'Custodian of Online Communities': How Moderator Mutual Support in Communities Help Fight Hate and Harassment Online

Madiha Tabassum
Northeastern University
Boston, MA, USA
m.tabassum@northeastern.edu

Alana Mackey
Wellesley College
Wellesley, MA, USA
am116@wellesley.edu

Ada Lerner
Northeastern University
Boston, MA, USA
a.lerner@northeastern.edu

Abstract

Volunteer moderators play a crucial role in safeguarding online communities, actively combating hate, harassment, and inappropriate content while enforcing community standards. Prior studies have examined moderation tools and practices, moderation challenges, and the emotional labor and burnout of volunteer moderators. However, researchers have yet to delve into the ways moderators support one another in combating hate and harassment within the communities they moderate through participation in meta-communities of moderators. To address this gap, we have conducted a qualitative content analysis of 115 hate and harassment-related threads from r/ModSupport and r/modhelp, two major subreddit forums for moderators for this type of mutual support. Our study reveals that moderators seek assistance on topics ranging from fighting attacks to understanding Reddit policies and rules to just venting their frustration. Other moderators respond to these requests by validating their frustration and challenges, showing emotional support, and providing information and tangible resources to help with their situation. Based on these findings, we share the implications of our work in facilitating platform and peer support for online volunteer moderators on Reddit and similar platforms.

1 Introduction

Social media platforms, such as Reddit and Facebook, have emerged as powerful hubs for individuals seeking and providing support across diverse communities. These platforms facilitate the formation of online spaces where users can connect with like-minded individuals or those facing similar challenges, finding solace and understanding as they share experiences, seek advice, and offer empathy. These digital communities span various topics, including mental health, chronic illnesses, parenting, etc. The immediacy and accessibility of these platforms enable individuals to find support at any time, transcending geographical boundaries and fostering a global network of shared experiences.

These online communities experience various forms of toxicity, ranging from hate speech to cyberbullying. Volunteer moderators are frontline guardians in these communities, playing a pivotal role in maintaining their safety and well-being by fighting hate and harassment. Community moderators dedicate countless hours to enforcing community guidelines, curbing the spread of harmful content, and sanctioning offenders. On Reddit, for example, moderators provide, on average, $3.4 million worth of unpaid labor each year [29].

Volunteer moderators face various challenges in managing their community, including being personally targeted by harassers on the internet [5, 38], emotional burnout [15, 47], and lack of support from the platform [16]. Yet, there is little research examining how volunteer moderators seek out support in navigating through these challenges. While a few research studies looked at how moderators within a team collaborate to manage their community, revealing that they share frustrations and seek advice and affirmation on their actions from their peers [11, 18, 46], in this work, we examine what we term *moderator support* communities: online communities where moderators come together to discuss issues they encounter while moderating their communities and to request and receive support and advice from one another. These communities allow moderators to address moderation issues, with engagement from a diverse moderator community with different backgrounds and expertise.

In this paper, we complement prior works by examining mutual support among Reddit moderators in these moderator support communities: r/ModSupport and r/modhelp. We specifically focused on mutual support around fighting hate, harassment, and abuse, as it directly affects the moderator's

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

ability to keep the members safe and maintain a secure and respectful online environment. Both in these forums, Reddit moderators share topics or questions related to moderation to seek insights and advice from both their peers and Reddit administrators. We systematically analyzed moderators' discussions in these subreddits to understand:

- **RQ1**: How do moderators use "moderator support" communities for support in managing community safety? On what kinds of online hate and harassment topics do moderators ask for help or advice?
- **RQ2**: What types of advice are given in these communities by other moderators?
- **RQ3**: What role does this type of moderator-to-moderator support play in moderators' ability to protect the safety of their communities?

To answer these questions, we conducted a qualitative content analysis of 2,740 comments in over 115 threads about hate, harassment, and abuse, drawn from r/ModSupport and r/modhelp. Our analysis contributes to the Human-Computer Interaction (HCI) and Usable Security and Privacy research community in several ways:

- We are the first to our knowledge to provide a detailed characterization of different types of support exchanged among moderators in 'moderator support' communities to fight hate and harassment in online communities.
- We offer implications for design to facilitate peer and platform support for online community moderators in managing their own and community's safety and protecting them from online harm..

2 Related Work

Volunteer moderation in online communities

Moderation in online communities can be defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [20]. Volunteer moderators are the main drivers for moderation in many social media platforms, like Reddit, Facebook, Twitch, etc. These moderators wear many hats. They act as the custodians of community rules, explaining them to newcomers and reminding established members by enforcing the norms and guidelines [19]. They're also detectives, identifying violations and rule breakers and taking action by removing contents and punishing the violator [10]. But the moderator's job isn't just about enforcing order; they're also cheerleaders, fostering a sense of belonging and encouraging participation [46, 51].

Researchers have explored volunteer moderation across different platforms, offering varied insights into their roles and experiences [5, 10, 16, 24, 46]. Several looked at the digital labor of volunteer moderators and found that they spent a significant amount of time ensuring their community's growth and safety [29, 33]. In an interview with volunteer moderators, Dosono et al. found that moderators spend, on average, 2-3 hours daily managing and moderating their respective communities [15]. In addition to manual labor, they also shed light on the emotional labor moderators endure dealing with hate, harassment, and negativity in their subreddits. Steiger et al. shared the same sentiment, emphasizing the psychological impact of moderation in establishing and preserving personal boundaries to avoid burnout and navigating complex interpersonal conflicts within the community [47]. Schöpke et al. pointed out that in addition to disgruntled community members, psychological distress also stems from struggles with other moderators in the team [44].

Prior research with moderators also demonstrates the challenges moderators face in balancing free speech and community safety [25, 34], providing transparency of moderation decisions [7, 22], effectively communicating and collaborating with the moderation team [10], and with the moderation tools lacking the nuance required in considering contextual factors and corner cases [19, 23, 27]. Some others shed light on the challenges associated with moderation strategies based on interaction mediums, such as in voice-only communities in Discord [24] and live-streaming communities like Twitch [50]. Despite the vast majority of work looking at the challenges moderators face, little is known about the support moderators employ to navigate through these challenges. We extend the current literature by particularly looking at mutual support among moderators in managing community safety.

Peer communication and support in moderation

Multiple studies have investigated how social media platforms are used for general support-seeking, such as in navigating unemployment [17], job loss [8], dealing with specific physical health conditions [4, 35, 42], mental health [14, 37], or the death of a family member [6]. However, these studies also indicated that these spaces could lead to negative experiences, i.e., aggressive content, stalking, exploitation of shared information, etc.. These studies emphasized the support that online moderators provide to minimize such activities within their communities [2, 26, 41].

Some studies explored how moderators within a team work together to maintain supportive and safe online spaces. Chi et al. examined the communication and collaboration methods employed by volunteer moderators on Twitch, emphasizing the importance of both informal and formal communication to facilitate teamwork among moderators and streamers [11]. Seering et al. investigated how moderator teams interact in community development, observing that team members often discuss specific incidents to solicit advice or opinions on the most appropriate course of action, as well as to inform other moderators about actions taken [45]. In an ethnographic study involving Facebook moderators, Gibson et al. discovered that some moderators view their team as a source of validation and

confidence in their decisions, which helps alleviate the anxiety associated with volunteer moderation [18]. Moderators also express their frustrations within their moderation teams as a means of seeking social support [15, 18]. Beyond internal collaboration, moderators from various subreddits on Reddit joined forces in a large-scale collective action by temporarily shutting down their subreddits, demanding improved support from platform administrators [32, 39].

