"There Has To Be a Lot That We're Missing": Moderating AI-Generated Content on Reddit

TRAVIS LLOYD, Cornell Tech, USA
JOSEPH REAGLE, Northeastern University, USA
MOR NAAMAN, Cornell Tech, USA

Generative AI has begun to alter how we work, learn, communicate, and participate in online communities. How might our online communities be changed by generative AI? To start addressing this question, we focused on online community moderators' experiences with AI-generated content (AIGC). We performed fifteen in-depth, semi-structured interviews with moderators of Reddit communities that restrict the use of AIGC. Our study finds that rules about AIGC are motivated by concerns about content quality, social dynamics, and governance challenges. Moderators fear that, without such rules, AIGC threatens to reduce their communities' utility and social value. We find that, despite the absence of foolproof tools for detecting AIGC, moderators were able to somewhat limit the disruption caused by this new phenomenon by working with their communities to clarify norms. However, moderators found enforcing AIGC restrictions challenging, and had to rely on time-intensive and inaccurate detection heuristics in their efforts. Our results highlight the importance of supporting community autonomy and self-determination in the face of this sudden technological change, and suggest potential design solutions that may help.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Online Communities, Reddit, AI-Generated Content, Generative AI, Moderation

1 INTRODUCTION

Recent advances in generative AI [1, 66, 72] have led to the proliferation of AI-generated content (AIGC) across the internet [84], causing online communities to adjust their policies and practices. Generative AI use, whether disclosed or not, may alter the social dynamics of online communities. For example, AI-generated text with subtle inaccuracies poses a challenge for knowledge-sharing communities, like Stack Exchange [55], and communities where members engage in cooperative work, such as Wikipedia [28]. In addition, the perceived inauthenticity of AIGC can be disruptive to social communities where people come together for human connection, such as some Facebook groups [25]. This disruption adds to the existing challenges faced by online communities, such as coordinated disinformation campaigns, which generative AI may make cheaper, more effective, and better targeted [31].

As many communities rely on content moderators for stewardship [32], it is important to understand and document moderators' firsthand experiences responding to AIGC. Effective moderation can help communities stay civil during times of disruptive change by clarifying and enforcing community norms [44, 45, 80]. In this new age of generative AI, it is likely that existing moderation strategies will need to adapt in order to remain effective.

This paper qualitatively investigates the experience of moderating AIGC in self-governing communities, in order to explore one dimension of generative AI's impact on online communities. Not all communities will see AIGC as a concern, but we focus on those that do to illuminate how platform designers and community stewards might support such communities in the face of dramatic technological change [79]. How to provide such support is a natural question for the HCI

Authors' addresses: Travis Lloyd, tgl33@cornell.edu, Cornell Tech, New York, NY, USA, 10044; Joseph Reagle, j.reagle@northeastern.edu, Northeastern University, Boston, MA, USA, 02115; Mor Naaman, mor.naaman@cornell.edu, Cornell Tech, New York, NY, USA, 10044.

and Social Computing research communities, a central goal of which has been to encourage healthy communities as they strive to meet a variety of individual and collective needs [13, 45, 47, 48].

To this end, we conducted fifteen in-depth, semi-structured interviews with Reddit moderators. We focus on Reddit, a social sharing and news aggregation site, so that we can compare and contrast relatively independent self-governing communities that, nonetheless, share a platform. Reddit is a popular platform to study [70] because of its variety of communities and distributed governance structure: subreddits set and enforce their own rules to supplement site-wide policy [12, 26]. Reddit thus allows us to explore how different attitudes and practices towards AIGC organically emerge across communities. We use a qualitative investigation to explore perspectives on what constitutes AIGC and what makes it problematic or challenging. Our goal is to better understand the concerns of moderators in communities that have restricted the use of AIGC, not to provide a representative view of all Reddit moderators or communities. Our work thus explores the following research questions:

- **RQ1:** What are volunteer community moderators' attitudes towards AIGC in communities that restrict its use?
- **RQ2:** What experiences have these moderators' had with AIGC in their communities?

