Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19

LIA BOZARTH, University of Washington, USA
JANE IM, University of Michigan, USA
CHRISTOPHER QUARLES, University of Michigan, USA
CEREN BUDAK, University of Michigan, USA

Past work has explored various ways for online platforms to leverage crowd wisdom for misinformation detection and moderation. Yet, platforms often relegate governance to their communities, and limited research has been done from the perspective of these communities and their moderators. How is misinformation currently moderated in online communities that are heavily self-governed? What role does the crowd play in this process, and how can this process be improved? In this study, we answer these questions through semi-structured interviews with Reddit moderators. We focus on a case study of COVID-19 misinformation. First, our analysis identifies a general moderation workflow model encompassing various processes participants use for handling COVID-19 misinformation. Further, we show that the moderation workflow revolves around three elements: content facticity, user intent, and perceived harm. Next, our interviews reveal that Reddit moderators rely on two types of crowd wisdom for misinformation detection. Almost all participants are heavily reliant on reports from crowds of ordinary users to identify potential misinformation. A second crowd–participants' own moderation teams and expert moderators of other communities–provide support when participants encounter difficult, ambiguous cases. Finally, we use design probes to better understand how different types of crowd signals—from ordinary users and moderators—readily available on Reddit can assist moderators with identifying misinformation. We observe that nearly half of all participants preferred these cues over labels from expert fact-checkers because these cues can help them discern user intent. Additionally, a quarter of the participants distrust professional fact-checkers, raising important concerns about misinformation moderation.

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

Additional Key Words and Phrases: crowd wisdom, misinformation, online moderation, crowdsourced flagging, crowdsourced fact-checking

ACM Reference Format:
Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 155 (April 2023), 33 pages. https://doi.org/10.1145/3579631

1 INTRODUCTION

The wisdom of the crowd, or the collective knowledge of a group of people, was found to match or even exceed the knowledge of experts in a wide range of domains, including financial markets,

Authors' addresses: Lia Bozarth, liafan@uw.edu, University of Washington, Seattle, Washington, USA; Jane Im, imjane@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Christopher Quarles, cquarles@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Ceren Budak, cbudak@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

2573-0142/2023/4-ART155
https://doi.org/10.1145/3579631

education, and health services [2, 4, 25, 72]. With the rising apprehension about the widespread misinformation and its costly consequences [45], many recent studies have also explored ways to leverage crowd wisdom for misinformation detection [5, 9, 27, 57, 75, 78]. Broadly, researchers have evaluated whether crowdsourced workers can substitute professional fact-checkers and identify misinformation at scale [5, 9, 27, 57], and whether crowd signals available on social media platforms can be aggregated to detect misinformation [44, 47, 76]. These crowd signals include *crowdsourced flagging* which refers to users using platform-wide report systems to flag (i.e., report) content as misinformation to platform stakeholders (e.g., platform admins), and *crowdsourced fact-checking* which is defined as users fact-checking other users' posts via commenting (e.g., a user replies to another user's content with "this is fake"). Crowdsourced flagging signals are only available to platform stakeholders (e.g., platform admins). In contrast, crowdsourced fact-checking signals are public but not always easily accessible to platform admins unless they discover the comments. In practice, Facebook has relied on crowdsourced flagging to identify potential misinformation on its platform [58], and Twitter has recently introduced the BirdWatch program to facilitate crowdsourced fact-checking [59].

Thus far, most existing [58, 59] and theorized applications [5, 9, 27, 44, 57, 75] of crowd wisdom for misinformation detection are from the perspective of platforms. Yet, platforms like Facebook and Reddit often offload misinformation moderation to their communities [33]. Currently, however, there is a lack of knowledge on the role and value of crowd wisdom in misinformation moderation for self-governing online communities, particularly from the perspective of community moderators.

Indeed, online communities vary significantly in their sizes, purposes, rules, norms, and user bases [12, 22, 28]. This suggests that the characteristics of crowd wisdom for each community vary as the quality and availability of crowd wisdom are directly impacted by crowd composition [56, 57]. Likewise, the decision-makers (i.e., community managers and moderators) of these communities also differ in their personal beliefs, moderation philosophy, and expertise [40, 67]. These differences may affect how they perceive and moderate misinformation. Given the extent of variance in online communities, *can we find shared practices of misinformation moderation across these diverse online communities? Further, what is the role of crowd wisdom in the moderation workflow? More importantly, how can we help platform stakeholders (e.g., platform admins and community managers) improve the existing moderation process based on such understanding?*

To tackle these research questions, we focus on Reddit—a platform that is uniquely fitted for the study of crowd wisdom, given the heavy reliance of the platform on communities (i.e., crowds) themselves to self-moderate. Further, we also focus on COVID-19 misinformation, given that many Reddit moderators had publicly expressed that COVID-19 misinformation was rampant on Reddit, platform-wide actions against it weren't sufficient, and that they had significant difficulty moderating COVID-19 misinformation [3, 18].

Here, we conduct semi-structured interviews with 18 Reddit moderators to understand i) how moderators conceptualize misinformation (e.g., how they decide whether a post is misinformation), ii) their moderation workflow and practices, and iii) the role of crowd wisdom in the workflow. Further, we also seek moderators' thoughts on novel moderation queue designs that aggregate various public crowd signals to improve the moderation process. These crowd signals include crowdsourced fact-checking and moderation actions of other moderators, which are not easily accessible in the current Reddit moderation queue's interface (see detail in Section 3.4). Using design probes, we iv) explore these crowd signals' potential use and caveats. Finally, we v) compare the values of these crowd signals to labels from professional fact-checkers.

Our analysis reveals a general workflow model that encapsulates the moderation processes adopted by most participants to moderate COVID-19 misinformation. Further, we show that the moderation process revolves around three elements: content facticity, user intent, and perceived

harm. The earlier part of the moderation process is centered on identifying misinformation content (i.e., content facticity) and users who are intentionally spreading false narratives (i.e., user intent), and, moderation actions taken in the latter part of the process are affected by all three elements. Next, we show that Reddit moderators broadly rely on two types of crowd wisdom for misinformation detection: the wisdom of ordinary users and the wisdom of other moderators. Specifically, we observe that almost all Reddit moderators are heavily reliant on crowdsourced flagging by ordinary users to come upon potential COVID-19 misinformation (i.e., the "maybe pile"). When encountering difficult-to-identify, ambiguous cases, moderators often seek support from their own modteam or from expert moderators of other communities. Nevertheless, we also observe significant issues with crowd wisdom in the current moderation workflow (e.g., user abuse of the report system). Close to half of the participants preferred crowd signals over labels from professional fact-checkers when asked about their thoughts on new moderation queue designs. This is largely because crowd signals can assist participants with evaluating user intent. Additionally, a quarter of the participants distrust professional fact-checkers, raising important concerns about misinformation moderation. Finally, we synthesize the findings to conclude with concrete strategies such that platform stakeholders can better leverage crowd wisdom for misinformation moderation.

2 RELATED WORK

Online governance is a complex, multi-faceted process that involves many different actors and competing interests [30]. Past work has demonstrated that the success of self-governance is critically dependent on the contributions of both moderators and users in these communities. Here, we first review existing studies focused on the moderation practices of community moderators. We then discuss possible implications for misinformation detection based on the findings from these past studies. Next, we discuss the wisdom of laypeople and its uses in misinformation detection. We observe that most of the related studies were from the perspective of platforms and users, and there is minimal work exploring how moderators use crowd wisdom for misinformation detection. Nevertheless, we hypothesize the extent to which past findings may be applicable to our work.

2.1 The Moderation Practices of Community Moderators

Moderators' Roles and Moderation Practices. Community moderators have many roles and responsibilities. These include making rules to facilitate positive community norms, collaborating with other moderators, regulating problematic content and user behavior, and even undertaking collective action to promote platform-wide changes [48, 49, 67, 68]. Though, past studies suggested that community moderators spend most of their time regulating misbehavior and facilitating open/constructive discussions [68]. Additionally, the occurrence and strictness of moderation are dependent on the community's goals and the individual moderator's moderation philosophy and personal values [28, 67, 68]. Indeed, what's considered rule-breaking in one community may be viewed as normal user behavior in another [28]. Further, work by Seering et al. [67] showed that some moderators viewed themselves as the "referee" or "juror", emphasizing their role in overseeing and regulating communities. Others referred to themselves as a "neutral representative" who only stepped in when absolutely necessary, while a small group of "anti-censorship" moderators preferred minimal or no moderation.

These studies drive us to posit that subreddit goals and moderators' moderation philosophies likely affect misinformation moderation. For instance, "juror" moderators are likely to be more actively taking actions against misinformation posts and the corresponding posters, whereas "anti-censorship" moderators are likely to be avoidant of punitive moderation actions. Similarly, some communities may be specifically created to host high-quality content and conversations, and therefore, require moderators to be more strict against misinformation [81].

Facilitating the Moderation of Problematic Content and User Behavior. Past work [43, 49, 81] demonstrated that moderators employ various approaches and mechanics for moderation. For instance, most moderators introduce community-specific rules (e.g., no off-topic or low-quality content) and then promote user behavior aligned with community values through the enforcement of these rules [12, 28, 33, 81]. Moderators also rely on various technical systems—native to the platform or third-party supported—for routine tasks [34, 80]. As an example, both Reddit and Twitch have auto moderators to automatically flag or remove problematic content. Third, community moderators often rely on good-faith members of the community to set positive examples, promote community values, and assist with moderation [39, 43]. Though, controversial communities often attract bad-faith users who actively hinder moderation [19]. Finally, past work showed that moderators' reliance on these methods is dependent on the community-level characteristics [11, 33]. For instance, larger communities rely on more algorithmic methods, such as bots, for moderation [66], while smaller communities are moderated chiefly through manual efforts [32]. And, niche, highly-specialized subreddits such as r/science are reliant on expert community members to assist with moderation, such as identifying misinformation [39].

We explore the extent to which moderators of different communities rely on crowd wisdom for misinformation detection as opposed to algorithmic approaches or other means. Based on past work described above, we posit that larger subreddits with high volumes of new content are likely to rely more on algorithm approaches to detect misinformation than crowd wisdom [66]. Likewise, we speculate that small subreddits are largely dependent on manual efforts for moderation. Further, past work [30] showed that bad-faith users often purposefully generate erroneous crowd signals to mislead platform stakeholders. As such, we hypothesize that controversial communities (e.g., r/political_revolution) that attract norm-violating users are less able to rely on crowd wisdom for misinformation detection [19].

Next, community moderators face many moderation-related challenges, and past literature has explored various ways to assist community moderators with regulating problematic content and user behavior [11, 35]. This includes studies that demonstrated making rules more transparent and salient can reduce problematic user behavior [35, 43, 50]. Other researchers have also explored ways to improve moderation through expanding existing algorithmic tools [11, 34]. Notably, Chandrasekharan et al. [11] introduced an advanced auto moderator that leverages cross-community learning. That is, it aggregates moderation decisions made by moderators of many different communities to learn and predict the moderation decisions of a given community. Unlike these related studies, we aim to improve moderators' misinformation moderation process through better leveraging crowd wisdom. The potential advantages and caveats of using crowd wisdom for misinformation detection are discussed in detail in Section 2.2.

2.2 Crowd Wisdom in Online Communities

Existing work focused on leveraging crowd wisdom for misinformation detection can be broadly assigned into two categories. The first [5, 9, 27, 57, 78] focused on the wisdom of *synthetic crowds*. These studies employed laypeople from crowdsourcing platforms (e.g., Amazon Mechanical Turk) to create crowds for content labeling tasks. Thus far, past work demonstrated that labels provided by synthetic crowds, when aggregated, are highly correlated with those provided by professional fact-checkers [5, 27, 57, 62], particularly when a crowd is balanced, and/or has received constructive treatments [9, 57, 61]. The second category of studies examined the value of *organic crowds*: communities of ordinary users present on social media platforms. These studies demonstrated that organic crowds provide two types of signals to assist platforms with misinformation detection. The first is crowds using the platform-wide report button to flag (i.e., report) posts as misinformation, which we refer to as *crowdsourced flagging*. The second is users commenting on other users' posts

and calling out these posts as misinformation (e.g., a user replies to another user's submission and says, "this post is fake"). We refer to the second type as *crowdsourced fact-checking*. Crowdsourced flags serve to directly alert platform stakeholders and are only available to such stakeholders. In comparison, crowdsourced fact-checking posts are public, but platform stakeholders like community moderators may not be aware of them unless they stumble upon the related threads. Below, we describe these two types of crowd signals in detail and discuss how they might assist community moderators with misinformation moderation.