While most previous works have focused on moderators' reliance on their team for guidance and assistance, this paper explores mutual support among moderators beyond their immediate moderation team. Our research offers a comprehensive overview of how moderators leverage the expertise and experience of the broader moderator community to address challenges related to combating hate and harassment within their community.

2.1 Moderation in Reddit

Moderation on Reddit operates through a combination of automated tools, subreddit moderators, and platform administrators. Each subreddit is overseen by volunteer moderators who enforce community guidelines by removing inappropriate content, issuing warnings, or banning users who violate rules. Reddit also employs a team of paid administrators who manage site-wide policies and legal matters and can issue site-wide bans when necessary. Communication between users and moderators is facilitated through Modmail, a shared inbox where users can report violations and moderators can address community concerns. Automated tools such as "automoderator" assist moderators in identifying, filtering and removing abusive content. The "modqueue" serves as a central hub within each subreddit, listing all content pieces that needs moderator review, including user reports, filtered posts, and comments.

3 Methods

In this section, we present details of the data collection, filtering and the data analysis procedure.

3.1 Data collection & sample generation

In our study, we focused on two subreddits, r/ModSupport and r/modhelp. These platforms are specifically created for moderators to engage in discussions covering a wide range of topics, such as moderation issues, tools, and instances of online abuse within the community, etc. Moderators utilize these forums to seek assistance and guidance from both administrators and fellow moderators. These two communities have substantial user bases: r/modhelp, established in 2009, is the largest moderator community on Reddit with 121k members, while r/ModSupport, established in 2015, has 72.8k members as of February 2023. Both subreddits exhibit high activity with over ten daily posts (original submission made by a user in a subreddit), providing a rich dataset for our research. We downloaded all available posts from these subreddits from inception to December 2022 from pushshift.io [40], a platform that is used to maintain an up-to-date public archive for Reddit. We omitted posts where the that were empty, deleted by the poster, or removed by moderators. The resulting corpus contains 41,256 posts.

We focused on posts where moderators discussed hate, harassment, and abuse-related attacks towards their community or themselves and/or were 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 [3] and Thomas et al. [49]: *"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."* Through an iterative process, we developed a set of 35 keywords and key phrases drawn from Thomas et al.'s taxonomy [49] to which we added the set of reasons that Reddit's report form offers users for describing content that breaks site rules to identify posts relevant to hate, harassment, and online abuse, provided in Appendix A.1. We searched the posts containing these keywords, which left us with 3,321 posts in our final dataset. More details about the selection of subreddits and the process of generating keywords can be found on [48].

Final Sample: Two researchers randomly sampled a post from the dataset. They manually reviewed and discussed the post. If the post was unrelated to online hate, harassment, and community safety, it was considered a false positive and replaced with a new, randomly sampled post. Otherwise, they downloaded and coded the entire thread (the entire discussion that unfolds from the post, including the post, all the subsequent comments, and replies) associated with the post. This process of random sampling and coding continued until we reached saturation, following the guidelines in prior research [43] In total, we coded and reached saturation with 115 relevant threads (2740 comments) sampled from our final dataset. The sample is also used in another paper of our authorship [48] to systematize adversarial attacks on Reddit that are happening by exploiting platform features and identifying challenges moderators encounter for such exploitation. In this paper, we scoped our analysis to understand mutual support among moderators in mod-support communities to fight hate and harassment in the community they moderate, which was not investigated and reported in [48].

3.2 Data analysis

The goal of the data analysis was to evaluate each thread in the sample for the type of requests in the original post and the support exhibited in the associated comments. We used an open coding process to identify the types of support sought by the moderators in the posts. To develop the codebook for the types of support exhibited, we drew from the offers and provisions of support, as defined by Cutrona and Suhr's Social Support Behavioral Code [13], and adjusted our codebook to include only support codes and subcodes present in our dataset. Additionally, we added new codes and subcodes that emerged from our analysis that were not present in Cutron and Suhr's Support Code. For instance, we have added a 'clarification' subcode under the information support code and the 'unsupport' code in our codebook. This process consisted of having three researchers go 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. Two researchers then coded sets of 20-25 threads at a time, meeting between sets to compare codes, resolve disagreements, and revise the codebook until no new code emerged. Both coders coded and discussed the same set of threads and agreed on the codes, so we do not report inter-coder agreement. Finally, the codes were grouped into categories in order to characterize the kind of support requested and received presented in section 4. The research team held regular meetings to review and discuss the analysis results and the categories generated from the analysis.

3.3 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 posts, comments, posters 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).

3.4 Limitations

In this research, we only focused on public-facing posts and comments on Reddit specific English-language subreddits, within the context of mutual support in managing community safety. As such, it is uncertain how applicable our findings are to other platforms such as Facebook, Twitch, Discord, etc., which may have different moderation structures or to other contexts that are not related to hate and harassment. Moreover, moderators on Reddit may also seek support in other ways, such as in private subreddits or channels, which are not covered in our work. The aim of this study is not to establish generalizability but rather to examine a specific phenomenon within a particular context. Due to the nature of our research, many variables remain unknown. We do not have access to any data regarding the demographics of the moderators we studied, leaving their gender, education level, occupation, age, and location undisclosed.

4 Results

This section presents the results of our analysis of the Reddit data from two moderator support communities, r/ModSupport and r/modhelp, in reference to our research questions. In the rest of the paper, we use the following terminology:

- 'Community' refers to communities/subreddits moderators moderates unless otherwise specified.
- 'Moderators' or 'posters' indicate moderators from various subreddits who posted to seek assistance in moderator support communities.
- 'Commenters' or 'Redditors' are individuals who responded to these posts.
- 'Admins' are Reddit administrators unless otherwise specified.

4.1 Purpose of posting

Moderators who sought support tended to engage with the moderator support community, reaching out to everyone for guidance. They discussed various issues, such as the hate or harassment they encountered, the difficulties in ensuring the safety of their community, or simply expressing their frustrations to be acknowledged. They openly complained about their problems, shared personal experiences, and recounted specific instances where they explicitly sought support and advice. We have found that the support requested by the moderators falls into three major categories: suggestion/advice (63 threads); clarification of platform, features, and rules (31 threads); and feature/tool requests directed at Reddit administration (13 threads). These categories are not mutually exclusive, and some posts fall into multiple categories. When asking for support for any of these categories, moderators often shared frustration with their situation and challenges

(30 threads). In fourteen threads, the poster did not explicitly ask any questions or request any specific tool/feature; instead, they simply expressed their frustration with admins, AEO (Reddit anti-evil operations team that identify and address violations of Reddit's policies on the platform), or Reddit (i.e., lack of response from admin, wrongful action by AEO, etc.) and the platform's lack of support. Table 2 in Appendix A.2 displays the types of support requested by moderators in our sample with examples. In the subsections below, we describe each type of support seeking in more depth.

