community or enduring harassment.

4.1.6 Mod-Targeted Harassment and Burnout

Moderators on Reddit are more prone to targeted harassment due to their role in enforcing subreddit rules and decisions, their increased visibility, and the potential for disagreements with users, which often escalate into personal attacks in the relatively anonymous online environment [11]. Moderator mentioned harassing messages (i.e., death and rape threats toward mods and their families, threats of physical harm, threats of hacking and doxing, etc.), organized targeted attacks of false reporting and downvoting all their posts and comments, and being hacked, doxed, or stalked in campaigns that sometimes escalated to moderators' other social media accounts. In extreme cases, harassers persisted over months or even years.

"After banning a user in my sub he sent me a private message containing my phone #, family address, and other personal info, saying they would 'pay me a visit'. I moderate a local sub so the threat felt very real. I am unashamed to say I am still afraid of this person and I don't think I'll ever reach closure around this situation nor will I feel at peace."

To minimize these risks, some moderators discussed using separate accounts for moderating and personal Reddit activities. Some wanted the ability to seamlessly switch between their dedicated moderation account and regular user profile or recommended that Reddit should conceal moderators' identities when interacting with rule violators.

In addition to targeted harassment, moderators see more offensive and harassing content and behaviors, which leads to emotional burnout [52, 59]. One moderator said: *"Automod prevents the public from seeing this stuff, but I still see it."* Previous research with online content moderators [51] and community moderators [21] showed that moderators cope with burnout by taking various approaches, including peer support, separating moderation from home, engaging in physical exercise or distraction, and using tools such as muting audio, ceasing immediate viewing when content was confirmed as sexual abuse content, etc. to minimize impact.

Special Events: Special events relevant to a particular community can create temporary increases in toxic content, leading to increased moderation burden both practically and emotionally. Discussing Pride Month, a moderator of an LGBTQ community asked: *"Is anything being done to address the significant increase in hate and harassment towards LGBTQ users, moderators, and subreddits?"* Current events could lead to similar but unplanned surges, as described by a mod of a geographically focused subreddit after a local mass shooting: *"It turned out to be unexpectedly demanding in terms of time and emotional toll. We were immersed in the tragedy without any chance to take a break, and piled on top of that was the overwhelming toxicity from brigaders."* Reddit organizes mod reserves, which provide temporary extra moderators for such events, but moderators exercise caution when considering this option, since so much of moderation depends on understanding subtle context or familiarity with the community's culture and values, reinforcing findings from past work [43].

Some attributed inaction to financial causes. *"Reddit possesses the financial resources to offer an immediate response for mod support in local surges like these. They opt not to since mods will just suffer and do it without pay."*

4.2 Complex Trade-offs

Moderators identified complex tradeoffs involving conflicting pressures between welcomingness, privacy, free speech, and other factors when working to combat hate and harassment.

Accessibility vs. Harassment: Ban evasion—in which offenders create new accounts and continue problematic behavior—was a commonly discussed challenge. One common mitigation is to establish a minimum account age or karma requirement to prevent new accounts from participating in a subreddit. However, this also restricts genuine (i.e., non-ban-evading) new users from contributing, which is a

serious concern for some moderators: *"Because we mod a local sub, we do not want to set age or karma limits. It would prevent new posters from posting, thus creating an unwelcome environment for folks who, say, recently moved to our city."*

Another approach to reducing harassment is to make a community private, such that only users approved by moderators can view, post, or comment. This option seriously limits accessibility to new users, as well as necessitating extra mod effort to manually approve each user. *"Changing our sub to private was an option that we thought of, but the labor needed for that and the risk of losing members was of major concern. We would possibly lose existing members or genuine new subscribers who want to talk, grieve, or find info. "*

Harassers also sometimes misuse features to induce this tradeoff. Mods described harassers who create accounts with offensive usernames and gild target's posts, which causes the platform to display the gilder's (toxic) username alongside the post. Moderators cannot remove the gild giver's username from the comment—their only moderation options are to delete the entire comment, which silences the target and makes the community less accessible to them, or leave the post up still branded with the harasser's username.