Crowdsourced Flagging. Crowdsourced flagging is one of the most common ways for platform stakeholders to solicit help from ordinary users to identify problematic content [30]. Past work argued that crowdsourced flagging provides a practical mechanism that scales with the magnitude of online content [16, 30]. Moreover, it also affords stakeholders such as moderators the legitimacy to remove flagged content [16]. In recent years, social media giants, including Twitter, Facebook, and Reddit, all updated their reporting systems to enable users to flag content as misinformation. Though, there is limited transparency in how these platforms use crowdsourced flags. Notably, Facebook stated that it uses flags as one of the signals to determine whether a piece of content needs to be reviewed by third-party fact-checkers [58].

Prior work also identified meaningful caveats with crowdsourced flagging [15, 30, 65]. This includes findings about flags providing "little room to express the degree of concern, contextualize the complaint, or take issue with the rules" [30]. Further, flags may be biased due to user ideology [15] and the users who regularly flag may not be representative of the user base [30]. Additionally, flagging can be gamed or abused by bad actors [65]. These caveats likely affect the accuracy of flags. As an example, a recent Twitter internal study suggested that only 10% of COVID-19 content flagged as misinformation was in violation of Twitter policies. Researchers suggested that flag abuse can be mitigated through various means, such as distinguishing high-quality flaggers from low-quality ones [64].

Thus far, existing studies have largely focused on the values and caveats of crowdsourced flagging from the perspective of platforms. Unlike platform admins, moderators of communities embedded in these platforms have fewer resources to address flag abuse due to more limited administrative privileges [30], highlighting an important challenge. Yet, work that explores how moderators use crowdsourced flagging in their moderation practices is sparse [20]. Notably, Diakopoulos and Naaman [20] interviewed moderators of online newsrooms, and their study suggested that communities with diverse user bases who do not share the same interests and values tend to experience flag abuse. For instance, a user may flag content containing viewpoints that they disagree with but does not necessarily violate community rules. This result is related to past work on trolling and group conflicts [19, 23], which showed that moderators of controversial or minority communities often experienced disruptions from non-community members. Given the contentiousness of politics [6], we hypothesize that crowdsourced flagging is less accurate and more frequently abused in partisan communities. In contrast, apolitical, mainstream communities are likely to have higher quality flags.

Crowdsourced Fact-checking. Ordinary users also fact-check misinformation posted by other users via commenting. Thus far, the prevalence of crowdsourced fact-checking posts is not well understood. For instance, results from [51] demonstrated that ordinary users generated the vast majority (96%) of all fact-checking tweets on Twitter. This is primarily because the number of ordinary users far exceeds the number of professional fact-checkers. Another study [37] showed that approximately 15% of the posts containing false claims made on Reddit had at least one crowdsourced fact-checking comment. Further, results from [1] suggested that 5% of all posts on Reddit had at least one crowdsourced fact-checking comment. In [77], the authors identified Twitter users who have

been actively engaging in fact-checking and explored ways to increase these users' fact-checking activity levels through recommendation systems.

Scholars have suggested that crowdsourced fact-checking posts can serve to correct viewer misconceptions and curtail the spread of misinformation [53, 69, 77], particularly when the refuting sources are reputable, or when the user who wrote the fact-checking post is perceived to be credible. Though, in practice, crowdsourced fact-checking is not always appreciated. For instance, [56] explored the reception of crowdsourced fact-checking across three different subreddits: r/politics, r/the_donald, r/hillaryclinton. They observed that it was more prevalent and better received by users in r/politics than r/the_donald and r/hillaryclinton. Based on the results, we posit that crowdsourced fact-checking is less available and potentially less effective for partisan online communities.

Here, we observe that related work primarily focused on the value of crowdsourced fact-checking from the perspective of ordinary users. Similar to crowdsourced flagging, there is limited work that explored whether moderators consider crowdsourced fact-checking when moderating misinformation. We argue that this is largely because, unlike crowdsourced flagging, crowdsourced fact-checking is not readily available to moderators through platform mechanisms. Indeed, platform report systems are well integrated into the existing moderation pipelines, whereas additional, non-trivial algorithmic means are needed to identify fact-checking posts [1, 37, 77]. In this paper, we discuss the feasibility of moderators leveraging crowdsourced fact-checking and also introduce related design probes (see Section 3.4.2). We use these design probes to better understand whether crowdsourced fact-checking can assist moderators with misinformation detection. Additionally, we explore whether, similar to ordinary users, moderators also value fact-checking posts containing reputable sources and are written by credible users.

3 METHOD

3.1 Research Context

This project uses semi-structured interviews to understand how Reddit moderators regulate misinformation. The Reddit platform is uniquely fitted for studies focused on crowd wisdom. Unlike other social media giants such as Twitter and Facebook, Reddit adopts a more decentralized approach to content regulation [33]. While it had, in rare circumstances, quarantined and banned problematic subreddits and users en mass in the past, it is generally hesitant and slow to take content moderation actions [33, 48, 67]. Instead, it heavily relies on communities called subreddits to self-govern. Though, it provides some tools (e.g., spam filter settings and crowd control functionalities) to facilitate community-specific moderation [49, 67].

Ethical Considerations: Our study has received an IRB exemption for posing minimal risks of criminal or civil liabilities to participants. Nonetheless, research with human subjects often risks exposing participants to unintended consequences. To mitigate harm, we pseudonymized the subreddits that our participants moderate. Furthermore, we also provide aggregated participants' demographic data instead of providing the exact information.

In the following sections, we first illustrate the process we adopted to generate the list of subreddits that we used for purposeful sampling and our recruitment procedure. We then summarize the subset of moderators (and the subreddits they moderate) that participated in our study. Next, we describe our data collection process, which includes both our interview protocol and design probe. Finally, we discuss our qualitative data analysis approach.

3.2 Subreddit Selection and Recruitment

3.2.1 Subreddit Selection Pool. Reddit has over 3 million subreddits [1]. Our goal was to focus on subreddits of reasonable size where COVID content has been actively moderated and where COVID misinformation could have been an issue. To do so, we narrowed down the sample pool to subreddits that: 1) had $\geq$ 1000 subscribers (approximately 98th-percentile), 2) frequently hosted COVID-related posts, and 3) contained COVID-related posts that subreddit moderators took moderation actions on.

We used two processes to identify subreddits that contained frequent discussions around COVID. First, we identified the subset of subreddits where COVID was the primary discussion topic (e.g., r/coronavirus, r/covid19positive) by extracting the subreddits that contain at least 1 of the COVID-related keywords in their subreddit description field [2]. This list of keywords was obtained through related work [14, 42, 71]. Then, the first author manually reviewed all matched subreddits and filtered out subreddits that were not specifically about COVID. We also filtered out subreddits that were private, restricted, quarantined, or banned (e.g., r/nonewnormal). Our second process aimed to find subreddits that were not specifically about COVID but where COVID-related content nevertheless frequently occurred (e.g., r/health, r/massachusetts). To do so, we first parsed all submissions posted in all subreddits between April 2020 and June 2021 using pushshift.io. We then identified subreddits where at least 1000 submissions and at least 5% of all posted submissions in the time period were related to COVID. A submission was considered COVID-related if it contained at least one of the COVID keywords. We also filtered out adult 18+ subreddits (e.g., r/bostonr4r). These two processes collectively resulted in 572 candidate subreddits.

Next, we removed all subreddits without significant COVID moderation. Any subreddit that had fewer than 10 COVID submissions removed by moderators was taken out of the sample. This resulted in a final selection pool of 424 subreddits.

3.2.2 Recruitment Process. We used stratified sampling across three distinct subreddit attributes of interest to select participants for recruitment. These attributes are described below:

Subreddit Type: We assigned subreddits into two groups: COVID-specific subreddits (e.g., r/coronavirus, r/covid19positive) or non-COVID specific subreddits (e.g., r/health, r/stimuluscheck, r/massachusetts). We observe that 67 or 15.8% of the 424 subreddits are COVID-specific.

Subreddit Size: The largest subreddit in our selection pool had 2.4 million Redditors and the smallest had 1.3K. Prior literature showed that large subreddits commonly use automated approaches to address content that contains obvious violations; moderators of small subreddits voiced less need for it [34]. As such, we expected the subreddit size to play a role in the degree to which moderators rely on automated, computer-aided or manual approaches to identify misinformation. To capture this nuance, we assigned subreddits into small ($\leq$33.3-percentile), medium (33.3 to 66.7-percentile), and large (>66.7-percentile) based on each subreddit's percentile ranking of the number of subscribers. Specifically, we classified subreddits that had fewer than 45K users as small, subreddits with 45K to 105K users as medium, and subreddits with more than 105K users as large.

Subreddit Ideological Leaning: Prior studies identified ideology as a significant factor in how individuals view COVID-related misinformation [10]. Additionally, past work showed that users' ideological biases could impact flag quality [15]. Here, we used the method proposed by [60, 73] to compute the ideological leaning of subreddits. Briefly, given a subreddit s, we defined liberal-leaning users in s as users that i) had posted more frequently in liberal subreddits (r/politics,

[1] See the full list of subreddits: https://frontpagemetrics.com/list-all-subreddits

[2] Full keyword list: {covid, covid-19, covid19, covid_19, corona, coronavirus, corona virus, wuhan virus, wuhan, wuhanvirus, china virus, chinavirus, 2019ncov, ncov19, ncov2019, sars-cov-2, ncov, kungflu, pandemic, lockdown, outbreak, quarantine}

r/liberal, r/progressive, r/democrats) than conservative subreddits (r/the_donald, r/conservative, r/republican), ii) had a higher average karma score [3] in liberal subreddits than conservative subreddits, and iii) their average karma score in liberal subreddits was greater than 1 [60]. We also defined conservative-leaning users using a comparable definition. We then calculated the ideological-leaning of s using the difference between the numbers of liberal and conservative-leaning users, divided by the total number of subscribers of s. This metric was then standardized to a scale of [-1, 1] across the 424 subreddits in our selection pool. We refer readers to the original papers [60, 73] for a detailed description of this approach. Here, we labeled a subreddit as conservative-leaning if its userbase was more conservative than 95% of all subreddits in our selection pool. Similarly, a subreddit was labeled as liberal-leaning if its userbase was more liberal than 95% of all subreddits in our pool. We observe that subreddits including r/republican, r/askthe_donald, r/conservatives, r/lockdownskepticism, r/ccp_virus are conservative-leaning, and subreddits such as r/political_revolution, r/coronavirusmn, r/joebiden are liberal-leaning. Additionally, subreddits such as r/vancouver, r/COVID19, r/miami are not ideologically aligned. Finally, we note that a subreddit's moderators' ideology is not strictly aligned with the majority of its userbase.

We stratified the 424 subreddits according to the three dimensions listed above. This resulted in 18 (type $\times$ size $\times$ ideology) distinct strata. Subreddits were then assigned into waves; each wave containing one subreddit from each stratum. Note that wave size varies because certain strata had fewer subreddits than others (e.g., we had zero large, conservative-leaning, COVID subreddit, and only one large, liberal-leaning, COVID subreddit).

Finally, we invited moderators to participate in the study using five waves. For each subreddit, the moderation team was contacted through *modmail*, Reddit's system for messaging moderators. Responses from the four participants in the first wave were used to fine-tune the interview protocol prior to subsequent waves. Altogether, 78 subreddits were contacted between March and May 2022, which led to a total of 18 interviews (response rate of 23%).

3.3 Participants Description

Tables 1 and 2 contain the descriptions of the mods and subreddits that participated in our interview. As shown, we interviewed a total of 18 moderators accounting for 20 subreddits (note that during the interview, a couple of the interviewees discussed two distinct subreddits that they are moderating, both of which experienced COVID misinformation). As shown in Table 1, half of the subreddits are COVID-specific subreddits. Further, there are eight large subreddits, eight medium subreddits, and four small subreddits. We also observe that half of the subreddits are liberal-leaning, nine have no particular leaning (not-aligned) or are apolitical, and only one is conservative-leaning. Moderators from conservative-leaning subreddits overwhelmingly declined to be interviewed. Additionally, the vast majority of COVID-specific subreddits are not ideologically aligned.

Next, Table 2 contains a summary of the participants we interviewed (See 3.4.1). To ensure participant privacy, we only provide aggregated demographic information. The average participant age is 39, and only one person identifies as female. Most of the moderators reside in the United States, and two participants are located outside of the U.S. Further, two-thirds of the participants moderate one or two subreddits, and only two out of the 18 moderate five or more subreddits. In addition, moderators on average have 2.6 years of moderation experience on Reddit. The least experienced moderator has two months of experience, and the most experienced mod has over 6 years. Further, an average participant spends one hour moderating per day. Focusing on moderators'

[3]Reddit utilizes a rating system where users can anonymously upvote (positive karma) or downvote (negative karma) content that they perceive to be high-quality or low-quality [29]. Karma points for a user in a community reflect the upvotes and downvotes the user receives there.