4.1.1 Requesting suggestion/advice

In most posts within this category, moderators sought suggestions and guidance on combating instances of hate and harassment they were experiencing (50 threads) or anticipated in the community they moderate (5 threads). The most common issues moderators faced were attackers spreading hate and harassment via mass downvoting, false reporting, spamming, harassing posts/comments/PM, etc., and a lack of support from Reddit in fighting these attacks. Sometimes, these attacks are specifically targeted to moderators of the community. For instance, the moderator mentioned: *"One user is personally attacking one of our moderators. They have even created a username with her name and the C word. He targets any comment she makes on the internet and says he will continue to escalate until she deletes her Reddit account."*

Moderators' inquiries spanned from seeking advice on handling specific attacks to asking for feedback on the actions they had taken or intended to take to stop or prevent such attacks. For instance, one moderator mentioned: *"My subreddit's theme attracts nazis, racists, and transphobic. I have issued numerous bans and begun taking mod applications to have more help dispelling this type of behavior. Yet, I doubt that addressing this issue through bans will actually solve anything. Are there any mods who have effectively prevented rampant bigotry in their community? How is it done?"*

In a few instances (4 threads), moderators asked for guidance regarding how to help a community member at risk. These community members were in vulnerable situations, like being abused or threatening suicide, or receiving targeted attacks like doxing and revenge porn. Moderators used the resources they had at hand, like reporting to Reddit and sharing supporting resources with the user. However, in all cases, we observed moderators feeling responsible for community members beyond that and asked for advice from other moderators. For instance, one moderator said: *"A user in my sub seems to be in an abusive and threatening situation. I understand that they need help, and I would use the Endangerment or suicide/self-harm form to report it, except that would not fit this circumstance. Also, I am outside of the US so I can't report this to my country's authorities. Do I need to contact (Reddit) support so that they may disclose the user's location to the police? How would I do that?"* Finally, in a few other threads (4 threads), moderators asked for advice on addressing wrongful actions taken by the AEO, best practices for adding new mods and managing their own safety.

4.1.2 Requesting clarification

Within the support-seeking post, 31 were posted where moderators posing queries or seeking clarification on various matters. Seventeen threads indicated moderator confusion regarding platform functionalities and features including how a specific feature works, appropriate ways to use a feature, the differences between features, when to use which feature, and how the platform makes decisions. Examples include inquiries like, can mod log be edited? what are the differences between different types of bans? which reporting category to use to report a particular instance? how AEO works?, etc. For instance, one moderator sought clarification from the admin, stating, *"Admins should disclose to us why they delete posts and comments. we should be notified with a simple message indicating violence, threat, dox, just something to guide us so we can better moderate according to their standards."*

In some other posts (12 threads), moderators sought clarification around Reddit policy and rules around specific matters(i.e., what constitutes brigading? What is the policy around doxing? etc.). For instance, one moderator asked: *"Recently, we have received multiple requests from a certain company to delete confidential data (email threads, sales reports, email addresses). I don't want to remove these because the posts are well-written and produce healthy conversation; they just attach a photo of information that this company does not want to be public. Am I legally required or under any obligation according to Reddit's terms to delete these posts?"*

Finally, in two threads, moderators asked for clarification on the modmails they received from Reddit to understand why they had received such warnings.

4.1.3 Tool/feature request

In several threads (13 threads), moderators specially requested tool or feature support from the admins. In most of the threads in this category (7 threads), moderators asked for features/tools to prevent attacks against the community. For instance, once moderators requested a tool to automatically flag and remove the accounts that evade bans and automatically remove posts/comments from ban evaders and block them from doing further harassment. Another moderator wanted the feature to only allow subscribers (who were subscribers on or before a certain date) to comment on a post to reduce the risk of brigading. In four of the threads, moderators requested features to help them detect and report offenders (i.e., the ability to see edited comments, the ability to add explanations while reporting, etc.). In a few threads (2 threads), moderators specifically asked for features to reduce moderators' harassment, such as the ability to seamlessly switch between

their dedicated moderation account and regular user profile, concealing moderators' identities when interacting with rule violators, and limiting the number of modmails someone can send to the mods in a certain period. One moderator stated:

"We were just inundated with harassing messages from a single user sending 30 messages in a minute, which I did not realize was possible. I cannot comprehend any circumstance in which this spam messaging would be wanted or acceptable, so why not take away someone's ability to do this?"

In summary, moderators use r/ModSupport and r/modhelp as a space for expressing and discussing various issues in the context of their experience fighting hate and harassment in their subreddits. These discussions encompassed topics such as attack prevention techniques, the safety of community members, moderation tools and platform policy, moderators-admin communication, platform moderation, and automated tools used by platforms. In our sample, most moderators posted using the account they were using to moderate. Some moderators who were facing targeted harassment used throw-away accounts to ask for advice to prevent stalking. Using the mod account is not surprising as these subreddits are specifically created for moderators and overseen by administrators, likely fostering a level of trust among posters despite occasional rude comments. Moreover, moderators often needed to discuss specific challenges encountered while moderating particular types of subreddits as exemplified by statements like *"I manage a subreddit focused on mental health, and there's a user actively encouraging our suicidal users to commit suicide."* Interestingly, some moderators believed they faced specific difficulties due to moderating niche communities: *"truly feels like Reddit admins ignore the NSFW community. Our issues are falling on deaf ears."*

4.2 Support received

We observed a wide range of types of support offered in response to moderators' requests and the challenges.

Most of the comments were positive, providing information, clarification, and validation to moderators. They often engaged in meaningful discussions, asked follow-up questions, and expressed gratitude. We observed only a few instances where commenters exhibited unsupportive or negative behavior. However, the fact that we rarely observed negative behavior might be explained by it having been effectively moderated, since it would be reasonable to expect these moderator-focused communities, one of which is moderated directly by Reddit administrators, to be promptly moderated.

We identified four categories of support among supportive comments: informational, validation, emotional, and instrumental. We describe negative behavior under a separate 'unsupport' category. Table 3 in Appendix A.3 depicts the types of support exhibited in moderator support communities in our sample with definitions and examples. We characterize these types of support in detail in the following subsections.

4.2.1 Information support

Cutrona and Suhr defined information support as providing information about the problem or how to deal with that [13]. In mod support communities, Redditors shared insights, ideas, and suggestions with fellow Redditors, aiming to assist them in better understanding their situations and making more informed decisions. Redditors provided information support by providing strategic advice, clarification and explanations, assessment of the situation, and referral to other people and resources. 108 threads out of the 115 threads received some form of information support from fellow Redditors.

Strategic advice: One of the main ways information support is demonstrated is through offering suggestions and advice [13]. In our dataset, Redditors provided strategic advice to handle ongoing attacks, how to prevent future attacks, and how to protect moderators and their communities from hate and harassment. We observed such a form of support in 61 threads. Sometimes, Redditors provide direct advice applicable to the posters' situation. Other times, they shared experiences of handling a similar situation and their personal moderation practices instead of giving direct suggestions to the posters. For instance, one Redditor shared their moderation practice in response to a moderator experiencing Brigading: *"Crossposting is the root of brigading and that is the main issue. We impose bans on xposters and lock and remove their posts. This process is written in our sidebar. We know it is controversial, but has functioned effectively in our community for years. The number of users from other subs who crosspost or abuse our sub has decreased significantly."*

These suggestions, however, are not always unanimous, triggering back-and-forth discussions. For example, a query about how to handle a troll induced the following discussion:

"C1: Document all of their actions in the next few months. Take screenshots of everything they do. Post all this evidence when they come back to disparage the mods, then ban them. C2: Too much effort. Take a firm stance. If they post another controversial comment against the mods, give them a temporary 30-day ban. Explain in the ban comment space that their trolling is not welcome in your sub and that the next time it happens, however minor, they will be permanently banned. Be banned or behave—those are their options."