Privacy vs. Harassment: Reddit allows users to remain relatively anonymous, which can be both a strength and a weakness. Anonymity can protect privacy and allow for open discussions [12, 39], but it can also shield harassment, hate speech, and trolling. A popular way to maintain anonymity on Reddit is by creating a throwaway account. Many people use throwaway accounts to ask sensitive questions, confess secrets, or share stories that they don't want to be associated with their regular Reddit account. However, throwaway accounts are also misused for harassment and evading bans. One way to prevent this is to verify the throwaway account to confirm that they are legitimate users using a formal verification method. However, such verification can undermine the purpose of using a throwaway account. One mod suggested: *"If you are having an issue with new accounts trolling your community, why not require them to message you from an established account to prove their throwaway is legitimate?"*. Another replied: *"I hope that solutions exist to preserve anonymity if a user has genuine fears about their Reddit identity being exposed. I don't believe throwaway accounts should trust or rely on mods to keep their identity confidential."*

Reddit stores only the latest version of edited posts or comments. Mods described abusers who post offensive comments, then edit them to appear benign to evade punishment by concealing evidence. However, this lack of edit history is also seen as a valuable privacy feature:

"The privacy implications could be drastic if this change was enacted. Redditors could more easily gather another's personal information. Ongoing disclosure and editing is part of Reddit's approach to privacy, which is why you can edit posts and comments even when your account is suspended."

Finally, for the sake of privacy and to deter moderators from misusing their authority to suppress legitimate reports of misconduct, Reddit doesn't reveal who reported an issue to moderators. Unfortunately, mods noted that this also protects those who abuse the report button, and many wished for a system that could hold spam reporters accountable without disclosing genuine reporters' identities to mods.

Transparency vs. Harassment: Similar to prior work [33], mods noted a tradeoff between transparency and harassment mitigation. Moderators discussed the need for clear and concrete rules and transparency on how the platform makes decisions on violations. Though this information would help moderators take moderation actions that align with platform policy, excessive transparency can inadvertently empower harassers by revealing patterns that help them evade detection and circumvent moderation measures [20, 33, 34]. One mod described a toxic sub with *"[a] desire for an explicit rule not to better prevent harassment, but so they could find a loophole. They followed the letter of the spelled out rules but still allowed harassment."*

5 Discussion

Our discussion draws out to major themes from our results. First, we discuss the understudied relationship between mods and admins, examining it in terms of power dynamics and placing moderator behavior and understanding of the platform in that context. Second, from our results detailing feature misuse for hate and harassment, we generate a taxonomy of such misuse, analyzing the ways feature misuse strengthens harassment and makes it more resilient to moderation. We then articulate design themes for creating platform features that are more difficult to misuse in these ways, in the process arguing for the value in this context of taking a security mindset-based approach to platform design.

5.1 The Relationship of Mods and Site Admins

In our investigation, the relationships between mods and admins came up often. In broad terms, both admins and moderators have similar goals to deter targeted attacks, harassment, and other negative behaviors on Reddit and foster a positive and safe environment for users. However, we observed complex and often discordant relationships between them.

Moderators' perceptions of their relationship with admins crossed many emotions: frustration about insufficient communication, confusion about unclear guidance, anger about perceived inconsistent or inappropriate actions, and fear about the possibility that they, their users, or their whole community might be punished or banned unfairly. Mods often view admins as capricious, unreliable, or arbitrary yet powerful entities. These inconsistent and arbitrary actions of admins affected moderators' ability to effectively moderate their community, the safety of community members, and the perceptions of Reddit as a place for positive engagement. For in-

stance, moderators get inaccurately banned for reporting violations, their posts arbitrarily get removed without explanation, and so on, affecting their ability to interact with and ensure the safety of their community. Despite these challenges, moderators often find themselves in a position where they need to be reliant on admins. We have observed this dependency on admins both in the high frequency with which moderators advised each other to escalate issues to admins and the frequency with which admins came up when discussing challenges. This dependency underscores the significant power disparity between admins, who wield ultimate authority, and moderators, who possess lesser status and authority. Moderators feel compelled to engage with admins frequently to ensure the growth and safety of their community despite many of the challenges they face being a result of admin actions. Yet, we observed moderators seldom receive reasonable responses or support, if any, from the admin. We encourage future work to explore the power dynamics of such relationships as they vary across platforms with different moderation structures and approaches, especially given that power relations of this sort are particularly likely to harm marginalized groups with less status and power, to begin with.

The relationship between moderators and admins may also be affected by the inconsistent role of admins in moderation. Massanari asserted, “There seems to be a deep reluctance on the part of the administrators to alienate any part of their audience, no matter how problematic, as it will mean less traffic and ultimately less revenue for Reddit [41]" after investigating the response of Reddit to a public case of non-consensual distribution of intimate images of a celebrity within the platform. Reddit administrators responded to the case by saying they: “feel it is necessary to maintain as neutral a platform as possible and to let the communities on Reddit be represented by the actions of the people who participate in them. [10]" Yet, we observed moderators perceived administrators actions as particularly untrustworthy and unfair for niche and marginalized communities. One possible avenue for future exploration could be clearly distinguishing between admins’ and moderators’ involvement and responsibilities in moderation. This would involve specifying the hierarchy of community moderation, identifying which parties are involved, outlining their duties, and establishing protocols for intervention. For instance, depicting the circumstances under which administrators would intervene and when they would defer to and respect community moderators judgment in matters of moderation.