Our exploration of **RQ1** reveals that our participants restrict the use of AIGC because they are concerned that it may negatively impact the content quality, social dynamics, and governance processes in their communities. Many participants say that AIGC is antithetical to their communities' purposes and values. Though our participants acknowledge some potential benefits of AIGC, they generally oppose its use and see several specific threats to their communities. Participants say that AIGC is poorly written, inaccurate, and off-topic, and thus threatens to reduce the quality of content in their communities. Our participants also report that AIGC threatens to impact community social dynamics by reducing opportunities for human connection, straining community relationships, and violating shared community values. Finally, AIGC may increase the scale and sophistication of existing types of problematic behavior, making the jobs of moderators more difficult.

Our exploration of **RQ2** reveals that even though some participants feel that clarifying rules and norms has rendered AIGC no more threatening than other problematic content, AIGC has increased the workload on already overburdened moderators, as rules about AIGC are hard to enforce. Moderators believe they can detect AIGC, but the process is time-consuming, as they rely on imperfect heuristics that look for certain "tells" in content and behavior. Moderators sometimes use automated detection tools to help with enforcement, but none are reliable enough to fully automate the process. Additionally, moderators acknowledge the sensitivity of accusing a community member of posting AIGC, and emphasize the care that goes into such decisions.

We discuss the implications of these findings for members of online communities, community moderators, and platform designers. We explore the relationship between community values and attitudes towards AIGC and suggest that the arrival of AIGC will drive communities to seek out new ways of demonstrating authenticity. We discuss platform designs that can encourage authenticity and keep members aware of evolving community norms towards AI. We also discuss the potential for bias in moderators' current detection approaches and explore design solutions that could aid this task by synthesizing various relevant sources of information. Finally, as platforms begin to integrate generative AI features, we suggest that they make these features optional so that communities can retain agency over the types of interactions that they want to encourage.

2 RELATED WORK

Understanding the challenges posed by AIGC requires an understanding of online communities, their health, and the challenges of moderation. The literature, below, shows that that online

communities vary in their reasons for forming and their values; this suggests that community attitudes towards AIGC will be context-dependent. The literature on community health highlights the challenges of dealing with deception, low engagement, and harassment; these might be amplified by AIGC. Finally, research emphasizes the challenging nature of content moderation work; the difficulty of identifying and addressing AIGC might further strain moderators and their relationships with their communities.

2.1 Online Community Types and Values

Online communities form for different reasons, have different values, and thus may feel threatened by AIGC in different ways. Early work by Preece [68] coined the term *empathic communities* to refer to sites of both emotional and factual communication, where member participation is motivated by a desire for something that professionals cannot provide: authentic, firsthand accounts from others with similar lived experiences. Recent work has shown that these communities can perceive certain types of interactions as inauthentic that are completely acceptable in other contexts [82]. We suspect that AI-written responses may be seen as problematic in empathic communities due to their lack of lived experience. Lave and Wegner provide another useful concept, the *community of practice* [49], which denotes a group of people united by a desire to improve their skills by doing, sharing, and receiving feedback from one-another. Since these communities deeply care about process, they might regard contributions as inauthentic if they are made using generative AI rather than their traditional practice. A final useful concept is Jenkins' idea of *knowledge communities* in which "members work together to forge new knowledge, often in realms where no traditional expertise exists,"[40]. These communities value accurate knowledge and expertise, which may be threatened by generative AI's current tendency to produce inaccurate "hallucinations" [2]. Kraut and Resnick supplement this conceptual literature with practical considerations for platform designers [47] hoping to encourage such communities through systems of trust and reputation [74]. These systems of trust may need to evolve to account for a world where contributions made with AI are indistinguishable from those produced by expert community members.