Redditors frequently suggested configuring auto mods (i.e., restricting new accounts, screening new users, filtering posts and comments based on keywords, etc.) to defend against harassment and attacks. Other suggestions include strictening the community rules, actioning offenders (deleting posts/comments and ban), adjusting features (i.e., making the subreddit private, only allowing pre-made fliers, etc.), using existing third-party bots (i.e., safest bot, totesmessengerbot, etc.) or developing custom bots to detect and defend against abuse. Furthermore, there were a few instances where commenters recommended some best practices to reduce harassment and ensure posters and their community's safety, such

as using separate accounts for modding and general Reddit use, using a throwaway account when asking questions about something delicate about the community or community members, etc. For example, in response to a moderator sharing his struggle with a stalker, one Redditor suggested: *"Use a separate reddit account for posting outside of your subreddit. Only use your existing Reddit account for moderating in your subreddit(s). Other mods have spoken of doing this even if they haven't been harassed to avoid potential stalkers. It is impossible for stalkers to track you if you do not have posts outside of the sub(s) that you moderate."* Another Redditor suggested: *"Delete comments indicating suicidal intentions after sending the self-harm report to prevent any potential bad actors from contacting the user privately and exacerbating the situation."* to a moderator dealing with a suicidal user.

Situation assessment: According to Cutrona and Suhr, information support can also be offered by helping support requesters reassess or redefine their situation [13]. On 46 threads, we observed Redditors providing support by assessing someone's situation from their experience with Reddit and moderating communities. It includes analyzing why and how an attack or harassment may have happened, why admins or AEO may have taken a particular action, why someone's approach to handling an attack is working or not working, etc. For instance, one moderator posted about receiving *"anti-semitic language, as well as violent threats and homophobic and transphobic language"* attached with the reports in his subreddit, and one commenter assessed the attack by saying: *"most likely these are deliberate trolls attempting to get a reaction from someone. It's possible that all of the messages are from one user trying to appear a bigger threat than they are. Unless you have given them any personal info it is highly unlikely that they will follow through on any of their threats."*

In another instance, one moderator was concerned about the trolling attack with new accounts and shared their approach to handling trolling, and a commenter assessed why their approach failed to stop the trolls. He said: *"It's completely reasonable to ask an account to be a week old before posting. It gives genuine users a chance to become more familiar with your community before actually posting. You are being deceived in two ways: actual new users in your subreddit must both be new to Reddit or use a pseudonym, and they must have a topic so urgent that they cannot wait your probationary period to post it."* Similar to the strategic advice and clarification category, assessments were not always uniform and triggered back-and-forth discussions.

Referral: Cutrona and Suhr describe referral as directing an individual to other sources of help [13]. On 78 threads, we have observed at least one Redditor referring poster to other sources who can help with their specific queries or issues.

Referral to admin: For 71 out of the 115 threads, at least one commenter referred the person facing specific challenges to admin to solve their problem. Moderators were referred for report abuse, ban evasion, organized harassment/brigading, targeted and persistent harassment, and stalking. The high volume of referrals to Reddit admin is not very surprising, as we observed that the moderators often were unable to prevent and defend against those attacks with the tools and power they had at their disposal. The commenters also echoed the same challenge when they referred the poster to the admin.

Commenters offered such support by providing the link to the report form and message link to the admin, providing ideas on the option to select from the report form to reflect OP's situation and the information to include in the report. For instance, in response to how to modmail admins about username abuse, one Redditor suggested what information to include in the modmail to report abuse of award feature: *"Be sure to report the harassing username of the award giver. In no way, mention the username of the awardee and also give them a link to the harasser's Reddit profile. Don't report it in case someone commits an error."*

Moderators were advised to begin by reporting encountered issues using the platform's report form. However, if the individual who made the report did not receive any response, experienced inaccurate actions taken by Reddit, found that their efforts to address the abuse were ineffective, or if the situation was urgent, such as ensuring the safety of a community member facing issues like suicide threats, being in an abusive relationship, or encountering pedophilic activity, moderators were advised to directly message the admin via modmail.

We have observed a genuine effort from volunteer moderators to capture the admins' attention, even if that requires more time and energy from them, and that was reflected when they were providing advice. For instance, some Redditors suggested reporting every single example of report abuse and ban evasion even though it is time-consuming to report large scale attacks manually. Some also suggested not deleting the reports from the mod queue to protect the evidence, although it clogs up the mod queue and makes moderating problematic.

Referral to others: In the extreme cases of stalking and harassment like doxing and physical threats, the posters were referred to the legal authorities, e.g., the police and FBI. Some commenters also explained the law and how to approach the authorities: *"In Australia, it is a federal offense to use an online service to harass, threaten, or be offensive with a punishment of up to 3 years in prison. If you can locate his state, call that state's police (even though it's a federal crime, it is too minor to be enforced by the Australian federal police.) and initiate a report with all screenshots."*

Moderators were also referred to other subreddits and people who could help them with their issues. For instance, legal help subreddits for how to take steps against harassment, suicide watch-related subreddits for resources on how to help a user, automod-related subreddits to help set automod and code custom bots, mod reserves to get additional moderators in an emergency, etc. For example, in response to a moderator struggling with mass downvoting, one commenter suggested: *"I know of a subreddit called /r/x that used automod settings to*

prevent downvote brigading, maybe they could weigh in? You could reach out to their mods?" A few times, Redditors suggested seeking support from their community by exposing the abuse to the community. For instance, one Mod complained about another sub plagiarizing their post, and someone suggested: *" I'd recommend alerting your sub members to the situations and have them report as soon as it happens."*

Clarification: In addition to Cotrona and Suhr's information support behavioral codes advice, situation assessment, and referral, we have observed Redditors providing information that clarifies someone's confusion or misconception about the Reddit platform, features, and policy and was observed in 61 threads on our dataset. Redditors help moderators by explaining how the Reddit platform and different features work, what the rules are surrounding online attacks and harassment and how to report those to the Platform administrators. For instance, to clarify a moderator's confusion about *"Regarding doxing, where is the boundary between an influential individual's Twitter/social media and the socials of a local small business owner?"*, one Redditor explained: *"A published tweet isn't doxing. An example would be to say 'the individual that did x lives at 123 x avenue and their name is Jane Doe, here is a photo of them and phone number!"'*

However, these clarifications often come from the commenters' experiences and understanding of Reddit and may not match with reality. We have observed conflicting conversations among Redditors on how something works. For instance, two Redditors were in conflict about the difference between admin removal and moderators' removal of posts:

"C1: When it's deleted by the admin... it's deleted completely from the site. When it's deleted by a moderator, individuals can still view it if they use a particular link.
C2: To the best of knowledge, a deletion done by a moderator is the same as a deletion done by the admin.
C1: That is dependent on how they delete it. Admin can remove content off the entire website. Moderators can only obscure it from general users.
C2: I've looked into this and I'm afraid you're wrong."