Table 1. Subreddit Summary.

subreddit code	subreddit type	subreddit size	subreddit user political leaning	description
SR1	covid	large	liberal	U.S. regional covid-specific subreddit
SR2	covid	large	not-aligned	A subreddit for users who have covid-19
SR3	covid	large	not-aligned	A subreddit dedicated to covid-19 monitoring and discussions with a focus on high-quality posts
SR4	covid	large	not-aligned	A subreddit with highly strict rules dedicated to scientific news and discussions of covid-19
SR5	covid	medium	not-aligned	U.S. regional covid-specific subreddit
SR6	covid	medium	not-aligned	U.S. regional covid-specific subreddit
SR7	covid	medium	not-aligned	A satirical covid-specific subreddit
SR8	covid	medium	not-aligned	U.K. regional covid-specific subreddit
SR9	covid	small	conservative	A subreddit focused on discussing lockdowns & other pandemic policies
SR10	covid	small	liberal	U.S. regional covid-specific subreddit
SR11	not covid	large	liberal	Liberal political subreddit
SR12	not covid	large	liberal	U.S. regional subreddit
SR13	not covid	large	liberal	One of the largest political analysis and discussion subreddits
SR14	not covid	large	not-aligned	U.S. regional subreddit
SR15	not covid	medium	liberal	Political question answering subreddit focused on user opinion
SR16	not covid	medium	liberal	Political analysis and discussion subreddit with a focus on recent events
SR17	not covid	medium	liberal	A subreddit dedicated to documenting and debunking political conspiracies
SR18	not covid	medium	liberal	A political subreddit focused on a particular candidate
SR19	not covid	small	liberal	A progressive news and talk show subreddit
SR20	not covid	small	not-aligned	U.S. regional subreddit

COVID expertise, we observe that moderators' self-assessed COVID-19 knowledge score on average is seven (response to the question, *"How would you rate your level of COVID-19 knowledge? On a scale of 1 to 10 where 1 is no knowledge, and 10 is expert level"*). Surprisingly, self-assessed scores are comparable for moderators of COVID subreddits and non-COVID subreddits. Moderators also state that they obtain COVID information primarily from academic sources, reputable news sites, and government websites. Several mods also obtain information from their social and professional connections. Finally, five moderators (or 27.8%) are working or have worked in the health and healthcare industry.

3.4 Data Collection

The first and third authors conducted semi-structured interviews from early March to mid-May 2022. All participants were interviewed via Zoom. We obtained the audio recording of each interview after participants gave affirmative consent. The average interview duration is 1 hour and 10 minutes, the minimum duration is 56 minutes, and the maximum is 1 hour and 50 minutes.

3.4.1 Interview Protocol. The initial protocol consisted of seven sections, which explored: (1) the participant's moderation experience and self-assessed COVID knowledge, (2) the extent to which COVID misinformation has been a problem for the subreddit and Reddit as a whole, (3) how potential misinformation come to the moderator's attention, (4) how moderators decide whether content is misinformation, (5) the moderation actions taken by the moderator, (6) whether moderation of COVID misinformation is different from that of other problematic content, and (7) the usefulness of additional crowd-based wisdom, as shown in Section 3.4.2. We also collected demographic information from each moderator.

Protocol Update: After the first wave of four interviews (Table 2), all authors reviewed the high-level observations to discuss whether and how the protocol should be revised. We found that the initial protocol didn't reveal the extent to which moderators relied on manual moderation as opposed to automated moderation. We concluded additional insights were needed to better understand the relative importance of moderators' manual efforts to detect misinformation as opposed to

Table 2. Moderator (Interviewee) Summary. Participants from the first wave of interviews are marked by *.

Mod code	subreddit code	In Health industry	average mod time (hours per day)	mod exp(subs)	mod exp(years)	covid knowledge (1-10)	covid information sources
P1*	SR1	No	1.0	5+	3.0	8.0	Mod experience; journal articles
P2	SR2	Yes	1.5	1	<0.5	8.0	Mayo clinic; John Hopkins
P3	SR3,SR4	Yes	<0.5	4	1.5	8.0	Professional training; cohort
P4	SR5	No	0.5-1.0	2	2.0	6.0	State government sources; local news media; Reuters and AP
P5	SR6	Past	<0.5	1	2.5	8.0	Medical journals; New England Journal of Medicine; the Lancet; Google search
P6	SR7	No	2.0	3	1.5	6.0	CDC; institutional research
P7	SR8	No	1.5	5+	2.0	7.0	Prepublishing bio archive
P8	SR9	No	<0.5	2	2.0	7.0	Medics; people in biomedicine; news and preprints
P9*	SR10	No	<0.5	1	2.0	8.0	Medical journals
P10	SR12	Yes	1.5	2	4.5	8.0	The CDC or the DHS, state government sources
P11*	SR11	No	1.0	1	4.0	6.0	Contact who work for the FDA
P12	SR14	No	1.0	2	1.0	8.0	CDC
P13*	SR15	No	0.5-1.0	1	<0.5	5.0	Word of mouth; articles on the internet and Reddit
P14	SR17	Past	2.0	2	3.5	8.5	News sources, PBS, NPR, BBC, and other big news outlets
P15	SR18	No	1.0	5+	6.0	7.5	Own experience; own doctor
P16	SR16,SR13	No	2.5	5+	4.5	4.5	News on Reddit; another moderator who's a covid-19 researcher
P17	SR19	No	1.0	1	2.0	6.0	Progressive media
P18	SR20	No	0.5-1.0	2	5.0	7.5	CDC and NIH

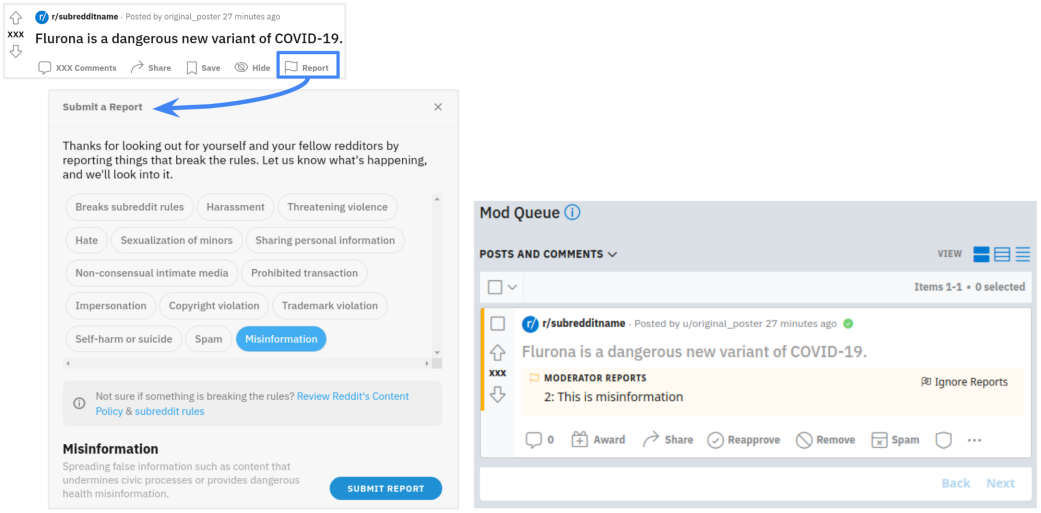
relying on other means, such as automated tools and crowd wisdom, to detect misinformation. Similarly, we observed that moderators relied on content characteristics to identify misinformation and user attributes to identify misinformation spreaders. However, the relative values of the two (user vs. content) were unclear. As such, we updated the protocol to ask about (8) the relative importance of automated moderation vs. human moderation, and (9) whether the moderators rely more on user account information than submission/comment content in making decisions concerning misinformation moderation. Further, we included additional design updates, which will be discussed in Section 3.4.2.

3.4.2 Design Probe. In this work, we are interested in whether and how moderators might use additional types of crowd wisdom to detect and moderate misinformation. We use a simple example of misinformation throughout our design probes: "Flurona is a dangerous new variant of COVID-19". We pick an easy and banal example because we want moderators to focus their attention on understanding our widget instead of getting distracted. Here, we first describe Reddit's current report system and modqueue (see Figure 1). We then introduce an alternative modqueue design that contains a widget with additional types of information.

Currently, when a user reports a post as misinformation on Reddit (see Figure 1a), the reported content and the number of reports show up on the modqueue as depicted in Figure 1b [4].

For our design probe, we introduce a mockup widget that contains three components and additional types of information (see Figure 2) that may be helpful to moderators in deciding whether the initial misinformation reports are accurate. The three components are *crowdsourced*

[4]Moderators have access to the number of users who reported the original post as misinformation through the modqueue. However, the reporters' usernames and account information are unavailable. The modqueue also doesn't include any explanation from the reporters as to why they marked the original post as misinformation. This is because Reddit's existing report UI doesn't include any input field (see Figure 1a). Moderators can click on the original post to go into the thread to find out more about the context of the reports. Moderators can also click on the reported user to examine their history (e.g., account age, total karma, and previous posts).



(a) Default User Report System

(b) Default Moderation Queue (Modqueue)

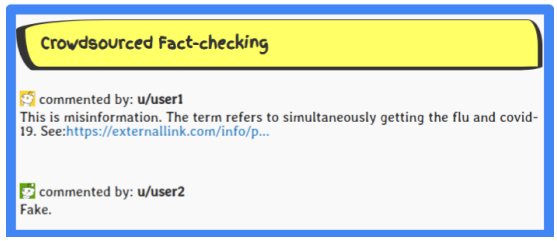
Fig. 1. The current Reddit native implementations of the user report system and the moderation queue. When users report a submission (or a comment) as misinformation (see Figure 1a on the left), the reports show up on the moderation queue (see Figure 1b on the right). Here, the modqueue indicates that two users reported the submission by *original_user* as misinformation.

fact-checking, *similar posts*, and *expert labels*. We chose these three categories based on potential usefulness and technical feasibility discussed in prior literature, which we describe below. We used static mockups made via animated Google Slides to showcase how the three components can be incorporated into the existing modqueue. Our interview process probes 1) to what extent moderators find the widget useful, 2) what additional information can make the widget more useful, and 3) the potential drawbacks of the widget.

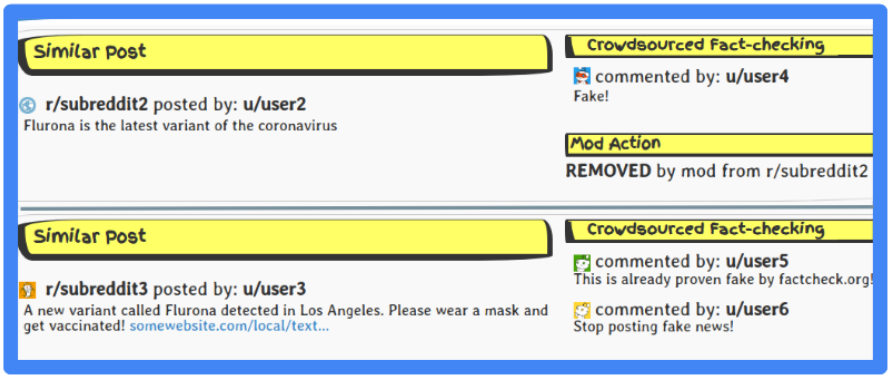
Crowdsourced Fact-checking: Similar to prior work [1, 37], we define *crowdsourced fact-checking* as the comments that reply to the original post and claim that the post is misinformation. For example, as shown in Figure 2b, the comments *"This is misinformation. The term refers to simultaneously getting the flu and covid. See [link]"* and *"Fake!"* are both crowdsourced fact-checking comments (Figure 2b). Unlike Reddit's native modqueue, with this widget element, moderators can learn which users wrote these crowdsourced fact-checking comments. Further, we posit that the additional text/evidence can inform moderators whether the original post is indeed misinformation [53, 69, 77].