Finally, in the midst of the tension between mods and admins, we noticed genuine appreciation from moderators when support was provided by admins and their recognition of moderators’ efforts. More attention from platforms and admins in publicly acknowledging moderators’ contributions, whether through shout-outs in newsletters, badges, or other forms of recognition, may help improve the admin-mod relationship.

5.2 Designing Against Adversarial Misuse

Table 1 summarizes the platform feature misuse moderators identified. The research team analyzed each type of feature misuse in terms of the abuse types it enabled (informed by our dataset) from Thomas et al.’s taxonomy of online hate and harassment [55]. Table 2’s left-hand seven columns summarize this analysis. Through an iterative process reviewing the coded segments for the ’feature misused for harassment’ code, two of the researchers arrived at eight capabilities enabled by the feature misuse. The same researchers annotated different types of feature misuse to determine whether or not they enabled any of the capabilities in our dataset, with the rest of the research team validating the annotations. The adversarial capabilities that emerged from the analysis include (1) amplifying harassment, (2) silencing/narrative control, and (3) making harassment more resilient to moderation (divided into six capabilities). Finally, based on our systematization and responding to Freed et al.’s call for frameworks to design against UI-bound adversaries [24] we articulate general lessons for designing less abusable features to guide designers of not just Reddit but all platforms.

5.2.1 Amplify Harassment

Some features empower potential adversaries to execute large-scale attacks or aid in escalating harassment. Crossposting enables attackers to share posts from one subreddit to another, expanding their reach to a broader, potentially hostile audience. This, in turn, can lead to instances of brigading and a heightened level of harassment directed at the original poster or the target subreddit. We have also observed malicious individuals establishing subreddits to orchestrate coordinated attacks against specific communities. Additionally, adversaries can generate unlimited alternative or throwaway accounts to intensify harassment directed at the target. Adversaries can follow targets, and whenever targets post, it appears in their feed, providing them easier and immediate access to the target’s posts to harass. Finally, post and comment history escalate harassment by enabling attackers to follow around the target and harass them everywhere they engage across Reddit. Beyond Reddit, adversaries can amplify harassment on any platform, allowing users to create radicalized communities without proper platform moderation.

5.2.2 Silence/Control Narrative

There are several methods commonly employed to stifle opposing voices on Reddit. One approach involves reporting the content as abusive and triggering platform tools like AEO to remove the content or suspend the poster’s account. Another widely used tactic is to downvote posts or comments, reducing their visibility to the audience. We have also observed adversaries employing a unique method to silence by creating offensive usernames and gilding specific comments. It forces

moderators to remove the comments entirely, as there's no way to remove only the offensive gild givers' names from the comments. Moderators have also discussed how adversaries misuse the block feature to disrupt a target's ability to participate on Reddit. For instance, if a user higher up in a comment thread blocks you on Reddit, you lose the ability to reply to any users below you, including those responding to your comment. Adversaries exploit this feature to lock out the target from the conversation. Moderators have also highlighted the consequences of such feature misuse in terms of controlling the narrative of a conversation. For example, someone can disseminate misinformation more effectively by silencing active Redditors who typically counter misinformation with factual information, taking advantage of the aforementioned features.

5.2.3 Making Abuse Harder to Moderate

Past work has viewed abuse from various lenses, such as vulnerability (e.g., of specific populations), design (e.g., nudging to reduce toxic behavior), human and cultural factors (e.g., the need for humans in the moderation loop), and prediction (e.g., automated detection of hate speech). We add another lens that has been little used to date: harassment as a *security problem*, wherein intelligent adversaries (harassers) misuse features to harass and to resist moderation. 13/16 of the misused features we identified could be used to make abuse more resistant to moderation. We argue for the value of this lens in borrowing existing design tools from technical security work—most fundamentally, the adversarial perspective-taking of the security mindset [48]—to inform the design of platform features hardened against adversarial misuse. The rightmost columns of Table 2 summarize the ways moderators in our dataset witnessed feature being misused to resist moderation of abusive behavior. The subsequent paragraphs elaborate on these ways. We then present a set of design themes to inform designers working to design or re-design platform features that are resistant to adversarial misuse by harassers resisting moderation.

Incapacitating Moderators: In this adversarial model, attacks directed against moderators can be viewed as incapacitating them by consuming their time, focus, or emotional energy or simply by getting their accounts banned. Feature abuse here includes spamming modmail, filling the modqueue with spurious reports to make it tedious to act on real ones, or trying to get moderators banned via false reporting. Any system that relies on volunteer or professional human moderation is potentially vulnerable to similar denial-of-service attacks. Beyond Reddit, a variety of types of features might be misused to this end, such as direct messaging inboxes, notifications, moderation queues, and any kind of automated or user-report driven moderation system that might be adversarially redirected towards spuriously banning moderators.