4.2.2 Validation support

Another form of support we observed is validation support (on 50 threads), where Redditors validate posters' experience and/or needs. Cutrona and Suhr describe validation as a way of providing esteem support, i.e., communicating confidence by validating the recipient's perspective regarding a situation [13]. We included validation as a separate support code due to its prevalence in our dataset and established two ways validation is offered: confirming recipients' experience (i.e., confirming frustrating situations with AEO and Admin, confirming facing similar attacks and abuse, etc.) and endorsing suggestion/request (i.e., showing agreement that the requested tool or clarification is needed, endorsing someone's suggestion to prevent an attack, etc.). The validation support comment generally started with sentences like *'can second this'*, *'I have the same experience'*, *'I think that would be super helpful as well'*, etc. For instance, in response to someone sharing their concern about the increase in the frequency of abuse during Pride month, someone replied: *"This conversation has been very valuable. We've already experienced a spike in reporting for "misinformation" or "harassment" on any post with gay or trans content."*

In a few instances, we have observed Redditors validating platform updates. For instance, one moderator mentioned: *"I love the new blocking feature. The improvements they have made to blocking have greatly improved my quality of life. I find myself being stalked less frequently around Reddit."*

4.2.3 Emotional support

Curtona and Suhr defined emotional support as the provision of love, care and empathy [13]. Redditors provided emotional support by showing their fellow Redditors appreciation, care, understanding and encouragement on 27 threads. Redditors appreciated the moderators for their work and for managing particular subreddits that help users in need. For instance, one moderator shared his experience of harassment running a local subreddit during a mass shooting in that area. To appreciate the moderator, one Redditor responded: *"Y'all performed wonderfully in the aftershock of that catastrophe. Thank you for your work."* The moderator expressed gratitude by saying: *"Thank you for the acknowledgment and your gracious words. Your reply got a bit buried but even if it's just to let us know we are on the right track, I really value the response. "*

Redditors showed sympathy and care to fellow Redditors, especially when someone harassed or stalked online. For example, in response to a moderator sharing harassment experiences, one Redditor said: *"I am so sorry you were put through all of that pain. You're only trying to assist folks. You're a really great person, I don't believe I would continue moderating a sub with users that harassed me or others in that way."*

Redditors also provided emotional support by showing understanding of fellow Redditor's emotional condition. For instance, one moderator shares his/her experience helping a suicidal person while having a suicidal tragedy in the family. One Redditor empathized by sharing his own experience: *"I feel you. I nearly went through that with a loved one. They only began speaking to me again after their failed attempt. Before that, I had attempted to help her numerous times through panic attack. But we made no progress. Things only changed after she opened up again, it was horrific."*

4.2.4 Instrumental support

Instrumental support is comparable to providing 'tangible assistance' by offering goods and services [13]. Beyond just offering advice and knowledge (information support), we observed Redditors in our dataset providing tangible resources

or offering to provide specific services to help with the recipient's situation (38 threads).

The tangible resources include automoderator codes, custom bots, and materials that could potentially solve OP's problem. For instance, in response to a moderator asking for advice to clean up Bigotry in their subreddit, one Redditor responded: *"If it would help, there is an extensive filter in /r/<redacted> to stop *a ton* of slurs and bigoted terms. I am happy to share the Automod code to you."* However, we have observed instances where Redditors could not share resources that they think could be helpful to the poster. For instance, in response to a moderator sharing their struggle fighting with a well-known spam bot, one Redditor wanted to share a document that described the individual behind it and their strategy, various websites and domains they owned, and other details. However, they later realized the document was from a private Reddit community and said: *"I'm unable to even archive the page. I can attempt to contact one of their moderators and see if they'll allow me to share information regarding the post."* In a few cases, Redditors provided instrumental support by sharing their willingness to get directly involved in the issue posters face and trying to find a solution. These include offering to collect information for the poster and talk over DM to help build solutions. For instance, one Redditor wanted to set up an automated bot to help suicidal users with resources, and another Redditor showed support by saying: *"Let me know if you'd like me to do some more research for y'all :) Looks like you are already quite busy. "*

4.2.5 Unsupport

On 30 threads, at least one Redditor showed unsupportive behavior by questioning, demeaning, blaming, or bullying the poster or other commenters. Although several such comments were merely intended to insult or harass or were irrelevant to the thread's discussion, we noticed that certain unsupportive remarks served as a form of community self-moderation. In several instances, Redditors helped uphold community standards by criticizing posts or comments that violate guidelines, such as calling out someone for engaging in harassment within the thread. We also observed commenters holding posters accountable for their actions by citing their own experiences with the poster's subreddit or particular incidents moderated by the poster. For example, one moderator discussed conflicts with some users regarding a ban and expressed worry that it might escalate into a potential brigading attack on their subreddit. In response to that, one commenter responded with: *"Don't act like a victim, you messed up."*. Another commenter said in the same thread:*"Great. /r/x is a very toxic subreddit that bash men constantly. The sub's moderators are a complete disaster. I'm pretty certain u/x had an awful childhood experience that caused them to behave poorly towards men."* These initial comments sparked a cascade of criticism, where numerous commenters joined in to express their disapproval of the moderator seeking support and adding negative remarks about them. Such a response from the mod-support communities could lead to more stress, exclusion, and burnout.

In several instances, we have observed the moderators of r/ModSupport and r/modhelp removing rude/harassing comments and/or banning the unsupportive Redditors. For example, one commenter was being rude and demeaning to the poster. Though moderators did not remove the comment, the commenter was temporarily banned with the statement: *"Your harsh words are absolutely unnecessary here. One mod is distressed and uncertain how to handle the harassment that they are enduring. I'll give you a three day ban from r/ModSupport. I hope this time allows you to think through some things."*

Summary: moderators primarily found support from their fellow moderators. Reddit administrators also engaged with 49 out of 115 threads by offering guidance and clarification. Nevertheless, in the majority of cases, they referred poster to the Reddit report form or suggested contacting them via modmail regarding the abuse rather than directly providing information or assistance. In a few instances, admins elucidated the platform's rules, explained how Reddit operates or delved deeper into the reasons behind the poster's specific problem. For example, in response to a user's complaint about wrongful suspensions of community members, an admin replied: *" Thank you for this post, and sorry you're frustrated. We have multiple teams that employ mixed methods of human review and automated tools to prevent offensive content from reaching Reddit users. However, both will commit the occasional error, just like mods do. For that reason we have appeals. I investigated the 3 suspensions you spoke of and in 2 of them (including the suspension of your fellow mod), were revoked through appeal. In the third case the suspension timed out on its own. "* Overall, admins displayed empathy towards the moderators, and in 29 of the threads where admins provided a response, at least one moderator expressed gratitude towards them. Conversely, in 21 instances, at least one commenter engaged in arguments and exhibited frustration with admin responses.

5 Discussion

5.1 Stress, coping and community support

Prior research has associated volunteer moderation with stress, trauma, and burnout [15,44,50]. In our analysis, we found that moderators emotional distress stemmed from personal harassment, secondary trauma resulting from combating harassment, and a lack of prompt assistance from platform administrators when needed". Dosono et al. found that moderators cope with emotional stress by "Building solidarity from shared struggles,"sharing frustrations with team members, or connecting with community members facing similar challenges [15].

In general, online groups can help individuals cope with a wide range of stressors by providing access to a larger and

more diverse community of support, compensating for the lack of support available in immediate social groups [12]. In this work, we've observed moderators utilizing the moderator support communities to cope with stress using two strategies following Lazarus and Folkman's transactional model of stress and coping [28]. Some posters managed their stress by seeking information and solutions to their problems that cause stress (problem-focused coping), others by expressing negative emotions such as sharing frustration and anger with their peers to manage their emotions (emotion-focused coping), and some by combining both approaches. Receiving instrumental and informational support in moderator support communities may assist moderators with problem-focused coping. Receiving validation, empathy, and care may help moderators with emotional-based coping to feel less isolated, especially when stress comes from immediate team members. In this work, we have observed commenters providing emotional and validation support even when seeking such support was not the primary intent of the post, helping moderators cope with stress and navigate the emotional labor they experience to sustain their community.

On the other hand, we also observed moderators receiving negative and unsupportive responses when seeking help. Moreover, moderators often referred to administrators for solutions, and their lack of response was a source of stress: "We've reached out to the admins but received no response yet. This situation is really stressing me out, which is the last thing I need during finals.". Not receiving immediate or effective feedback or encountering negative interactions within mod support groups may negatively affect coping with stress. Future work should investigate both the positive and negative influence of peer (un)support on moderators' stress management and its impact on community safety.