Design Update: After the first wave of interviews, we updated our widget to include a popup that contains user account detail. This was done because, in our first wave of interviews, moderators disclosed that they often rely on user account age, total karma score, post history, and other account characteristics to determine whether a user is purposefully spreading misinformation (i.e., bad-faith) as opposed to being genuinely confused (i.e., good-faith). As shown in Figure 2e, moderators can

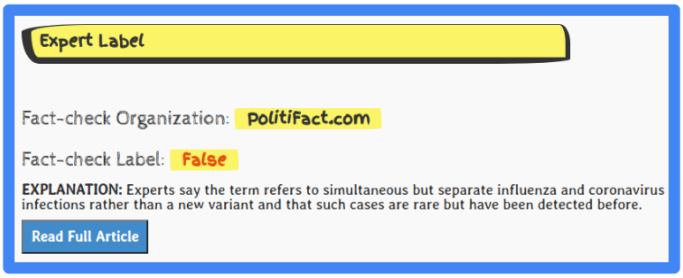
(a) Alternative Modqueue Design



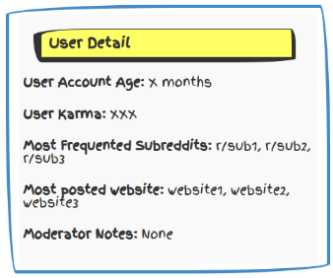
(b) Widget Component: Crowdsourced Fact-checking.



(c) Widget Component: Similar Posts



(d) Widget Component: Expert Labels



(e) User Popup

Fig. 2. Widget Mockup. The placement of the widget in the modqueue is given in Figure 2a. The widget has three different components given in Figures 2b-to-2d. Each component includes additional information that may help moderators decide whether a user-reported post contains misinformation and whether the poster is intentionally pushing a narrative. Note, moderators can hover over any of the usernames in the components *crowdsourced fact-checking* and *similar posts*, and the user popup element shown in Figure 2e will appear. The elements *user popup* and *mod action* were added in the second wave of interviews.

click on any username and learn its account age, total karma, most frequented subreddits, most frequently posted websites, and moderator notes [5] on the account.

Similar Posts: Misinformation can be shared across many different subreddits by the same user or by different users. As depicted in Figure 2c, the *similar posts* component contains posts comparable to the original content that is reported as misinformation. Additionally, it also contains crowdsourced fact-checking associated with these similar posts. If a similar post was removed by a moderator, the moderation action is also shown. Unlike the previous widget component, which relies on the users of a single community to provide fact-checking comments, the *similar posts* element gathers relevant crowdsourced fact-checking comments across the entire Reddit platform. In other words, moderators of one community can have access to the crowd wisdom of other communities. This component is motivated by past work [44, 47, 76], which suggests that aggregated crowdsourced fact-checking across entire social media platforms can be used to detect misinformation.

Design Update: For the second wave of interviews, we also included the user detail popup. We also added the *mod action* element based on participants' feedback. If similar content is removed by moderators of other subreddits, our widget will make this action known.

Expert Labels: Thus far both widget components *crowdsourced fact-checking* and *similar posts* contain information from Reddit. Here, as shown in Figure 2d, the *expert labels* component includes the label (e.g., "false", "missing context", "true") and explanation from third-party fact-checking organizations (e.g., Politifact, Snopes).

Technical Feasibility of the Widget: Several studies have proposed promising techniques to extract crowdsourced fact-checking using part-of-speech based rule parsing [1, 56], or pretrained deep-learning models [37]. Next, the information needed for the user popup is available via Reddit's native API. Several existing third-party tools, such as Context Bot [6] and Safer Bot [7] already include functionalities comparable to the user detail popup.

For the similar posts component, prior literature has proposed many deep-learning based paraphrasing models to identify similar content [8, 38, 70, 82]. For instance, many past studies relied on the pretrained paraphrase models (e.g., *paraphrase-MiniLM-L3-v2*) from the SentenceTransformer library [8]. Variations of these models have already been adopted by existing work focused on building automated fact-checking systems [8, 70]. Further, efforts like pushshift.io also make it easier to identify moderator removals [9].

Finally, the expert labels component requires a dataset of claims that were already verified by fact-checkers and a method to determine if a user-reported Reddit post contains a claim that matches one of the verified claims. Currently, there are two large datasets of verified COVID-19 claims extensively used in related work: *IFCN* [7, 51, 74] and *Google Fact Check Tools* [13, 26]. Once we obtain these verified claims, we can again use existing paraphrase models to determine if a user-reported Reddit post contains a claim that's already fact-checked.

[5] Moderator notes allows moderators of the same community to keep track of the participation histories of members of their communities. See https://mods.reddithelp.com/hc/en-us/articles/4635680764557-Mod-Notes-and-User-Mod-Log

[6] https://github.com/FoxxMD/context-mod

[7] Saferbot is a third-party tool that allows Reddit moderators to define a list of problematic subreddits, and then ban users in their own communities that participated in those problematic subreddits https://www.reddit.com/r/Saferbot/wiki

[8] These models can be fine-tuned using sentence pairs $(s1, s2)$ where $s1$ and $s2$ are paraphrases of each other (e.g., "this is a happy person" and "this person is very happy"). See https://www.sbert.net/index.html.

[9] Pushshift fetches and stores the submissions and comments soon after they are submitted. If a post is later removed by moderators, the removed post has the text "[removed]" instead of the original text. Though, it's worth noting that auto-removed posts are not captured by Pushshift.

3.5 Qualitative Analysis

The first author transcribed and corrected all interviews via rev.com. We used inductive coding [63] to analyze the transcripts. The first and second authors independently reviewed and conducted line-by-line open coding of 8 semi-structured interviews (four from the first wave and four from the remaining waves). Then, the two authors met multiple times to discuss and resolve disagreements to create one consistent codebook. This process consisted of merging codes into one, creating new subcodes, refining the original code's description, and removing codes that were found to be unnecessary by both authors. The other authors were also involved in several group meetings to review the codes and discuss initial themes. During the discussions, we decided to organize the codes according to moderation workflow activities (see work activity affinity diagram described in [31]). For example, top-level codes concerning the automoderator [34] were sorted to be next to each other. The workflow activities were informed by the interview transcriptions, prior studies on Reddit moderators' moderation practices [11, 34, 43], and the first author's own experience creating a new subreddit and moderating it.

With the merged codebook, the first and second authors each independently coded half of the 18 interview transcriptions (note that the eight initial transcriptions were recoded). During the independent coding process, the two authors regularly checked in with each other to discuss potential updates to the codebook, including adding new codes, updating descriptions of existing ones, or breaking an existing code into several ones. This allowed the codebook to stay consistent. The authors reviewed all codings to ensure they reflected the final codebook. Next, the first author assigned the codes into final themes using affinity diagramming, and the second author then reviewed each assignment. The two authors resolved all disagreements to reach a census.

In sum, we had a total of 137 codes. We identified various themes related to i) participants' moderation practices and workflow, ii) the role of crowd wisdom in the workflow, iii) the challenges of relying on crowd wisdom for misinformation moderation, and iv) moderators' perceived potential use and issues of each widget component. We describe these themes in the next section.

4 FINDINGS

4.1 Moderation Workflow and Practices

In this section, we first describe the general moderation workflow model, which encompasses the moderation practices of almost all subreddits. Further, we also show that moderation workflows commonly revolve around three elements: content facticity, user intent, and harm. Finally, we discuss variations in moderation practices as a result of moderator and subreddit characteristics.

4.1.1 Brief Overview of COVID-19 Misinformation Moderation Workflow. Our interviews reveal a general misinformation moderation workflow model that encapsulates all subreddits' moderation practices except one subreddit, SR15, which does not moderate misinformation. This model is summarized in Figure 3. It broadly consists of 6 subprocesses (blue diamonds). Each subprocess is employed by Reddit admins or subreddit moderators. Here, we provide a high-level overview of this model.

As shown, when a user makes a post in a given subreddit, it is immediately analyzed by Reddit's site-wide automated procedures (subprocess (a) in Figure 3). For instance, P14 disclosed that "Certain websites are banned from Reddit from linking ... I know that Great Awakening. That one's banned. You can't link." If this post doesn't invoke any moderation actions from the site-wide algorithms, it can still be identified as *potentially* containing misinformation or its poster can be labeled as a potential misinformation spreader by three distinct subreddit processes (subprocesses (b), (c), and (d) in Figure 3). These three processes are i) automated approaches at the subreddit level like the automod, ii) crowdsourced flagging, and iii) manual scrolling (i.e., moderators manually scrolling

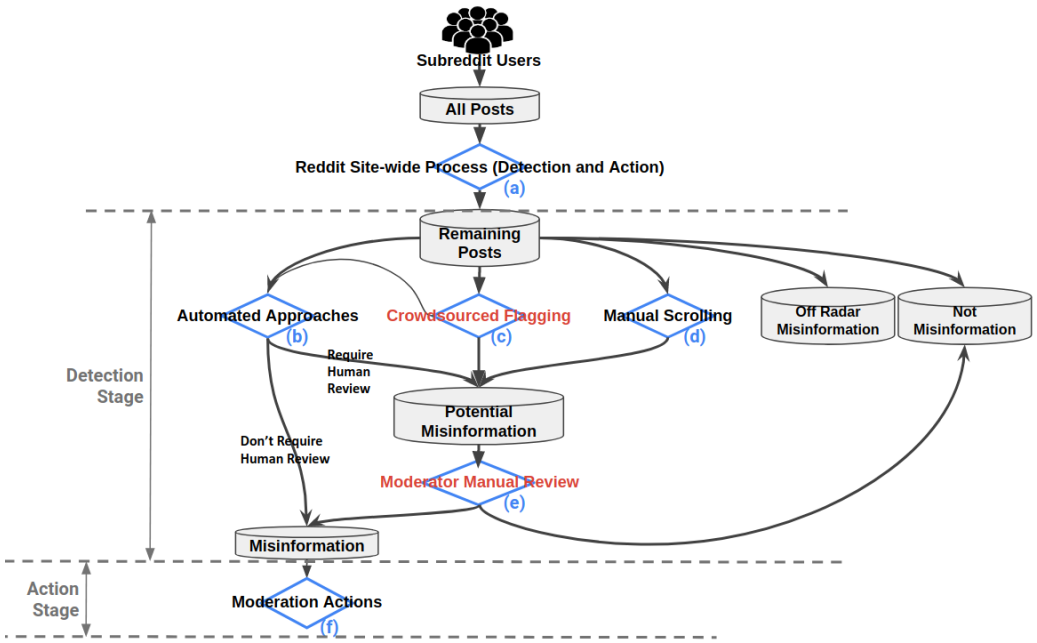


Fig. 3. Synthesized Moderation Workflow Model. Blue diamonds labeled *a*-through-*f* correspond to subprocesses used by the moderators in misinformation moderation. In the *detection stage*, moderators rely on the content and user characteristics to evaluate/classify content facticity and user intent. In the *action stage*, moderation actions' occurrence and severity depend on content facticity, user intent, and harm. The subprocesses with red labels incorporate crowd wisdom.

through their own subreddit and reading the submissions and comments). We discuss these three approaches in detail in Section 4.2.1.

Next, moderators predominantly rely on manual review (subprocess (e) in Figure 3) to determine whether or not the post indeed contains misinformation and whether or not the poster is indeed purposefully spreading misinformation (we described this more in detail in Section 4.1.2). Though, a handful of participants also noted that a few of the automated approaches they used (e.g., Saferbot) can bypass the manual review process.

Lastly, if the post is evaluated to be misinformation or if the poster is perceived to be a misinformation spreader, actions (subprocess (f) in Figure 3) are taken against the post and its poster.

Our analysis also suggests that this moderation model can be generalized to other types of misinformation and problematic content. In fact, a majority of the participants noted that their moderation of COVID-19 misinformation is comparable to other types of problematic content. The remaining participants noted that their moderation priority and difficulty differed across different types of problematic content.

4.1.2 Moderation Revolves Around Content-level Facticity, User Intent, and Harm. When describing their workflows, two-fifth of the interviewees mentioned that moderation of COVID-19 misinformation is an informal process or that their moderation practice likely varies from other subreddits due to Reddit not providing any official guidelines or policies on COVID-19 misinformation. Several also wished to have official policies on misinformation and additional technical support from Reddit.

> "Because obviously moderators across the site all have different rules in their subreddits and there are lots of different levels of what moderators feel is misinformation. So I feel like if Reddit stepped up and set specific policies about what people should remove, I think that would be very helpful." -[P12]

Nevertheless, our qualitative analysis reveals that misinformation moderation generally revolves around three elements: *content facticity, user intent*, and *harm*. Interestingly, the three elements that emerged from our analysis are largely aligned with the elements of the misinformation ecosystem described in [79]. Though, interviewees were focused more on *harm* than the *dissemination mechanisms* of misinformation. We observe that, in the *detection stage* (Figure 3), moderators rely on content-level and user-level attributes to evaluate content facticity and user intent. Moreover, in the *action stage* (Figure 3), the occurrence and severity of moderation actions are dependent on all three elements. Below, we describe each element separately. However, in moderation, these three elements are interconnected. For instance, a few participants observed that users who post obvious misinformation tend to operate with deceitful intentions.