Non-Standard Delivery Mechanisms for Toxic Content: Harassers embedded toxic content in "non-standard" locations or fields, such as usernames or flairs, or delivered it to targets in "non-standard" ways, such as by following a target or gilding their messages to their insulting username in notifications or alongside targets' posts. These "non-standard" approaches contrast "standard" approaches, such as posting, commenting, or private messaging. Furthermore, malicious users misused their ability to edit content by including offensive language after the moderators approve their innocuous posts. Such non-standard mechanisms may make moderation more difficult for a variety of reasons, including a lack of first- or third-party tooling for efficiently viewing and moderating such complaints or a lack of platform moderation features for reporting such content (e.g., one moderator complained about the fact that there is no option to report users for having hateful usernames). Beyond Reddit, such non-standard delivery mechanisms might exist anywhere users can embed content (e.g., profiles, tags, customized URLs, profile pictures, emoji reactions, etc.) and anywhere they are able to cause content they control to be shown to targets or audiences (e.g., notifications, replies, re-blogging, profiles, etc.).

Hiding Content from Moderators and/or Admin: Moderators discussed a variety of ways in which harassers misused features in order to hide toxic content or evidence of harassment from moderators and admins, including editing a toxic post to be innocuous (or deleting it) after the target has seen the toxic content, sending toxic content via direct message, cross-posting a target's post to a trolling subreddit or creating a subreddit to harass a particular subreddit and it's users. In many cases, these misuses rely specifically on features of the platform that may be described as privacy features, such as the fact that the history of edited or deleted posts are not preserved, inducing challenging tensions for designers.

Hiding Content from the Target: In this attack, the content is hidden not from mods or a broader audience, but from the target themself. The harasser posts a toxic comment replying to the target, then immediately blocks them, making their comment invisible to the target and preventing them from responding. Mods noted that this attack not only prevents the target from responding to toxicity, but also portrays them to others as unable or unwilling to do so. They also noted the potential to create a seeming "consensus" about misinformation or toxic beliefs by preventing those most likely to debunk or combat those ideas from responding. Blocking is an archetypical privacy feature, and the design of Reddit's blocking feature is natural and expected, yet this attack is a striking example how protective features can be leveraged by clever attackers. Beyond Reddit, any feature that allows users to control the visibility of content, such as private posts or accounts, circles, visibility settings, and more may implicate this type of attack. One example can be seen on Twitter, where quote retweeting from a private account has come to be widely associated with toxic or harassing behavior [28].

Hide Attacker Identity: Two features enabled harassers to hide their identities more easily: the deletion of posts/comments/accounts and Reddit's liberal policy with

Misused Features	Types of Abuse from [55]							Adversarial Capabilities (this paper)							
										Resilience Against Moderation					
Community engagement: Ability to create, or engage with community. *Mod/Admin Engagement:* Ability to interact with Mod/Admin. *Privacy control:* Ability to control one's privacy, including controlling the persistence of as well as others' experience of one's content and account. *Tracking:* Ability to track or learn aboutsomeone else's activity.	Toxic Content	Content Leakage	Overloading	False Reporting	Impersonation	Surveillance	Lockout	Amplify Harassment	Silence/Control Narrative	Incapacitate Moderators	Non-Standard Delivery	Hide Content from Authority	Hide Content from Target	Hide Attacker Identity	Leverage Platform Incentives
Community Engagement Features															
Create account (usernames/profile pic)	✓	✓			✓						✓				
Post/Comment	✓	✓	✓		✓										
Flair (post & user)	✓	✓									✓				
Crosspost	✓	✓	✓					✓				✓			
Create subreddits	✓	✓	✓					✓				✓	✓		
Direct messages (DM)	✓	✓	✓									✓			
Gilding	✓	✓							✓		✓				✓
Voting			✓						✓						
Mod/Admin Engagement Features															
Modmail	✓	✓	✓	✓						✓					
Report			✓	✓			✓		✓	✓					
Privacy Control Features															
Edit posts/comments (no history)	✓										✓	✓			
Delete posts/comments (no history)												✓		✓	
Delete accounts														✓	
Alt/throwaway accounts allowed			✓		✓			✓						✓	
Blocking									✓				✓		
Tracking Features															
User post/comment history						✓		✓							
Follow (a user)	✓	✓				✓		✓			✓				

Table 2: Taxonomy of platform feature misuse for hate and harassment. The categorization on the left-most column depicts the primary purpose for which the users used these features in the context of our dataset. A ✓indicates the feature being exploited to conduct an attack, amplify harassment, silence the target, or develop resistance against moderation in our dataset.

alt/throwaway accounts to keep anonymity and avoid association with the primary account. To make matters worse, when a malicious user deletes their (alt/throwaway) account, the harassing posts and comments published from that account persist and are not automatically removed. More than our other categories of moderation resistance, this one induces difficult tradeoffs, given the particular value of privacy and anonymity, as discussed in Section 4.2. Beyond Reddit, any features that allow anonymous, pseudonymous, or multiple-username participation such as guest posting, allowing multiple accounts, or the ability to change usernames or other account details might raise similar challenges.