5.2 Empowerment through mutual support

Ammari and Schonebeck introduced the concept of networked empowerment that highlights how social media facilitates the process of empowerment through access to people going through a similar situation [1]. In the context of our study, network empowerment describes how moderators use mod support groups to find and learn from other moderators going through similar experiences, share resources to support moderation challenges, and empower each other through mutual support of insights, validation, and solidarity.

Networked empowerment is built on Zimmerman's model of psychological empowerment, which includes three components: the interactional component, the interpersonal component, and the behavioral component [52]. Moderators' discussions examined in our results can be mapped onto these components, allowing us to understand the roles that various types of support play in supporting moderators' empowerment and thus their work to protect and strengthen their communities.

The interactional component describes people's awareness and ability to act toward goals. In the moderator's support group, this component involved moderators receiving informational support and instrumental support from other moderators. Through this, moderators gain insights into various strategies for addressing issues, learn from other's experiences, discover resources for managing community safety, grasp a deeper understanding of platform features and policies, and broaden their perspectives by considering alternative viewpoints offered by experienced moderators. It could be particularly helpful in situations with solitary moderation or inexperienced moderator teams.

The intrapersonal component describes "how people think about themselves," which includes someone's perception of their ability to solve the problem at hand and perceived competence in taking the actions necessary to do that. Moderators pose questions to seek guidance from their peers and discuss their approaches to problem-solving. In return, moderators received informational, instrumental, and validation support that may help moderators increase their competence in dealing with the problem at hand. Prior research indicates that moderators feel more confident in taking moderation actions when they receive advice and affirmation from fellow team members [18]. Additionally, emotional and validation support helps moderators to think positively and cultivate a positive mindset, considering that their feelings and experiences are acknowledged and accepted by others. Conversely, we note the high frequency of threads (71/115) in which moderators are encourage to refer problems to admin. This may suggest that mods often do not feel empowered to solve the problem at hand through their own actions, and this feeling and its propagation through referral-type support may weaken moderator empowerment by weakening the intrapersonal component.

The behavioral components describe someone taking action to directly influence outcomes. This is demonstrated by moderators joining moderator support groups, asking questions, moderators assisting each other by providing answers, guidance, and resources, and even volunteering to directly address others' issues through informational and instrumental support. For instance, when one moderator expressed a desire to combat bigotry, another moderator offered to provide Automod code they use in their subreddit to filter out content containing slurs and other bigoted terms. This directly influences posters' ability to eliminate content containing bigoted language from their subreddit.

In addition to supporting moderators' personal goals, we observed moderators using mod support communities as a space to advocate for changes that would benefit the whole moderator community, by highlighting platform-wide issues, requesting tools and admin support, and discussing what is needed to empower moderators. In response, the moderators received support from the mod support community, validating said cause and strengthening their voices for change. While Ammari's model of networked empowerment highlighted how networks can empower individuals by supporting

their personal goals, our findings introduce another dimension: empowerment by driving changes that could help the entire network.

5.3 Design implications

Here, we present implications for design to facilitate mutual support among moderators and reduce the challenges they encounter in managing their personal and community safety.

5.3.1 Peer support

Develop a formal repository for moderators to share common and contemporary advice, resources, tools, etc. In online communities, moderators often encounter similar challenges or have similar goals for their communities. For instance, moderators from multiple communities may deal with the same spam bot or have the same goal of identifying community members who crosspost posts to trolling subreddits, etc. Redditors actively create content and mechanisms to address ongoing issues and share those with others to assist with moderation, automation, or fighting contemporary hate and harassment attacks [30]. Currently, these resources are primarily disseminated through informal channels such as word of mouth or direct requests for support, potentially leaving valuable resources unnoticed by moderators who could benefit from them. In some cases, individuals may need to reach out directly to the creators of these resources, as they are not publicly accessible. We understand that not all content can be freely shared due to security concerns—for example, sharing automod code used to safeguard against specific attacks could inadvertently expose vulnerabilities to adversaries. However, there are still valuable resources, such as guidelines to contact external support, guidelines for how to deal with suicidal community members, insights into an ongoing sitewide attack, tutorials about new moderation tools, etc., that could benefit other moderators without posing significant risks to anyone.

Reddit features a 'Wiki' section within each subreddit, enabling moderators and approved contributors to generate and collaborate on content. However, its utilization varies among different subreddits. While some wikis focus on elaborating the subreddit's rules or guidelines, others serve as repositories for FAQs or resources pertaining to the subreddit's topic. Even if a subreddit's wiki holds valuable content, individuals who could benefit from it may be unaware of its existence. The wikis of r/ModSupport and r/modhelp subreddit offer resources related to common moderation issues. Nevertheless, we observed moderators sharing useful resources to deal with contemporary issues that are not documented in these wikis. Establishing official repositories for moderators to share resources could ensure that valuable knowledge is accessible to all the moderators. However, such a repository could potentially open up a new avenue for malicious actors to target communities. For example, an adversary might create a tutorial for a tool and embed harmful code snippets within it. If a moderator lacking technical expertise were to execute this code, they or their community could inadvertently fall victim to an attack. Furthermore, someone could create a guideline document containing harassing language. We've noticed that resources shared in discussion threads often receive endorsements from multiple moderators, either through comments or by upvoting recommendations made to the poster. This serves to validate the reliability of the resources. The proposed repository should also incorporate a mechanism for users to endorse or dispute specific resources, enabling moderators to conduct their own assessments on whether and how to utilize said resources.

Develop mechanisms to facilitate building formal and informal relationships. Prior research has underscored the significance of receiving support from one's social circle, including family and friends, particularly for individuals employed in emotionally demanding fields [9, 21]. This social support serves as a means for individuals to express themselves, fostering a sense of belonging and comprehension, ultimately enhancing their mental well-being. Earlier studies revealed that volunteer moderators cope with emotional stress by conferring with fellow moderators within their respective teams [15, 50]. Our findings expand the prior research, indicating that moderators often alleviate frustration by venting within larger moderator communities comprising individuals from different subreddits who may be facing similar challenges. O'Leary et al. suggested that mechanisms that connect peers based on shared characteristics, beliefs, and needs can notably enhance peer-support [36]. We recommend that platforms should deliberately incorporate features to facilitate informal communication among moderators. For instance, platforms could establish official chat channels exclusively for moderators, enabling them to discuss issues, exchange experiences, and forge relationships in a casual setting. Additionally, providing moderators with the option to customize labels, where they can share information about themselves, such as interests, experience, the type of communities they moderate, etc., may facilitate the formation of sub-channels and lead to communication. Furthermore, establishing official mentorship programs where experienced moderators can guide and support newer moderators can help build relationships. Such initiatives can be particularly beneficial for novice moderators without experienced moderators in their team.

It is important to acknowledge that such features could potentially be exploited by adversaries to inflict additional harassment on moderators. Therefore, it's imperative that these support systems are meticulously designed, taking into account potential avenues for exploitation and incorporating robust defense mechanisms. Developing such features entails navigating complex challenges, and future research should explore support systems that enable moderators to connect and bond with one another in an unexploitative and safe manner.