Content Facticity: When asked "Are there specific types of Covid-19 misinformation that you're particularly vigilant against (very important to identify and moderate)?", one-third of the participants responded that they prioritize obvious, blatantly false claims that contradict facts with an established scientific consensus or are supported by a lot of existing data, such as misinformation about vaccines being dangerous. Some moderators described obvious misinformation as "you will know it when you see it"-[P3]. A couple of interviewees also said that *it* is posts that contain wild claims without providing any evidence to substantiate them.

In the *detection stage* (Figure 3), moderators relied on content-level characteristics to evaluate content facticity. A third of the subreddits relied on the automod and keyword regex matching to identify potential COVID-19 misinformation (subprocess (b) in Figure 3). For instance, P11 remarked that they "remove all posts that are from Fox News and also have COVID numbers because they've been a major source of misinformation early on." Some participants also considered content-level characteristics during manual scrolling (subprocess (d) in Figure 3). For example, a couple of participants asserted that they were more attentive to threads concerning topics (e.g., COVID-19) that were more likely to attract misinformation. Finally, half of the participants mentioned that, during manual review (subprocess (e) in Figure 3), if a post contained an URL, they relied on the reputation of the website as a signal to determine whether or not the post is misinformation.

In *action stage* (Figure 3), some moderators also took more severe actions against obvious misinformation. That is, the moderators would remove the post and also ban the poster, as opposed to just removing the post. Several participants noted that this type of misinformation is easy to spot and easy to ban. P18 asserted that their subreddit is more strict with obvious misinformation because the correct information is readily available and, as such, there is little value to having posts or discussions on this type of misinformation.

> "Vaccines were found to be safe, specifically Pfizer, Moderna, and Johnson & Johnson. They were found to overall be safe. To be completely candid, we just don't have time for that anymore, and there's enough of a knowledge base that is freely available." -[P18]

User Intent: Our analysis suggests that user intent is equally, if not more, important than content facticity. We observe that moderators depended on the following strategies to assess user intent in the *detection stage* (Figure 3). First, as a general moderation practice, almost all subreddits only allow users who meet specific account age and karma score requirements to post. Next, a handful of participants, most of whom moderate COVID-19 subreddits, remarked that they relied on several algorithmic tools (subprocess (b) in Figure 3) like Saferbot to scan users' post histories to determine

their frequented subreddits as a way to identify potential misinformation spreaders. During manual review (subprocess (e) in Figure 3), moderators also relied on a variety of user-level characteristics to determine whether or not a user was purposefully posting COVID-19 misinformation. These characteristics, ordered from the most frequently mentioned by the participants to the least, include: account age, total karma, whether the user had been making similar claims in the past in different subreddits, most/recently frequented subreddits, user activity gaps, most frequently posted websites, username, and user ideology. Finally, a third of the participants mentioned that they adopted several existing third-party tools (e.g,, Mod Toolkit, Reddit Enhancement Suite) or created custom tools to synthesize important user account and post history metrics to make manual review easier. Particularly, many moderators relied on moderator notes to keep track of users' problematic behaviors.

Next, most of the participants said that, in *action stage* (Figure 3), the severity of their moderation actions and their willingness to engage with the poster are dependent on the perceived user intent. Some participants emphasized the educational aspects for good-faith users. These moderators contended that removing a user's post and banning them would never enable this user to obtain the correct information. As a result, the user would remain misinformed.

> "They're genuinely misunderstanding the topic. They are repeating a fact that they heard, thinking that it's true. We haven't tried to do anything, but guide them to the truth through conversation. We don't ban them from the subreddit. We don't remove their messages. We interact with them and try to educate." -[P4]

> "Typically, we'll look at the user's history. And more often than not, it is somebody bouncing subreddit to subreddit and we'll just delete their comment, ban them and just move on; [if] they seem like a legitimate person, we will remove their comment and just let them know, 'Hey, we don't do that here.' " -[P10]

Still, many moderators believed that education efforts are futile for bad-faith users whose goal is to push false narratives. A handful of moderators also claimed that, in their experience, good-faith users who also posted misinformation were relatively rare in their communities.

Perceived Harm: The majority of participants also highlighted that they prioritize misinformation that they perceive would lead to harm at the individual level (e.g., causing a person physical harm) or at the societal level (e.g., causing harm to a collective). Examples of this type of misinformation include those that downplay the effects of COVID-19 (e.g., COVID-19 is just the flu or COVID-19 hasn't killed that many people), discourage others from getting vaccinated, or encourage people to pursue "alternative cures" such as hydroxychloroquine or ivermectin.

> "But it is dangerous that you think that [alternative cures] will cure Coronavirus. Because I really felt like this could really lead to someone getting hurt. And so, that's something I was really particular about." -[P17]

It's unclear whether or not harm evaluation is a part of the *detection stage* (Figure 3). Though, our analysis suggests that the topic of misinformation and volume are indicative of perceived harm. We also observe that perceived harm varies across the participants. Additionally, what moderators consider to be harmful changes over time. For instance, a few moderators viewed misinformation about masks not working against COVID-19 to be harmful early on, but less so now. Some of the moderators mentioned that this is because the volume of mask-related misinformation has significantly dropped and that the official guidelines on masks have also changed over time.

Finally, some moderators stated that, in *action stage* (Figure 3), their moderation actions are more severe against instances they perceived to be harmful.

4.1.3 Moderator and Community Characteristics, and Variations in Moderation. We also observed noteworthy differences in moderation practices as a result of moderator and community characteristics. Indeed, moderators vary in their moderation philosophies. Most notably, we observed that a few of the participants are strongly averse to punitive actions, including removals and bans (subprocess (f) in Figure 3). These participants emphasized the need for open discussions and limited censorship. For instance, one moderator referred to themselves as a caretaker rather than an authoritative figure. In contrast, another moderator stated that they had been very strict with COVID-19 misinformation because they believed that they were brought onto the moderation team to ban people. Further, the moderation process is also dependent on the moderators' expertise. In particular, we observed that most of the moderators who are or were employed in the health industry said that, during manual review (subprocess (e) in Figure 3), they often relied on their own knowledge to determine content facticity.

Moderation practices are also affected by community characteristics. For instance, P13 emphasized that the goal of their subreddit is to understand why Redditors believe what they believe, and it made little sense to moderate users' opinions even when they contain misinformation. A few of the non-COVID subreddits noted that they have specific rules against off-topic content and that these rules allowed them to take more aggressive actions with COVID-19 misinformation (subprocess (f) in Figure 3) because such content is unrelated to their communities.

Overall, observations here are aligned with many prior studies [28, 43, 67, 68] which found that moderator and community characteristics directly impact moderation practices.

4.2 The Wisdom of Two Crowds

In this section, we explore the role of two crowds in the generalized moderation workflow model: ordinary users and moderators. We show that moderators rely on the wisdom of the general crowd to detect potential misinformation (the maybe pile) and the wisdom of other moderators to decide ambiguous cases. Finally, we note that ordinary users and moderators are not two completely separate crowds. In our interviews, several moderators mentioned that expert users who demonstrated good character are often recruited to become moderators.

4.2.1 The Role of Ordinary Users. Our analysis shows that ordinary users assist moderators through *crowdsourced flagging* (subprocess (c) in Figure 3). Indeed, *crowdsourced flagging* is one of the most important, if not the most important, means used by participants to come upon potential misinformation and potential bad-faith misinformation posters.

> "So there were the users flagging. It is the number one way it [potential misinformation] comes to our attention. Number two is by scanning the background of new users coming into the community, [use Saferbot] to flag people that may be liable to post misinformation. The list of subreddits changes day to day; r/churchofcovid is one of them, r/nonewnormal." -[P1]

> "So things will come to our attention in two ways. Either someone will bring it to our attention by reporting it, because we've got specific report reasons that people can use. Or that we'll just be browsing the subreddit because we are still community members at the end of the day." -[P7]

Comparison to Other Methods: Some subreddits also rely on *automated approaches* and *manual scrolling* (subprocesses (b) and (d) in Figure 3) to identify potential misinformation. Approximately a third of subreddits relied on the automod to identify posts containing keywords or URLs that are associated with COVID-19 misinformation. Some used the Saferbot and its variations to detect potential misinformation posters. An even smaller number of participants created their own unique algorithmic tools. We observe that large subreddits and COVID-19 subreddits employed more

automated methods to identify potential COVID-19 misinformation. This is likely due to them having higher volumes of COVID-19 content. Moderators noted that some of these tools were very useful. Particularly, P3 noted that algorithmic approaches could preemptively capture potential misinformation before other users are exposed to it. Additionally, we also observe that subreddits with limited volumes of new content per day and acceptable moderation workload preferred manual scrolling. Results here are aligned with past work [32, 66], which found that subreddit size is a significant factor in the extent to which moderators relied on algorithmic means for moderation as opposed to manual efforts.

Among the three approaches, crowdsourced flagging is the *most widely used* method for identifying potential misinformation. Almost all subreddits that actively moderate COVID-19 misinformation rely heavily on crowdsourced flagging to find potential misinformation. Indeed, the vast majority of the participants remarked that moderation workload had been a significant challenge in their subreddits. As such, these participants cannot manage manual scrolling. While automated tools are highly scalable, some participants noted that the functionalities of many existing automated tools are limited and that algorithmic tools can be difficult to implement well. For instance, both P7 and P16 noted that algorithmic tools could often result in too many false positives. A few participants also mentioned trying out new tools but finding them too complicated to adopt. In comparison, crowdsourced flagging is well integrated into the moderation pipeline, and in the words of P5, "it's a very straightforward process". These circumstances likely contributed to participants' significant reliance on crowdsourced flagging.

4.2.2 Limitations of Crowdsourced Flagging on Reddit. Despite the heavy reliance on crowdsourced flagging, moderators also stressed that it has several significant weaknesses. That is, similar to past work [16, 30], we also observe that crowdsourced flagging is often misused/abused, the report rate can be low, and that a flag contains very limited information. However, we also show that flagging-related challenges faced by communities varied significantly. Our findings provide an additional layer of granularity to past work. We describe these findings in detail below. Additionally, we also discuss potential improvements derived from the interviews for addressing each limitation.

Prevalent Report Misuse and Abuse: The reporting system (Figure 1a) can be abused (i.e., used fraudulently) by bad-faith users, especially because the system is anonymous. All moderators have reported experiencing *report abuse*. Half of the participants also stated that report abuse is a significant problem, and it takes time and effort to address. Many moderators asserted that the *misinformation* option included in the report system UI is commonly misused as a "big disagreement button"-[P16]. Further, it tends to happen when users are arguing with each other, and one of them decides to report all recent posts made by another user as misinformation.

> "The report button often is used as a, 'Hey, I strongly disagree with this statement button.' So we'll go look in the modqueue and we'll see one user has been reported 20 times and everything's innocuous. Ten reports and it's all reporting the same user going back three days, four different discussions. That's usually what we'll see." -[P10]

Other times, bad-faith users abuse the report system to cause an additional workload for moderators.

> "We started getting a lot of trolling and spamming of moderator reports by people who didn't like the moderation. That's probably the single biggest problem that I've had in the whole process, when you get somebody who's a really dedicated troll who comes in, and it happens periodically." -[P9]

Some moderators also considered that sometimes this could be an honest mistake and the reporter sincerely thought that what they reported was misinformation. For instance, P6 mentioned that

people could accidentally report satirical content. Outside of the abuse of the *misinformation* option, some moderators also noted that the report options *harassment*, *self-harm and suicide*, and *sexualization of minors* are also frequently used as a way to harass targeted users or try to get them banned. In fact, one moderator stated that the *misinformation* option had been one of the less abused options in their experience.

Further, as indicated by past work [19, 23], moderators of controversial or partisan communities (e.g., P6) stated that they had experienced rampant report abuse. Interestingly, some moderators of apolitical, mainstream subreddits also mentioned that report abuse is significant in their communities. A possible explanation is that participants have varied perceptions of *significance* (e.g., a 20% abuse rate could be noteworthy to one moderator but not another).

Finally, some moderators noted that Reddit has been making an effort to address report abuse [10]. However, our analysis reveals that many participants felt existing efforts are insufficient.

Potential Improvements: identifying bad-faith users and limiting their ability to report. Most notably, a third of the moderators desired more transparency in the report system.