Platform Incentives: Above, we discussed gilding from accounts with toxic usernames as a non-standard delivery mechanism. Some mods saw this as effective not only because it is non-standard, but also because gilding is one of Reddit's revenue streams. They hypothesized pessimistically that monetary incentives would make the platform less likely to take action against accounts that spend money on gold, even for hostile purposes. Whether or not it changes Reddit's behavior, this certainly reduced trust in the platform by moderators. Given the importance of the relationship between mods and the platform shown in our results, such loss of trust has serious implications for user safety. Beyond Reddit, any feature perceived by users as creating incentives for the platform, including direct revenue streams (e.g., advertising, Important Blue Internet Checkmarks, API fees, premium subscriptions, etc.) or indirect ones that support revenue (e.g., anything that

increases engagement) may implicate such concerns.

5.2.4 Designing Platform Features to Resist Misuse

Our results show that many of the toughest challenges moderators face result from the misuse of features, and we have characterized above the particular ways that feature misuse works to enhance toxic behavior by amplifying its effects, silencing targets/controlling the narrative, and by hardening it against moderation in a variety of ways.

Freed et al. called for HCI researchers to create design frameworks for resisting UI-bound adversaries by adopting the abusers viewpoint to identify avenues for abuse [24]. Expanding their notion of UI-bound adversaries outside of the space of intimate partner violence, we take up their call to ask: how can we design platform features that "limit systems' abusability... while maintaining, or minimally impacting legitimate user experience" [24]. In analyzing the properties of feature misuse described above, we have derived four themes that can inform designers working towards this goal.

The Security Mindset: Popularized by security writer and cryptographer Bruce Schneier, the security mindset is a common attitude towards systems and their design that is broadly valued among technical security practitioners. Just as Freed suggests taking on the abuser's viewpoint, the security mindset emphasizing thinking like an attacker to reveal unexpected uses of a system or its features and enable their design or re-design into a less abusable form. We argue that this mindset is a perfect fit for hardening the features of online platforms like Reddit against misuse by harassers. Though issues of hate & harassment are often thought of under headings such as "Trust & Safety", they can be viewed, like technical security challenges, as potentially adversarial situations. Thinking of harassers as adversaries is not the only way to view them, and we are not arguing against past work that rightfully believes that toxic users are not inherently toxic and deploys techniques such as nudging to encourage better behavior [9, 40]. But the evidence from our results suggests that harassers frequently misuse features in ways that are explicitly adversarial, in that they intelligently attempt to make moderators' jobs more difficult, more burdensome, or subject to catch-22 tradeoffs. Though our taxonomy includes only the attacks and adversarial capabilities of features discussed by the moderators in our dataset, it is also important to consider how the exploitation of various features in combination gives an attacker even greater adversarial capability. For instance, an attacker might create multiple throwaway accounts with offensive usernames and harass the target by commenting on their post and following them. It forces the target to triage numerous notifications or comments, hindering their ability to engage fully online while also subjecting them to harassing content. We suggest that platforms would benefit from incorporating security mindset at all phases of the design process to identify and mitigate opportunities for adversarial misuse. As our results show, a good existing platforms might learn much from their users and, especially, their moderators, who are intimately familiar with misuse and could guide designers toward opportunities to harden the platform against them.

Tooling and Attack Surface: Attackers often used non-standard locations and methods to deliver toxicity. Just as database or web security practitioners consider every user-controlled value as potentially adversarial, we argue that designers should view any user-controlled data or actions as potential vectors for toxic behavior or its enhancement. Moderation tools should handle toxic usernames, profile pictures, gilding, or follows as easily as they handle "standard" vectors like posts and comments. And designers should include this full attack surface in security mindset analyses to identify potential challenges to moderation.

Trust: Much as conflict of interest rules require agents to mitigate not only actual conflicts of interest but also the appearance of them, platforms should be wary of losing user and moderator trust in situations where the platform's incentives appear to align against users. The unusual attack we identified—in which users create accounts with toxic usernames and then pay Reddit to attach those toxic usernames to targets' posts—are a striking example of this idea. Though Reddit may not show deference to paying harassers, moderators warily noted the obvious financial incentives. Platforms should watch for the ways that their incentives appear to users and moderators as they design their features and profit models, since the lost of trust in platforms could be a major problem for their ability to combat hate and harassment. Our results illustrate this by showing the importance of the relationship between admins and mods in supporting community safety.