5.3.2 Tool support

Establish an effective admin-mod communication structure. We have observed that moderators are frequently directed to contact platform administrators for their challenges due to a lack of sufficient tools for moderators to address these issues independently. However, moderators have encountered confusion and received unclear advice regarding what information should be included in their communication with admins for varying issues. On Reddit, moderators have the option to reach out to administrators by submitting report forms or sending modmail to r/ModSupport, a specialized channel designated for moderators. Yet, we have observed moderators frequently getting frustrated with admins not responding to them in a timely manner and, worse, receiving no response at all, even for critical issues like abuse of minors, suicidal users, etc. Previous studies also highlighted the lack of support from platforms in addressing moderation issues [16, 31, 45]. In 2015 and 2023, Reddit moderators went as far as participating in a blackout to demand support from platform operators and for additional moderation tools [32, 39]. Platforms should consider establishing an effective communication structure between administrators and moderators. Additionally, platforms should offer guidance on the format of this communication, specifying what information moderators should include in their report forms or modmails for particular issues. Platform moderators can employ automation to identify the most common issues they receive from moderators and provide communication templates for those. Furthermore, there should be a mechanism for platforms to prioritize time-sensitive issues where moderators require immediate attention. For instance, transmission of pornographic material to underage Redditors may require more immediate action than an accidental removal of a post by AEO.

Develop measures to empower moderators against personally targeted attacks. Prior research indicated that exposure to harassment is one of the primary reasons behind volunteer moderators quitting moderation [44]. In our analysis, we observed moderators seeking advice or support to cope with targeted personal harassment encountered while moderating their communities. In response, commenters suggested various best practices, such as using separate accounts for moderation and regular Reddit activities, refraining from sharing personal information on the platform, etc. Many of the moderators offering advice had themselves been victims of targeted harassment, including stalking and doxing. Notably, we did not come across any explicit mention of official guidelines for moderators instructing them on how to better protect themselves against such targeted personal attacks. We conducted an informal review of Reddit's moderator resources and found no specific guidelines on moderators' safety besides standard account security recommendations such as utilizing strong passwords and enabling two-factor authentication (2FA). We advocate for the provision of training and resources for moderators not only on moderation techniques, tools, and protecting community members but also on self-protection measures against targeted attacks. Furthermore, as articulated by moderators in some of the threads we analyzed, platforms should consider introducing features that empower moderators to safeguard themselves, such as the ability to moderate anonymously using pseudonyms or anonymized profiles and seamless transitions between moderator and regular user profiles.

Adequately explain moderation features in place where moderators employ them. Our research reveals that moderators often seek clarification regarding the functionality of platform tools, features, or policies. However, they periodically receive conflicting answers, as responses from fellow Redditors may be based on assumptions rather than official information. Some clarification requires input from admins, like how AEO works. However, deciding to what extent such information should be shared is complex as it could be used by bad actors to evade detection. Then again we have also observed moderators seeking clarification about moderation features they use despite explanations being available in the official moderator help center. One reason could be that explanations are not readily accessible in the places where moderators apply these tools. Additionally, there may be insufficient detail provided about certain features. For example, while the Reddit moderator help center outlines activities a banned user cannot perform, it does not provide information about other features they can still use in the subreddit where they are banned. We suggest that platforms offer explanations directly in the places where moderators are more likely to use that information, covering all the details moderators may require about the tool through thorough user research. However, like any interface design, the challenge lies in providing sufficient yet concise information without overwhelming moderators.

6 Conclusion

Overall, our study provides a detailed characterization of the different types of support requested and received in the 'mod support' communities in the context of fighting hate and harassment. Our findings highlighted how support exchanged among moderators in these communities empowers them in managing community safety. The results unveiled implications around designing peer and tool support for moderators to better equip them to protect their own and community safety.

Acknowledgments

We thank u/Watchful1, a moderator of r/pushshift, for their assistance with data collection and Ashley Schuett for their

help with data analysis. We also thank the National Science Foundation (grant 2334061 and 2317114) for supporting this research.

References

[1] Tawfiq Ammari and Sarita Schoenebeck. Networked empowerment on facebook groups for parents of children with special needs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2805–2814, New York, NY, USA, 2015. Association for Computing Machinery.

[2] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Trans. Comput.-Hum. Interact.*, 25(5), oct 2018.

[3] Sara Atske. The State of Online Harassment — pewresearch.org. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. [Accessed 02-16-2024].

[4] Ashley A Berard and André P Smith. Post your journey: Instagram as a support community for people with fibromyalgia. *Qualitative Health Research*, 29(2):237–247, 2019.

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

[6] Jed R Brubaker and Gillian R Hayes. " we will never forget you [online]" an empirical investigation of post-mortem myspace comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 123–132, 2011.

[7] 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.

[8] Moira Burke and Robert Kraut. Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1419–1430, 2013.

[9] Carolyn M. Burns, Jeff Morley, Richard Bradshaw, and José Domene. The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. *Traumatology*, 14(2):20–31, 2008.

[10] 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.

[11] Jie Cai and Donghee Yvette Wohn. Coordination and collaboration: How do volunteer moderators work as a team in live streaming communities? In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[12] Tsai-Yuan Chung, Cheng-Ying Yang, and Ming-Chun Chen. Online social support perceived by facebook users and its effects on stress coping. 2014.

[13] Carolyn E Cutrona and Julie A Suhr. Controllability of stressful events and satisfaction with spouse support behaviors. *Communication research*, 19(2):154–174, 1992.

[14] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80, 2014.

[15] 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.

[16] 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.

[17] Radhika Garg, Yash Kapadia, and Subhasree Sengupta. Using the lenses of emotion and support to understand unemployment discourse on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–24, 2021.

[18] Anna D Gibson. What teams do: Exploring volunteer content moderation team labor on facebook. *Social Media+ Society*, 9(3):20563051231186109, 2023.

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

[20] James Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

[21] Trond Idås and Klas Backholm. Risk and resilience among journalists covering potentially traumatic events. *The assault on journalism*, page 235, 2017.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Sarah Kendal, Sue Kirk, Rebecca Elvey, Roger Catchpole, and Steven Pryjmachuk. How a moderated online discussion forum facilitates support for young people with eating disorders. *Health Expectations*, 20(1):98–111, 2017.

[27] 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.

[28] Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[29] 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.

[30] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. "could you define that in bot terms"? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3488–3500, New York, NY, USA, 2017. Association for Computing Machinery.

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

[32] J. Nathan Matias. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1138–1151, New York, NY, USA, 2016. Association for Computing Machinery.

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

[34] 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.

[35] Hyun Jung Oh, Carolyn Lauckner, Jan Boehmer, Ryan Fewins-Bliss, and Kang Li. Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in human behavior*, 29(5):2072–2080, 2013.

[36] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. Design opportunities for mental health peer support technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1470–1484, New York, NY, USA, 2017. Association for Computing Machinery.

[37] Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of the 24th international conference on world wide web*, pages 315–321, 2015.

[38] Benjamin Plackett. Unpaid and abused: Moderators speak out against reddit. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html. [Accessed 02-16-2024].

[39] Jon Porter. Major reddit communities will go dark to protest threat to third-party apps. https://www.theverge.com/2023/6/5/23749188/reddit-subreddit-private-protest-api-changes-apollo-c 2023. [Accessed 02-12-2024].

[40] Pushshift. Unpaid and abused: Moderators speak out against reddit. https://pushshift.io/signupl. [Accessed 02-16-2024].

[41] Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. Understanding moderation in online mental health communities. In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 87–107. Springer, 2020.

[42] Shruti Sannon, Elizabeth L. Murnane, Natalya N. Bazarova, and Geri Gay. "i was really, really nervous posting it": Communicating about invisible chronic illnesses across social media platforms. 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.

[43] 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.

[44] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. Why do volunteer content moderators quit? burnout, conflict, and harmful behaviors. *New Media & Society*, page 14614448221138529, 2022.

[45] Joseph Seering. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.

[46] 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.

[47] 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.