> "Let the moderator see who's doing the reporting. Because I strongly suspect that a lot of those types of reports, the frivolous reporting, is the same subset of people. It's the same small group of people doing it." -[P5]

Though, some of the moderators conceded that Reddit admins are probably concerned about retribution and will likely not implement this transparency. As an alternative, others have also proposed to make available various metrics that can be computed from user account and report history: i) the number of reports made recently, ii) whether the reporter is a brand new user, iii) whether the reporter has previously participated in the community. These metrics can be used to inform moderators whether or not a report is made in bad faith without having to reveal the identity of the reporter. Further, some moderators also proposed to preemptively limit who can report and how many times they can report. For instance, P16 proposed that when a user is banned from a subreddit, they should no longer be able to report on that subreddit. P13 argued that the report feature should not be available for users who have never interacted with a subreddit before or whose activity on the subreddit falls below a certain threshold. Interestingly, one moderator said they would like to have the ability to contact bad-faith reporters to tell them to stop misusing the report system.

Potential Improvements: identifying good faith users. Some moderators mentioned that they keep notes of regular and expert users in their communities. In addition, some of these users may receive a fact-checker flair [11] or other flairs that are indicative of expertise. Our analysis suggests that moderators give more weight to inputs and reports from these regular and expert users.

> "Some of our longer time users are very keen to report troubling posts because many of them feel very invested in the space too. I talked one on one with a number of them. They are quick to report things." -[P8]

Given that the current report system is anonymous, account characteristics that are indicative of a reporter being a regular or expert user are unavailable. While none of the moderators proposed any specific ideas, we hypothesize that including additional reporter metrics can help moderators determine whether a report is made in good-faith. Some examples include i) whether the reporter had received a positive flair from the community moderators and ii) the percentage of previous

[10] https://www.reddit.com/r/modnews/comments/mlgsw5/safety_updates_on_preventing_harassment_and_more/

[11] User flair is an icon or text that appears next to Reddit usernames. Each subreddit has its own user flairs set up by the community's moderators.

reports made by the reporter that was approved. These metrics can be useful to moderators without revealing the identity of the reporters.

Low Report Rate: Another issue faced by some moderators is low report rate. As noted by P11:

> "A lot of our actions will happen in response to that [crowdsourced flagging]. If a user doesn't report a post, then we don't notice it most of the time. Finding those posts before they get to the front page of the subreddit and get exposed to the wider audience is definitely the hardest part." -[P11]

A couple of the participants speculated that perhaps many users are unaware of the report button and what it does, or unaware that reporting is anonymous. Further, users may also choose to engage with the misinformation poster directly rather than reporting them to the moderators and Reddit admins. That would keep the moderators uninformed of the existence of the misinformation.

Potential Improvements: Moderators have used various strategies to improve crowdsourced flagging rate. For instance, P9 stated that their subreddit has clear rules on the sidebar requiring users to back up their claims using reputable sources and that users in their community are very good at reporting unsubstantiated posts. Some moderators also adopted (or considered) ways to promote community awareness of the fact that misinformation is taken seriously and encourage users to report more.

> "To get more quality reports, it's possible to have auto moderator sticky a comment on all posts. You could attach a [moderator] comment saying, 'This post is likely to attract a fair amount of COVID-19 misinformation. Please report this kind of thing.' " -[P16]

Results here suggest that transparency and salience of rules and moderation practices can increase report rate [35, 43, 50]. Further, a couple of moderators also mentioned that they thought giving expert users flairs (e.g., a fact-checker flair) may incentivize these users to report more frequently. It's worthwhile to note that past studies [21, 54] have also explored various ways to increase the flagging rate by promoting a sense of personal responsibility in bystanders.

Limited Information: Finally, as stated in the previous section, moderation practice is heavily reliant on content-level facticity, perceived harm, and user intent. However, moderators expressed lacking information to efficiently make use of these three aspects, as Reddit's current report system only makes available the number of times a post is reported as misinformation. A few moderators, however, mentioned that the number alone is very informative.

> "So if [redacted number] people make a report on a submission or a comment, it gets removed and put into our modqueue for us to look at." -[P4]

Many of these moderators said that if the number of reports has exceeded a certain threshold, the reports are generally accurate. However, some mods also expressed wanting additional information from the reports. For instance, P10 noted that it would be nice to be able to contact the reporter for additional evidence. The result here resonates with prior work that asserts crowdsourced flagging is a thin expression [16, 30].

Potential Improvements: A simple and straightforward way to allow reports to have additional context is to include an input box where reporters can write some text to tag along with their reports. One moderator, P3, noted that their subreddit has an updated report UI with a *custom response* option. P3 said that users had used this option in the past to provide additional information to the moderators. However, not every subreddit has a *custom response* option. More importantly, past work that's focused on using flagging for misinformation detection has been more invested in classifying content facticity [15, 44]. It is unclear whether additional text provided by reporters can help moderators evaluate harm and the intent of the reported user.

4.2.3 The Role of Other Moderators. During manual review (subprocess (e) in Figure 3), moderators can read the content, view user post history, and use various means to discern content facticity and user intent. Our qualitative analysis demonstrates that moderators often rely on the wisdom of their modteam for ambiguous cases. Further, some participants also stated that they also rely on moderators from affiliated subreddits or reputable subreddits to moderate misinformation.

Deciding Ambiguous Cases: Some participants noted that in many cases, it's not difficult to tell whether the content is misinformation, and whether the user is purposefully spreading misinformation. When something is murkier, however, many of the participants would check in with fellow mods for inputs using Discord or Slack.

> "Generally, we'll talk it over as mods and we try to err on the side of leaving it up. We might even post a moderator comment saying, 'We had a discussion about this and we're really not sure if this is accurate or not, but it's plausible.' " -[P5]

The vast majority of moderators also said that they keep moderator notes on users that could potentially be problematic, such as users who participated in COVID-denial, antivaxxer subreddits. These notes are shared among moderators of the same community.

> "If it's something that's super on the fence where I'm legitimately like, 'This person could be doing it in good faith. I don't want to stifle that.' I would usually post it. The mod team has a Slack where we can communicate with each other. I would post it there and be like, 'Hey, does anybody have any experience on this? Do you know a little bit more?' If they gave the okay and the user was commenting in good faith and having conversations with people and they didn't have any past history, like have they ever gotten a little slap on the wrist for whatever, then it would probably just stay up. But, I leave notes to the other moderators, 'Oh, I was kind of iffy about this one. If they keep it up in the future, it might be something to look at.' "-[P11]

The notes allow moderators to identify users who repeatedly skirt subreddits' rules. Most of the participants disclosed that their moderation actions escalated for handling these users. For high-stake situations such as *permanent ban*, a few moderators also said that they tend to discuss with fellow mods to reach an agreement.

Cross-community Support: Participants also sought support from moderators of other subreddits. For instance, P7 mentioned that, when evaluating the intent of a user, they would sometimes message moderators of other subreddits that the user participated in. More commonly, there are Discord channels for moderators of various subreddits to "bounce off each other"-[P14]. For example, P1 mentioned that they are a member of a Discord channel for location-based subreddits in the same geographic region. This provided an extra information resource for moderators in that region and a way to keep track of malicious users who post in those subreddits.

In addition to asking for support, a couple of participants also lent their own expertise to other subreddits. In these cases, the subreddits that sought external help are often non-COVID subreddits (e.g., r/starwars, r/hockey). Moderators of these communities had little experience regulating COVID-19 content and lacked the knowledge to separate facts from misinformation. The following is an excerpt from the interview with P3, who had helped moderators of other subreddits.

> "What we would do is, we would hop into Discord with them [moderators of other subreddits]. They gave us access to their general Discord that they use for coordinating and planning. And then, if they had questions on whether something that someone was talking about was actually true or not, they could just bring it to us. And, we could take a look at it and answer, 'Yeah, that's accurate to the best of our knowledge, or to the best of the scientific literature.' Or, 'No, that's absolute nonsense.' " -[P3]

Past work observed cross-community interactions between moderators in high-stake scenarios such as collective actions, or when moderators have certain affiliations such as moderating similar communities [48, 49]. Our analysis shows that moderators also seek external support for everyday incidents related to misinformation.

Outside of misinformation detection, some participants also noted that they rely on other affiliated moderators to create and implement new subreddit rules and to tweak their moderation workflow (e.g., implement auto-ban bots) to target COVID-19 misinformation. Sometimes, moderators (e.g., P1) temporarily volunteered to moderate other subreddits to assist with the influx of user activities. This is consistent with prior work [68], which demonstrates that affiliated moderators rely on each other to develop and update moderation practices.

4.3 Moderators' Feedback on Alternative Modqueue Designs

In Section 3.4.2, we introduced a design probe of a widget with multiple components and additional types of information to assist moderators with misinformation detection. In this section, we discuss participant feedback in detail. Among our key findings, we show that participants considered *expert labels* to be very useful for identifying content veracity. However, crowd signals are useful for determining the intent of the parties involved. Further, we also show that the perceived usefulness of information is dependent on participants' perceived credibility of the information providers. Moreover, participants also discussed that the data and metrics they currently rely on (or proposed in the design probe) for determining credibility are not always reliable and can be gamed. Finally, we synthesize participants' feedback to highlight some additional improvements to the designs.

4.3.1 Comparing the Values of Different Components. Almost all participants found the widget or some components of the widget to be useful for moderating misinformation. The vast majority of participants viewed *crowdsourced fact-checking*, *similar posts*, and *expert labels* favorably, and almost all participants appreciated *user popup* (components given in Figure 2). We also observe that *expert labels* were considered to be the most helpful but, surprisingly, only for a mere majority of moderators. Of the remaining participants, one participant thought that the widget was not that useful; the others were evenly divided between *crowdsourced fact-checking* and *similar posts* in terms of which component they considered to be the most useful. More interestingly, several participants also asserted that *crowdsourced fact-checking* and *similar posts* are useful for identifying user intent, something that *expert labels* cannot do.

> "If it's a fact-checking organization you trust, that's a very quick read. Okay. This is a source that I trust. They're deeming it false. I can act on this quickly. 'Is this user problematic?' Then, I would say, *crowdsourced fact-checking* would be a good avenue to go in, to really investigate what's the intent of this user, just based on the sources they are posting and the consistency of their posts. *Similar post* is interesting and helpful for reposts, and just people spamming subreddits." -[P18]

In other words, our analysis demonstrates that the extent to which a type of information is useful is dependent on whether a moderator needs help evaluating content facticity or user intent. Further, we observe that while some participants mentioned that they had more difficulty with subtle misinformation, others had more trouble with skilled trolls. This suggests that all three types of information are needed to cover different use cases.

4.3.2 The Role of Perceived Credibility in the Perceived Usefulness of Information. Our analysis shows that the perceived usefulness of signals provided by different entities (e.g., users and professional fact-checkers) is dependent on the perceived credibility of the entities involved. Indeed, we observe that participants preferred crowdsourced fact-checking that i) was posted by regular or expert

users, ii) contained links to reputable sources, and iii) explained why the reported content is misinformation.

> "Yeah. That [user reputation] would absolutely add weight to making a decision probably easier. Google does with their reviewers. If I see, like Wikipedia, somebody who's done lots of good articles that I would know, I would rely on their report quicker. Or, if I see somebody has a new account, then I'm gonna look at it a little closer." -[P10]

Results here are aligned with [53, 69]. Additionally, some moderators noted that bad-faith users could also "fact-check". These moderators stated that they are suspicious of fact-checking comments from new user accounts with low karma, and also wouldn't trust crowdsourced fact-checking containing unusual sources or sources masked using URL shorteners. A few participants, however, stated that the number of crowdsourced fact-checking alone is a good indicator. Finally, a couple of interviewees noted that they wouldn't automatically agree with the fact-checking from regular users. However, they would assume that these users contributed in good faith.

Similarly, the majority of participants thought it is useful to have access to other subreddits' crowdsourced fact-checking and moderator actions on *similar posts*. However, to some, it matters which subreddits they were. A few moderators said that they preferred signals from reputable COVID-19 subreddits such as r/coronavirus and r/covid19. In addition, a couple of interviewees said that signals from subreddits they considered to be unreliable would have the opposite effect.

> "If r/conservative are saying like, "Oh, this is fake." That's going to change how I interpret their comments. Something that they think is fake is what I know is true." -[P11]

Finally, a quarter of the participants had reservations about *expert labels* because they viewed these fact-checking organizations as being too politicized or lacking medical expertise. Many participants also mentioned that it would be difficult to find sources and professional fact-checkers that are broadly viewed as credible. Our own analysis suggests that participants generally trusted academic sources the most, followed by government sources such as the CDC. Though, one moderator remarked that the accuracy of information from government sources is dependent on who wins the elections.

Overall, results here are consistent with the findings in Section 4.2 and with prior work [30]. Participants preferred information from entities that they considered to be good-faith and credible. Those signals were also given more weight during the moderation process.