Policy: Platform policy can be examined through these lenses and using the security mindset. For example, Reddit's liberal policy around alt accounts, as well as their privacy-preserving choice to discard edit history for posts, can be misused. As with technical features, misuse doesn't mean that a policy is wrong—these policies have broadly understood and proven benefits for legitimate users. However, as moderators suggest in their discussions of the nuances of such policies, exploring the design space of such policies holds great opportunities for mitigating misuse while retaining or even enhancing opportunities for legitimate benefits.

6 Conclusion

By exploring the challenges moderators face combating online hate and harassment, this paper revealed intricate power dynamics between moderators and platform administrators that complicate moderation; systematized ways that adversarial users can carry out attacks by misusing platform features without privileged access; and proposed a set of design themes for designing interfaces, incentive models, and platform policies using the security mindset to limit the potential for feature misuse while retaining functionality.

Acknowledgments

We thank u/Watchful1, a moderator of r/pushshift, for their assistance with data collection. We also thank Kelly Wang, Jay Rodolitz, Erika Melder, and Michael Ann DeVito for providing feedback on this manuscript. This research is supported by the National Science Foundation (grant 2334061 and 2317114).

References

[1] Everything in moderation: Case study: Reddit. bit.ly/4b1sHEc. [Accessed 10-13-2023].

[2] Moderation queue. https://support.reddithelp.com/hc/en-us/articles/15484440494356-Moderation-Queue. [Accessed 10-13-2023].

[3] Online harassment, digital abuse, and cyberstalking in america. https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/. [Accessed 10-13-2023].

[4] Reddit content policy. https://www.redditinc.com/policies/content-policy#:~:text=Respect%20the%20privacy%20of%20others,of%20someone%20without%20their%20consent. [Accessed 07-10-2023].

[5] Reddit help: Communities for moderators.

[6] Transparency report: January to june 2023. https://www.redditinc.com/policies/2023-h1-transparency-report. [Accessed 10-13-2023].

[7] What is modmail? https://support.reddithelp.com/hc/en-us/articles/15484162851860-What-is-modmail. [Accessed 10-13-2023].

[8] Online hate and harassment: The american experience 2023. https://www.adl.org/sites/default/files/pdfs/2023-06/Online-Hate-and-Harassmen-2023_0.pdf, 2023. [Accessed 10-13-2023].

[9] Zainab Agha, Karla Badillo-Urquiola, and Pamela J Wisniewski. " strike at the root": Co-designing real-time social media interventions for adolescent online risk prevention. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32, 2023.

[10] alienth. Time to talk. https://www.reddit.com/r/announcements/comments/2fpdax/time_to_talk/. [Accessed 10-13-2023].

[11] Hind Almerekhi, Supervised by Bernard J. Jansen, and co-supervised by Haewoon Kwak. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 294–298, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[13] Sara Atske. The State of Online Harassment — pewresearch.org. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. [Accessed 06-10-2023].

[14] Iris Birman. Moderation in different communities on reddit – a qualitative analysis study. 2018.

[15] Jens Brunk, Jana Mattern, and Dennis M. Riehle. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 429–435, 2019.

[16] Jie Cai and Donghee Yvette Wohn. After violation but before sanction: Understanding volunteer moderators' profiling processes toward violators in live streaming communities. 5(CSCW2), oct 2021.

[17] Danielle Keats Citron. *The fight for privacy: Protecting dignity, identity and love in the digital age*. Random House, 2022.

[18] Amanda L. L. Cullen and Sanjay R. Kairam. Practicing moderation: Community moderation as reflective practice. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[19] Lars de Wildt and Stef Aupers. Participatory conspiracy culture: Believing, doubting and playing with conspiracy theories on reddit. *Convergence*, 2023.

[20] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[21] Bryan Dosono and Bryan Semaan. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[22] Bryan Dosono and Bryan Semaan. Decolonizing tactics as collective resilience: Identity work of aapi communities on reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.

[23] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. Non-polar opposites: Analyzing the relationship between echo chambers and hostile intergroup interactions on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):197–208, Jun. 2023.

[24] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "a stalker's paradise" how intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[25] Sarah A. Gilbert. "i run the world's largest historical outreach project and it's on a cesspool of a website." moderating a public scholarship site on reddit: A case study of r/askhistorians. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.

[26] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. 01 2018.

[27] Rosalie Gillett, Joanne E. Gray, and D. Bondy Valdovinos Kaye. 'just a little hack': Investigating cultures of content moderation circumvention by facebook users. *New Media & Society*, 2023.

[28] Claire Goforth. Private quote tweets are the latest way trolls are targeting people for harassment. https://www.dailydot.com/debug/twitter-private-quote-tweets-abuse/, 2022. [Accessed 10-16-2023].