[48] Madiha Tabassum, Alana Mackey, and Ada Lerner. Investigating moderation challenges to combating hate and harassment: The case of mod-admin power dynamics and feature misuse on reddit. In *30th USENIX Security Symposium (USENIX Security 24)*. USENIX Association, August 2024.

[49] 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.

[50] 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.

[51] 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.

[52] Marc A Zimmerman. Psychological empowerment: Issues and illustrations. *American journal of community psychology*, 23:581–599, 1995.

A Appendix

A.1 Keyword list to filter post related to hate, harassment and online abuse

Base word	Word forms
Bully	Bully, Bullying, Bullied, Bullies
Troll	Troll, Trolls, Trolling, Trolled
Profane	Profanity, profane, profaned, profaning, profanities, profanely, profanatory, profanes
Offensive content	Offensive content, offensive contents, offensive post, offensive posts, offensive comment, offensive comments, offensive word, offensive words
Threat	Threat, Threats, Threatening, Threaten, Threatens
Violance	Violence, Violent, Violently
Incite	incite, inciting, incites, incited, incitement
Harassment	Harass, Harasses, Harassment, Harassed, Harassing
Dox	Dox, doxxed, doxing, doxes, doxx, doxxing, doxxes, doxed
Dogpile	Dogpile, Dogpiled, Dogpiles, Dogpiling
Raid	Raid, Raids, Raiding, raided
Brigade	Brigade, Brigaded, Brigades, Brigading
Mass downvote	Mass downvote, mass downvoting, mass downvoter, serial downvote, serial downvoting, serial downvoter, mass downvotes, serial downvotes
Abuse	Abuse, Abusive, Abusing, Abuser, Abused, Abuses
Impersonate	Impersonation, impersonate, impersonates, impersonated, impersonating
Stalk	Stalk, Stalks, Stalked, Stalker, Stalking
(Sexual| sexualization) & (Minor| minors)	(Sexual| sexualization| sexually| sexualize) & (Minor|minors)
Personal information	personal information, personal info, private information, private info, confidential information, confidential info
Self harm	self harm, self-harm
Suicide	suicide, suicidal
Racism	Racism, Racist
Bigot	bigotry, bigot, bigots, bigoted
Transphobe	transphobe, transphobes, transphobia, transphobic
Homophobe	homophobic, homophobia, homophobes, homophobe
Scam	scam, scammed, scamming, scams, scammer, scammers
AEO	AEO, anti evil operation, anti evil operations, anti-evil operation, anti-evil operations
Ban evasion	(ban|bans)&(evade|evades|evaded|evading)
Hate speech, Hateful, Hatred, Non-consensual intimate media, Revenge porn, Denial of Service, Explicit content, Vote manipulation	

Table 1: Keyword list to filter post related to hate, harassment and online abuse

A.2 Support requested in mod support communities

Types	Subtypes	Examples
Suggestion/advice (63 threads)	Combating hate and harassment attacks (55 threads)	*"Recently, we've had many users from other subreddits harassing our users over chat. As the only chat mod, I can't monitor the chat 24/7. How can I flag potential bad actors in real-time?"*
	Helping community members at risk (4 threads)	*"A user is expressing suicidal thoughts. I've informed Reddit, shared support resources, and offered to link him with a crisis counselor. I want to reach out to local authorities, but don't know who he is. Any advice would be helpful."*
	Managing personal safety as moderators (2 threads)	*"I have another account besides the one I regularly use. Should I use it as my mod account? Any suggestions? I do not want members to dig up my old content or stalk me."*
	Best practices to add moderators (1 threads)	*"We're searching for a new mod for our subreddit. What best practices do you follow when adding a mod?"*
	Addressing wrongful action taken by AEO (1 thread)	*"I noticed some comments containing "f***" were removed by AEO. But they were translations of video dialogue, not hate speech. Is it safe to approve these comments since AEO removed them?"*
Clarification (31 threads)	Platform functionalities and features (17 threads)	*Are there any differences between auto-mod shadowban and mod-ban other than the difference in the offender receiving notification?*
	Reddit policy and rules (12 threads)	*"Would it be considered as doxing if someone receives a DM saying, 'I've looked through your post history. It won't be difficult to locate you.'?"*
	Modmail warnings received from admins (2 threads)	*"The subreddit I moderate just received a wanrning about promoting hate without any details. Why are we getting warnings for rule violations that we're obviously not doing?"*
Tool/Feature support (13 threads)	To prevent attacks against community (7 threads)	*"Users in our subreddit are being targeted by followbots with offensive names. Please implement a system to prevent such abuse, perhaps a cooldown period for following users."*
	To detect & report offenders (4 threads)	*"I frequently encounter people posting harassing comments and then editing their text back to normal. Please show editing history for mods."*
	To reduce mod-targeted harassment (2 threads)	*" In my opinion, Reddit should hide all mod names when we interact with users who are receiving a ban."*

Table 2: Types of support requested in the moderator support communities to manage community safety

A.3 Support exhibited in mod support communities

Types	Definition	Examples
Information Support (108 threads)		
Strategic Advice (61 threads)	Provide recipient advice or ideas to stop ongoing attacks or prevent future attacks	*"Adjust your spam filter to the highest setting and secure your accounts as much as possible. If you're very concerned, you might ask the head mod to temporarily revoke mod permissions for everyone else, although this probably won't be necessary here."*
Clarification (61 threads)	Clarify recipient's confusion or misconception about the Reddit platform, features, and policy	*"Even if you ban someone, they can still view, vote, and report."*
Situation Assessment (46 threads)	Assess why recipient's is experiencing a particular situation	*"It seems this individual is based outside of the US, and some foreign ISPs are known to be quite lax regarding their customers' activities."*
Referral (78 threads)	Refer the recipient to some other sources of help	*"Talk to automod coders. They can help you deal with this. "*
Validation Support (50 threads)		
Confirm experience (41 threads)	Confirm recipient's frustration, concern or challenge is valid as they experience the same issues	*"I reported a harassing comment and got the same 'it doesn't meet the requirements' nonsense."*
Endorse suggestion/request (13 threads)	Show agreement with recipient's suggestion or request	*" Yes please!!! It would be incredibly helpful to have a way to explain why something is offensive or promotes hate."*
Emotional Support (27 threads)		
Appreciation (4 threads)	Show appreciation for recipient's work	*"Your subreddit is a lifesaver for nearly every moderator."*
Care/ Sympathy (11 threads)	Express sorrow for recipient's situation	*"I'm sorry that you had to experience this. Everyone deserves to be treated with respect, regardless of their identity or sexuality."*
Empathy/ Understanding (13 threads)	Express understanding of recipient's situation or disclose personal situation that communicates understanding	*"I understand how u feel. Moderating is a highly visible role, and suddenly many users know who you are, which can lead to some serious backlash."*
Encouragement (6 threads)	Provides recipient with hope and confidence	*"That's not okay. Don't give up. Last year, the subreddit I moderate was in even worse shape, but I persevered. You'll too."*
Jokes/memes (3 threads)	responding with joke or meme	*"Unofficial response from the admins (linked meme images)"*
Instrumental Support (38 threads)		
Tangible Resources (34 threads)	Share tangible resources, i.e., codes, tools, documents, etc. that can help to solve recipient's problem	*"Into the Automod, copy and paste this (Automod code). You will find 'Automod' option under mod tools"*
Willingness (8 threads)	Express willingness to perform tasks or provide services that would directly help with recipient's situation	*"Does my explanation make sense? I can look at it if you are worried about a particular element. "*

Table 3: Types of support received in the moderator support communities to manage community safety