4.3.3 Limitations of Crowd Signals: Fabricated Credibility and Over-reliance on the Widget.

Anticipating Bad-faith Users "Gaming the System": Interestingly, participants also anticipated various ways that bad-faith actors can mislead moderators by fabricating credibility. Indeed, some interviewees emphasized that user account attributes, such as account age and karma scores, that moderators have been relying on to evaluate intent and credibility are not necessarily authentic. For instance, year-old Reddit accounts with moderate karma scores can be purchased cheaply en mass. Users can also sit on their sock accounts until the accounts reach a certain age. Bad-faith users can also obtain more karma from free karma subreddits (e.g., r/freekarma4u). Likewise, with *crowdsourced fact-checking*, a few moderators were concerned about users misrepresenting sources. For instance, bad-faith users can provide links to tangentially related articles from reputable sources that would not necessarily refute the reported content. For *similar posts*, some moderators emphasized that it takes very little to create a subreddit and become a moderator. Incompetent or even malicious moderators may accidentally or purposefully remove factual content, which could potentially lead to a damaging cascade.

> "Because then suddenly, someone could be posting something and with a certain amount of concerted effort, someone could trick a moderator into banning something that they shouldn't have and then they would cascade through. It would be helpful if it was a true positive in terms of misinformation that needs to be banned. But it makes the effect of a false positive heavy." -[P6]

Over-reliance on the Widget can Undermine Moderation: A couple of participants argued that moderators may also become too reliant on user attributes (shown in *user popups*) rather than analyzing each user from their own perspectives. These interviewees expressed concern that good-faith users may be treated unfairly.

> "There are subreddits I could go to and make what I believe are perfectly innocent, truthful, non-controversial comments, but it's against their narrative and I'll be swarmed with downvotes ... Again, it's useful to have quick information [user popup], but just people have to know that it doesn't mean that they're trolling." -[P15]

Similarly, many participants also indicated that they preferred making their own moderation decision rather than relying on the actions of other subreddits' moderators (shown in *similar posts*) because each subreddit has its own goals and rules.

4.3.4 Improving Modqueue Designs: Credibility-focus Approaches. Interestingly, our analysis shows that most of the recommendations made by participants to help improve the modqueue designs are centered on intent and credibility. First, participants proposed several additional metrics to be included in the user popup. These include large gaps in user activities, user account's country of origin, and user activities in free karma subreddits. A few moderators also expressed interest in knowing whether a user was banned from other subreddits and having access to notes of other subreddits' moderators. Finally, one moderator asked for more advanced NLP techniques that detect abrupt changes in an account's writing style and topics of interest. Participants asserted that these metrics could help them separate sock puppet accounts and bad-faith users from organic users. Next, a couple of participants also suggested only including moderation actions from known reputable subreddits such r/coronavirus. Finally, several moderators indicated that the widget could include an additional search bar for academic sources such as PubMed or PMC search. These moderators believed that academic sources are more credible than third-party fact-checkers.

Additional Limitations and Concerns: The most common concern participants had with the widget is its technical limitations and AI trustworthiness. For instance, P3 questioned whether an algorithm could accurately identify crowdsourced fact-checking. P16 mentioned that other types of user comments can also be indicative of misinformation.

> "You're gonna see a lot more people saying like F*** off to the person posting misinformation. They post other things, verbal cues that aren't quite as obvious." -[P16]

In addition to these technical limitations, moderators also expressed concern about the social aspects of the widget. Notably, one moderator asserted that they would prefer users not to engage directly with misinformation posters via commenting as it would likely lead to conflicts. Rather, users should flag misinformation using the report system. Similarly, a few moderators were concerned that Reddit admins preferred to "keep communities from spilling into one another and causing drama", raising concerns about the *similar posts* section of our widget. Finally, a couple of moderators also noted that ordinary users (and Reddit admins) might have user privacy concerns.

5 DISCUSSION

5.1 Improving the Efficacy of Crowd Wisdom on Reddit

In this paper, we found crowd wisdom—in the form of crowdsourced flagging and modteam support—to be a powerful tool for detecting COVID-19 misinformation. Most of the participants heavily relied on crowdsourced flagging to identify potential misinformation, and many also relied on inputs from other moderators for ambiguous cases. Furthermore, we also identified concrete strategies to improve the efficacy of crowd wisdom.

First, similar to prior work [30], we observed that the use of crowdsourced flagging is limited beyond identifying potential misinformation. This is likely because the current report system neither allows the reporters to provide evidence along with the reports nor affords reporters to express their degree of concern. We argue that platforms should enable reporters to explain their reasoning for flagging. Platforms should also allow reporters to express their concern level along with their reports. This can be accomplished by including additional input boxes, scales, or checkboxes in the report system. An important caveat is worth mentioning: additional user input requirements may reduce the report rate [41]. Future work should also explore how this trade-off can be managed.

Similarly, while Reddit has various subreddits (e.g., r/modsupport, r/modguide) specifically designed to support moderators, our interviews revealed that moderators are dependent on other platforms, including Discord and Slack, to seek and provide networked support. This includes check-ins with each other for specific, difficult cases. Further, our analysis also suggested that moderators of smaller subreddits have less access to these networks. Reddit and other social media giants should consider providing additional platforms (e.g., r/mod_discord_finder) such that isolated moderators may find other moderator teams that can assist them in various moderation aspects. Likewise, our study also highlighted the need for future work to explore designs that can help moderators better identify and recruit expert users to be a part of their modteam.

Next, the perceived intent and credibility of ordinary users and moderators alike are important factors that significantly moderate the perceived usefulness of their inputs. As such, incorporating metrics that are indicative of credibility in the moderation workflow can improve the efficacy of crowd wisdom for misinformation detection. Specifically, many participants called for the reporting system to include additional information on the reporter (e.g., account age, karma, and whether or not the reporter is a regular member of the community). Moderators can use this information to better evaluate whether the reports are made in good faith. Similarly, moderators also only sought assistance from other moderators that they are affiliated with and trust. Likewise, the perceived usefulness of crowd-based signals in the widget is also dependent on whether or not these signals come from reputable users and moderators. For instance, some interviewees considered *crowdsourced flagging* by reputable users in their communities to have more weight in their decision-making. Lastly, platforms may be hesitant to make user information available in the report system or elsewhere due to privacy concerns. However, our analysis showed that platforms could provide various useful metrics without revealing the identity of these users. For example, mods mentioned wanting metrics like the number of reports made per user, which lets mods more easily decide whether they would take a report seriously without disclosing the reporter's username.

Finally, and interestingly, we also observed sophisticated anticipation from participants in that they had also identified ways that credibility-informing metrics can be gamed. Moderators suggested additional user account metrics that they believed to be harder to game. These metrics include sudden gaps in users' activities or abrupt changes in users' linguistic patterns. Though, how to best implement these account metrics at scale is still an open question.

5.2 One Size Doesn't Fit All

Our findings in Section 4 encompassed the moderation practices of many of the participants. Here, we reiterate the variations in moderation and highlight the complexity and the myriad of ways in which moderation happens.

Similar to related work [43, 68], we observed that moderation practices are dependent on moderator characteristics and subreddit attributes. Indeed, moderators' tech-savviness and trust in technology affected the extent to which they employed algorithmic tools to moderate misinformation. Similarly, moderators' expertise influenced how much they relied on their own knowledge to identify misinformation. Moderators' personal beliefs and philosophies also shaped the occurrence and severity of moderation actions they take after identifying misinformation. Likewise, community goals and rules affected whether and how misinformation was moderated. In particular, our analysis suggested that strict rules (e.g., not allowing posts containing non-reputable sources, not allowing off-topic content) and dedicated moderation enforcing the rules helped participants manage COVID-19 misinformation.

Overall, results show that the recommendations we made to improve misinformation moderation will likely have varied effects across different communities. For instance, moderators with low COVID knowledge may find labels from professional fact-checkers more useful than moderators with high COVID knowledge. Moderators who overlook repeated problematic user behaviors to preserve free speech will likely find less value in signals that are indicative of intent. Likewise, communities that do not moderate misinformation as that would contradict their goals will benefit minimally–if at all–from our proposals. Finally, moderators of communities with high levels of report abuse may benefit significantly more from transparent report systems such as the one proposed in this work.

For communities with moderation practices that significantly diverge from the norm, future work should explore other ways to better assist them with misinformation moderation. For instance, moderators who value free speech and are avoidant of punitive actions against problematic users indicated that they were, nevertheless, wary of bots that spread misinformation. As such, more advanced bot detection and removal tools can help these moderators reduce misinformation while still preserving free speech for human users. Additionally, researchers should also explore how these moderators can more effectively use non-punitive actions to regulate misinformation and its spreaders. For example, our study suggested that these moderators preferred to educate users and correct their misconceptions. Future work should therefore examine how moderators can best correct misinformation posted by users without having unintended consequences (e.g., causing a backfire effect [46]).

Finally, we note that while one size doesn't fit all [36], it does fit many. We discuss the generalizability of our findings in Section 5.4. The high-level findings and concrete strategies from our study can help many online communities with detecting misinformation.

5.3 Distrust in Fact-checking Services and the Limitations of Knowledge-based Misinformation Detection

One of the findings that we found surprising is the level of participant distrust in fact-checking services. That is, a quarter of all participants explicitly expressed their distrust in third-party fact-checkers. Among them, there were no noticeable differences between those moderating political subreddits versus non-political ones. Additionally, we also didn't observe a clear relationship between the participants' expertise and their distrust of third-party fact-checkers. Some of the participants considered fact-checking organizations to be polarized and biased. Others were concerned about fact-checkers lacking domain expertise, thus misinterpreting or mislabeling some subtle but

important medical information. While the broader public's distrust in third-party fact-checkers was observed in past literature [55], our study revealed that such distrust also exists in a significant portion of online platform decision-makers. This result has important implications for misinformation detection on Reddit. Most notably, moderators who do not trust labels from professional fact-checkers may have to rely more on their own research and knowledge, the wisdom of their communities, or even the Reddit platform itself to detect misinformation. As such, we argue that future work focused on improving crowd wisdom for misinformation detection could be especially valuable for these communities [17].

Alternatively, while there is likely no universal fact-checking site that all moderators trust, there seemed to be a hierarchy of sources that interviewees rank from the most reputable to the least reputable. Generally, participants tend to consider academic sources as the most authoritative. This is followed by official information from public health institutions. Though, moderators differ in whether they place more trust in federal institutions or local leadership. Given these considerations, Reddit could include facts and fact-checks from various sources, both academic and non-academic, and then let moderators choose which sources they rely on.

Beyond skepticism, our study demonstrated another limitation of professional fact-checking services. That is, while most participants considered labels from professional fact-checkers to be valuable for identifying content facticity, these labels cannot assist moderators with evaluating user intent. Interestingly, academic scholars have contended that intent is highly subjective and difficult to quantify [52]. Yet, platform stakeholders are nevertheless using various user characteristics to interpret intent and use this inferred intent to choose moderation actions. Given these considerations, we argue that future work in misinformation detection should also explore methods to identify the intent of misinformation posters in addition to content veracity. In particular, we observed that moderators often rely on a user's community associations and past behavior patterns to assess intent. Arguably, algorithmic models can mirror the practices of these moderators. For instance, academic researchers and industry practitioners can use social network analysis [24] to classify a user's community affiliations. Future work can also explore training automated models using features generated from past user behaviors and using these models to predict future user behaviors. Lastly, our studies only revealed some of the heuristics used by participants to discern user intent. Future work should more systematically identify such heuristics and develop methods to assist moderators with intent detection.

5.4 Generalizing Findings to Other Misinformation Domains

To what extent do our findings and strategies generalize to other types of misinformation? As mentioned in Section 4.1.1, for a majority of the participants, moderation of COVID-19 misinformation is similar to other types of problematic content. This suggests that, for many subreddits, the workflow model in Figure 3 is generalizable to other misinformation domains and to other types of problematic content or behavior. For instance, some moderators mentioned that their moderation practices for incivility and vitriolic users are comparable to those for misinformation. A few moderators also emphasized that they had limited time and resources and thus chose to employ a consistent set of protocols to handle all content without having to think about the types of content. In other words, our study suggests that researchers can adopt the workflow model from Figure 3 as the baseline for future research focused on understanding and improving online community moderators' moderation processes of problematic content and user behavior, such as incivility, trolling, and harassment.