[29] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[30] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5), jul 2019.

[31] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[32] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. Moderation challenges in voice-based online communities on discord. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[33] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30(1), mar 2023.

[34] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the looking glass: Study of transparency in reddit's moderation practices. *Proc. ACM Hum.-Comput. Interact.*, 4(GROUP), jan 2020.

[35] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[36] Tina Kuo, Alicia Hernani, and Jens Grossklags. The unsung heroes of facebook groups moderation: A case study of moderation practices and tools. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[37] Hanlin Li, Brent J. Hecht, and Stevie Chancellor. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *International Conference on Web and Social Media*, 2022.

[38] Renkai Ma and Yubo Kou. "defaulting to boilerplate answers, they didn't engage in a genuine conversation": Dimensions of transparency design in creator moderation. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[39] Supreet Mann and Michael C Carter. Emotional disclosures and reciprocal support: The effect of account type and anonymity on supportive communication over the largest parenting forum on reddit. *Human Behavior and Emerging Technologies*, 3(5):668–676, 2021.

[40] Jiayue Mao. The role of nudges in mitigating and preventing cyberbullying on social media. In *2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022)*, pages 1404–1408. Atlantis Press, 2022.

[41] Adrienne Massanari. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[42] J. Nathan Matias. The civic labor of volunteer moderators online. *Social Media + Society*, 5(2), 2019.

[43] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online communities: An examination of moderation work. 2016.

[44] Aiden R. McGillicuddy, Jean-Grégoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online

communities: An examination of moderation work. In *International Conference on Interaction Sciences*, 2020.

[45] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, GROUP '16, page 369–374, New York, NY, USA, 2016. Association for Computing Machinery.

[46] Benjamin Plackett. Unpaid and abused: Moderators speak out against reddit. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html. [Accessed 10-13-2023].

[47] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, 52, 07 2018.

[48] Bruce Schneier. The security mindset. https://www.schneier.com/blog/archives/2008/03/the_security_mi_1.html, 2008. [Accessed 10-16-2023].

[49] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media and Society*, 21(7):1417–1443, 2019.

[50] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 868–895. IEEE, 2023.

[51] Ruth Spence, Amy Harrison, Paula Bradbury, Paul Bleakley, Elena Martellozzo, and Jeffrey DeMarco. Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust and Safety*, 1(5), 2023.

[52] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[53] Nicolas Suzor, Sarah Myers West, Andrew Quodling, and Jillian C. York. What do we mean when we talk about transparency? towards meaningful transparency in commercial content moderation. *International Journal of Communication*, 13:18, 2019.

[54] Madiha Tabassum, Alana Mackey, and Ada Lerner. 'custodian of online communities': How moderator mutual support in communities help fight hate and harassment online. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, 2024.

[55] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267, 2021.

[56] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "it's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[57] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. "at the end of the day facebook does what it wants": How users experience contesting algorithmic content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.

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

[59] Donghee Yvette Wohn. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[60] Bingjie Yu, Joseph Seering, Katta Spiel, and Leon Watts. "taking care of a fruit tree": Nurturing as a layer of concern in online community moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA, 2020. Association for Computing Machinery.