Next, focusing on the differences in moderation processes, some participants noted that their moderation priority and difficulty varied across different domains (e.g., political vs. COVID-19 content) of misinformation. Though, there is no clear directionality. For instance, some moderators

commented that they had more difficulty with COVID-19 misinformation, whereas others said the opposite. Some participants similarly mentioned that they were more strict (e.g., taking more severe moderation actions) with COVID-19 misinformation, whereas others said that they were more lenient. Our analysis didn't reveal any clear indicators that can explain this variance. The differences highlighted here suggest that while moderators may rely on the same subprocesses (Figure 3) to identify various problematic content and user behavior, how they use these subprocesses likely differ based on the types of content. For instance, communities may choose to employ advanced automated approaches to help moderate hate speech because it has a high priority. But, the same communities may opt to rely on simple algorithms to identify potential COVID-19 misinformation because they consider it a low priority.

Lastly, we also observed that many participants had gone through unique knowledge acquisition steps in order to moderate COVID-19 misinformation. Particularly, as the pandemic progressed and more information became available, moderators needed to continuously update their knowledge on what's misinformation and what's not. A few others, however, noted that things had settled down more recently. Finally, a couple of moderators also commented that they had more difficulty digesting medical and scientific information than political content. Our analysis suggested that moderators needed a certain level of domain-specific literacy and knowledge in order to moderate misinformation of different domains.

6 LIMITATIONS AND FUTURE WORK

We note the following limitations in our study. First, while we covered a range of participants and subreddits, we had limited participation from conservative-leaning subreddits. Despite our best efforts, we were not able to obtain interview opportunities with these subreddits. It's possible that moderators of conservative-leaning subreddits have distinct experiences with crowd wisdom (e.g., they may have encountered more misreports [19] due to inter-community conflicts). Similarly, all but one of the participants identified as male. Female moderators may experience different or additional moderation challenges unobserved by our study [23]. Second, our analysis is focused exclusively on COVID-19 misinformation. While we are confident that many of the high-level findings are generalizable to other types of misinformation, additional future work is needed to explore the extent to which moderation practices vary across different information domains. For instance, we observed some evidence that moderators encountered more misreports on political content. This suggests that crowdsourced flagging is less reliable for detecting political misinformation. Similarly, our study was conducted during a time when there was no significant development (i.e., external shocks) in COVID-19. A few of our participants noted that their reliance on crowd wisdom had changed over time, and their moderation practices also varied amid external events. For instance, a few moderators considered users to be less knowledgeable and reliable early on than later in the pandemic. Given these considerations, future work should also explore the temporal variations in the reliability of crowd wisdom and its implications for misinformation detection in self-governing online communities. Next, we also highlight that our results are based on semi-structured interviews. Using quantitative analysis on larger-scale datasets would be interesting and important to complement our findings. For instance, future work can quantitatively examine the frequency of report abuse across different communities and determine which community-level attributes are indicative of report abuse. Likewise, we also relied on participants' self-reports to rate their level of COVID-19 knowledge. Future work can use more objective evaluations to assess moderators' topic expertise and research how different levels of expertise could impact moderation practices. Finally, we note that this work is largely focused on improving misinformation detection from the perspective of online community moderators. Yet, misinformation detection is only the

first step in the broader objective of combating misinformation. Much more work is needed to examine its characteristics and find the best ways to curtail its prevalence and influence.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (Grant IIS-2045432). The authors thank Bradley Lott and Ashwin Rajadesingan for their valuable feedback on the interview protocol.

REFERENCES

[1] Vlad Achimescu and Pavel Dimitrov Chachev. 2020. Raising the Flag: Monitoring User Perceived Disinformation on Reddit. *Information* 12, 1 (Dec. 2020), 4. https://doi.org/10.3390/info12010004
[2] Samantha A Adams. 2011. Sourcing the crowd for health services improvement: The reflexive patient and "share-your-experience" websites. *Social science & medicine* 72, 7 (2011), 1069–1076.
[3] Reddit Admin. 2020. Misinformation and COVID-19: What Reddit is Doing. https://www.reddit.com/r/ModSupport/comments/g21ub7/misinformation_and_covid19_what_reddit_is_doing/
[4] Hadeel S Alenezi and Maha H Faisal. 2020. Utilizing crowdsourcing and machine learning in education: Literature review. *Education and Information Technologies* (2020), 1–16.
[5] Jennifer Nancy Lee Allen, Antonio Alonso Arechar, Gordon Pennycook, and David Gertler Rand. 2020. *Scaling Up Fact-Checking Using the Wisdom of Crowds*. preprint. PsyArXiv. https://doi.org/10.31234/osf.io/9qdza
[6] Monica Anderson and Dennis Quinn. 2019. 46% of US social media users say they are 'worn out'by political posts and discussions. (2019).
[7] Philip Ball and Amy Maxmen. 2020. The epic battle against coronavirus misinformation and conspiracy theories. *Nature* 581, 7809 (2020), 371–375.
[8] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 215–236.
[9] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020), 1–26. https://doi.org/10.1145/3415164 arXiv: 2008.09533.
[10] Dustin P Calvillo, Bryan J Ross, Ryan JB Garcia, Thomas J Smelter, and Abraham M Rutchick. 2020. Political ideology predicts perceptions of the threat of COVID-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science* 11, 8 (2020), 1119–1128.
[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 1–30. https://doi.org/10.1145/3359276
[12] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–25. https://doi.org/10.1145/3274301
[13] Marina Charquero-Ballester, Jessica G Walter, Ida A Nissen, and Anja Bechmann. 2021. Different types of COVID-19 misinformation have different emotional valence on Twitter. *Big Data & Society* 8, 2 (2021), 20539517211041279.
[14] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance* 6, 2 (2020), e19273.
[15] Michele Coscia and Luca Rossi. 2020. Distortions of political bias in crowdsourced misinformation flagging. *Journal of The Royal Society Interface* 17, 167 (June 2020), 20200020. https://doi.org/10.1098/rsif.2020.0020
[16] Kate Crawford, Tarleton Gillespie, and Nms /new Media. 2014. What is a flag for?
[17] Ryan Cross. 2021. Will public trust in science survive the pandemic. *Chemical and Engineering News* 99, 3 (2021).
[18] Milmo Dan. 2021. Reddit bans Covid Misinformation Forum after 'go dark' protest. https://www.theguardian.com/technology/2021/sep/01/reddit-communities-go-dark-in-protest-over-covid-misinformation
[19] Srayan Datta and Eytan Adar. 2019. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, Vol. 13. 146–157.
[20] Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 133–142.
[21] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[22] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300372

[23] Bryan Dosono and Bryan Semaan. 2020. Decolonizing tactics as collective resilience: Identity work of AAPI communities on Reddit. *Proceedings of the ACM on Human-Computer interaction* 4, CSCW1 (2020), 1–20.

[24] David Easley, Jon Kleinberg, et al. 2012. Networks, crowds, and markets. *Cambridge Books* (2012).

[25] Matthias Eickhoff and Jan Muntermann. 2016. Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom. *Information & Management* 53, 7 (2016), 835–845.

[26] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Detecting misleading information on covid-19. *Ieee Access* 8 (2020), 165201–165215.

[27] Ziv Epstein, Gordon Pennycook, and David Gertler Rand. 2019. *Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources.* preprint. PsyArXiv. https://doi.org/10.31234/osf.io/z3s5k

[28] Casey Fiesler. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. (2018), 10.

[29] Sarah A. Gilbert. 2020. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. https://doi.org/10.1145/3392822

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

[31] Rex Hartson and Pardha S Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience.* Elsevier.

[32] Sohyeon Hwang and Jeremy D Foote. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[33] Waleed Iqbal, Gareth Tyson, and Ignacio Castro. 2022. Looking on Efficiency of Content Moderation Systems from the Lens of Reddit's Content Moderation Experience During COVID-19. *SSRN Electronic Journal* (2022). https://doi.org/10.2139/ssrn.4007864

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

[35] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 150:1–150:27. https://doi.org/10.1145/3359252

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

[37] Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. 2020. Modeling and measuring expressed (dis) belief in (mis) information. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 315–326.

[38] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–36.

[39] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and Opportunities in Online Science Communication. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300383

[40] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (Jan. 2020), 1–35. https://doi.org/10.1145/3375197

[41] Anja Kalch and Teresa K Naab. 2017. Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. (2017).

[42] Simranpreet Kaur, Pallavi Kaul, and Pooya Moradian Zadeh. 2020. Monitoring the dynamics of emotions during COVID-19 using Twitter data. *Procedia Computer Science* 177 (2020), 423–430.

[43] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an" Eternal September" How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1152–1156.

[44] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schoelkopf, and Manuel Gomez-Rodriguez. 2017. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. *arXiv:1711.09918 [cs, stat]* (Nov. 2017). http://arxiv.org/abs/1711.09918 arXiv: 1711.09918.

[45] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359,

6380 (2018), 1094–1096.
[46] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.
[47] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1867–1870.
[48] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1138–1151. https://doi.org/10.1145/2858036.2858391
[49] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (April 2019), 2056305119836778. https://doi.org/10.1177/2056305119836778 Publisher: SAGE Publications Ltd.
[50] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (May 2019), 9785–9789.
[51] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. *arXiv:2011.05773 [cs]* (Nov. 2020). http://arxiv.org/abs/2011.05773 arXiv: 2011.05773.
[52] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist* 65, 2 (2021), 180–212.
[53] Teresa K Naab, Dominique Heinbach, Marc Ziegele, and Marie-Theres Grasberger. 2020. Comments and credibility: how critical user comments decrease perceived news article credibility. *Journalism studies* 21, 6 (2020), 783–801.
[54] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.
[55] Sakari Nieminen and Lauri Rapeli. 2019. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review* 17, 3 (2019), 296–309.
[56] Deven Parekh, Drew Margolin, and Derek Ruths. 2020. Comparing audience appreciation to fact-checking across political communities on reddit. In *12th ACM Conference on Web Science*. 144–154.
[57] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci USA* 116, 7 (Feb. 2019), 2521–2526. https://doi.org/10.1073/pnas.1806781116
[58] P. Pourghomi, F. Safieddine, W. Masri, and M. Dordevic. 2017. How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach. In *2017 International Conference on Engineering MIS (ICEMIS)*. 1–8. https://doi.org/10.1109/ICEMIS.2017.8272957 ISSN: 2575-1328.
[59] Nicolas Pröllochs. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv:2104.07175 [cs]* (May 2021). http://arxiv.org/abs/2104.07175 arXiv: 2104.07175.
[60] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. 2021. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 525–536.
[61] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. 2021. Informed crowds can effectively identify misinformation. *arXiv preprint arXiv:2108.07898* (2021).
[62] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2021. Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19. *arXiv:2107.11755 [cs]* (July 2021). http://arxiv.org/abs/2107.11755 arXiv: 2107.11755.
[63] Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. sage.
[64] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z. Tan, and Amy X. Zhang. 2021. Modular Politics: Toward a Governance Layer for Online Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 1–26. https://doi.org/10.1145/3449090
[65] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–28. https://doi.org/10.1145/3415178
[66] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.
[67] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (March 2022), 621–640. https://doi.org/10.1177/1461444820964968
[68] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316 Publisher: SAGE Publications.

[69] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2022. If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 896–907.

[70] Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *CLEF (Working Notes)*.

[71] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907* (2020).

[72] Monika Skaržauskaitė. 2012. The application of crowd sourcing in educational activities. *Social technologies* 2, 1 (2012), 67–76.

[73] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*. 259–263.

[74] Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for COVID-19 disinformation categorisation. *PloS one* 16, 2 (2021), e0247086.

[75] Franklin Tchakounté, Ahmadou Faissal, Marcellin Atemkeng, and Achille Ntyam. 2020. A Reliable Weighting Scheme for the Aggregation of Crowd Intelligence to Detect Fake News. *Information* 11, 6 (June 2020), 319. https://doi.org/10.3390/info11060319

[76] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. *arXiv:1711.09025 [cs]* (March 2018). http://arxiv.org/abs/1711.09025 arXiv: 1711.09025.

[77] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, Ann Arbor MI USA, 275–284. https://doi.org/10.1145/3209978.3210037

[78] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2012. Social Turing Tests: Crowdsourcing Sybil Detection. *arXiv:1205.3856 [physics]* (Dec. 2012). http://arxiv.org/abs/1205.3856 arXiv: 1205.3856.

[79] Claire Wardle, Hossein Derakhshan, et al. 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco* (2018), 43–54.

[80] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[81] Jason Shuo Zhang, Brian C. Keegan, Qin Lv, and Chenhao Tan. 2020. Understanding the Diverging User Trajectories in Highly-related Online Communities during the COVID-19 Pandemic. *arXiv:2006.04816 [cs]* (June 2020). http://arxiv.org/abs/2006.04816 arXiv: 2006.04816 version: 1.

[82] Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5075–5086.

Received July 2022; revised October 2022; accepted January 2023