A Appendix

A.1 Final Codebook

Code	Definitions	Example Quotation
Lack of clarity on Platform rules	Lack of clarity on where reddit stands on rules in a particular situation.	*"I'm unsure if this is the forum to discuss this, but are license plates a piece of personal information? I know that some subs are more serious about it than others. Some have no restrictions against it. Is there a Reddit platform policy on this?"*
Lack of transparency of platform decision/action	Lack of transparency of how the platform or admins makes internal decision.	*"Recently, I received a permanent ban for "harassment" without any specific details provided, such as a link to the alleged offense or information on the recipient. Fortunately, the ban was overturned upon appeal. However, it was challenging to mount an appeal without knowing what I did wrong."*
Lack of clarity on platform tools/features	Lack of clarity on how the platform's tool/feature works.	*"I recently tested using an alternate account and was surprised to find that when you ban someone, they just can't post or comment. They can still access the subreddit, report content, award gold, and other god-knows-what."*
Lack of guidelines on reporting offenses	lack of clear guidelines on how to effectively record and report offenses and harassment.	*"When an individual submits 50 reports within a ten-minute timeframe, how should we notify the admins? Should we file a single report? Or should we report each instance separately?"*
Lack of clarity on third party tool	Lack of clarity of how a third party tool works.	*" I want to understand whether Safestbot bans people with activity on certain subs or if it bans those people with activity on my sub and the selected sub? Thank you."*
Lack of Mod safety guidelines	Lack of guidance for moderators in place to ensure their safety.	*"I used the same account for moderation and using Reddit. I shared artwork with my name on it. It did not occur to me that somebody might get hostile toward me for banning them. Dumb! "*
Lack of notification on feature/policy change	Poor communication from Reddit about new rules/tool/policy or updates of rules/tools/policy	*Reddit has obviously changed what they consider 'harassment'. Can mods get a clear definition on this and why there was a shift?*
Lack of tool/feature	The current tool or feature does not support needs. There is a need for additional or upgraded tools.	*"The current report form for hate based on identity or vulnerability lacks a text field for providing an explanation, unlike many other report categories. Consequently, it relies on the administrators' familiarity with all forms of hateful rhetoric."*
Issues with inconsistent/inaccurate AEO/Reddit/admin action	Moderators sharing challenges regarding inconsistent/inaccurate action from Reddit/AEO/Admin	*"The AEO/Admin team seems to turn a blind eye when users engage in doxxing against moderators, and they persistently misuse Google Translate to misconstrue harmless jokes and posts from our users as genuine threats or harassment, disregarding the context."*
Issues with lack of Reddit/admin response/action	Moderators sharing challenges regarding lack of response and actions from Reddit/AEO/Admin.	*"A close friend was a moderator and has endured persistent harassment for more than a year. Despite making numerous attempts to seek assistance from the admins, they never provided any help. Eventually, he reached a point of frustration and decided to leave Reddit altogether."*
Lack of/complexity of external support	Challenges with receiving support from external resources like police.	*"To file a police report, activities must constitute a "direct threat," and actions like stalking and doxxing alone do not meet this criterion. The individual in question does not know my whereabouts and has not explicitly threatened to engage in any illegal activities.*
Feature misused for harassment	The platforms features are being misused for harassment.	*"After investigating, we believe 2-3 former members of our sub are chronically downvoting. Now the majority of the posts on our subreddit's main page have no upvotes or negative karma, and someone pointed out that that it portrays the subreddit in a negative light."*
Harassment towards moderator	Moderators are receiving harassment and hate on Reddit either internally or harassment that escalates to other platforms.	*"We were recently bombarded by a single user who sent us the same harassing Modmail 30 times within a minute."*
Community Harassment	Harassment taking place in a subreddit that is affecting the whole subreddit, perpetrated by a group or individual (e.g., Massive downvoting, crossposting for the purpose of harassment, etc.)	*"Someone hacked our Discord server and misused moderator privileges through another alternate account within our subreddit, resulting in our community being temporarily shut down for a few days."*
Complex tradeoff	Moderators are in a situation where they have to compromise and make one choice, resulting in losing something, usually forgoing a benefit or opportunity.	*"C1: What purpose does a reporting system serve if it protects trolls (by hiding the name of reporters who abuse the report feature)? C2: What would be the point of having a reporting system if moderators could retaliate against those who use it?"*
Increase of Hate and Harassment during a special event	Moderators face and have to manage more hate and harassment-related attacks during special events such as Pride Month, voting, etc.	*"I mod of r/<redacted>, whenever something significant occurs, such as shootings or elections or something local, the situation often spirals out of control."*
Burn out	Moderators discussed being burned out and having to overextend their time and energy to manage community safety.	*"We've had moderators who resigned from their positions because it was adversely affecting their mental well-being and personal lives."*
Lack of options to share resources	The community members are unable to share resources with each other that could help the community.	*"Unfortunately, the document originates from a private subreddit, so I can't create an archive of the page. I'll attempt to contact one of the moderators to inquire if they're willing to authorize sharing any details from the post."*
Community norms	Moderators face challenges in creating and enforcing community standards, onboarding new mods, and dealing with internal issues.	*"The theme of my subreddit attracts nazis and those who perpetuate racism and transphobia. I edited our sidebar to reflect this message, I issued numerous bans, and I began soliciting mod applications so that we would have more help dispelling this type of behavior. Yet, I doubt that addressing this issue through bans will actually solve anything."*
Admin response	Admin's response to a thread.	*"We acknowledge that Pride month may attract increased trolling and harassment directed at LGBTQ users, moderators, and communities. Consequently, we are closely monitoring r/ModSupport write-ins pertaining to these issues, recognizing the severity of such harassment."*
Automod response	Automoderator's response to a thread.	*"It looks like you're asking about brigading. Brigading is when a group of users, generally outsiders to the targeted subreddit, "invade" a specific subreddit and flood it with posts, comments or downvotes, in order to troll, manipulate, or interfere with the targeted community. Your subreddit could be flooded by spam posts."*
Abusive moderator	Conversations about a moderator's abuse of power or someone citing a moderator as abusive.	*"Moderator from another subreddit is excessively utilizing the report function to spam, false accusations and is encouraging others to raid my subreddit."*

Table 3: Codebook of moderation challenges with definitions and examples