Reddit has been a frequent topic of study for empirical investigations into the different types and values of online communities [70]. Reddit's hierarchical governance structure, in which a set of site-wide policies are supplemented by community-derived and enforced rules, allows it to host many independent communities with distinct purposes and values. This model of community self-moderation [79] allows researchers to study evolving attitudes and practices as they emerge. For example, recent work on the use of governance bots has shown how new governance practices can impact Reddit community dynamics by affecting members' sense of virtual community [81]. Other research has studied community values across the platform through surveys of community members [89], analysis of public subreddit rules [26], and logs of of moderation actions [12]. Findings from these studies emphasize the diverse set of values held by different subreddits. While they identify some commonly occurring rules and values, such as rules against harassment and the importance of trust, they emphasize that rules are largely context-dependent. For example, Fiesler et al [26] found that subreddit topic is more predictive than size when predicting the types of rules that a subreddit will have. Recent work introduced the framework of *Community Archetypes* to describe commonly occurring types of communities on Reddit [69], which we use to make sense of the variation in community stances towards AIGC. On the whole, the findings from this body of work suggest that Reddit will not have a unified stance towards AIGC and motivates our inquiry into the specific perspectives of moderators from communities with rules governing its use.

2.2 Generative AI and Threats to Healthy Online Communities

The HCI literature has identified varied threats to online communities which might be amplified by the use of generative AI, such as deception, low engagement, and harassment. Past studies have shown online communities to be vulnerable to strategic influence operations, and recent speculative work has hypothesized that generative AI will increase the volume and believability of such efforts [2, 31]. Indeed, early empirical studies have found AIGC to be as or more deceptive than human-written content [67] and detected AIGC in fake online reviews [64] and scams [20]. Other studies of intentional deception have focused on the capability of algorithmically controlled social media accounts, or *social bots*, to deceive unsuspecting users. [24, 61]. While LLM-powered social bots have already been detected on Twitter [92], our work looks beyond social bots to include situations where *humans* use generative AI tools to create content. We hypothesize that humans posting undisclosed AIGC may be regarded as a subtler form of deception. Findings from AI-Mediated Communication (AIMC) studies suggest that AI affects interpersonal communication [35], causing people to regard others more negatively [71] and with less trust [38] when they perceive that the other is using AI. Past HCI research has explored interface designs that encourage closeness and authenticity by enabling *effortful communication*, or communication that demonstrates effort and care [95], between users. It remains to be seen if such design techniques can offset the negative effects that AIGC may have on interpersonal communication.

Another threat to online communities is that of declining participation. The HCI and CSCW literature provides techniques for encouraging community contribution [47] and commitment [73], such as establishing a group identity that allows strong ties to form between members. However, specifically studying how *AIGC* affects participation in online communities is a relatively new area of research [88]. Several studies of sites in the Stack Exchange network have demonstrated that the launch of ChatGPT caused an overall decline in website visits, question volume, and number of users, while increasing the complexity of questions asked [8, 77, 91]. The authors attribute these effects to the relative ease and speed with which users can get simple questions answered by LLM-powered chatbots off-platform. Interestingly, the authors of one study [8] contrast this effect with a measured null-effect on similar Reddit communities, which they suggest are able to avoid a decline in activity because of their emphasis on socialization, rather than pure information exchange. Recent work has also studied the effects of top-down bans on AIGC and found that banning the use of generative AI on Stack Exchange [86] led to a decrease in volume and quality of questions and answers. Our study builds on this line of inquiry by asking if the emergence of AIGC and consequent bans threaten participation in self-governing communities with varied values and needs.

Finally, online communities are often sites of harassment and abuse, a frequent topic in the CSCW literature [14, 34]; the effects of AI-powered variants, such as malicious synthetic, or deepfake, media [23, 75] are much less studied. Given the potential for AI-powered abuse, our work explores how community moderators are thinking about these threats and what they are doing to counter them.

2.3 The Challenges of Volunteer Community Moderators

Subreddits that decide to enact rules about AIGC will turn to already-strained volunteer moderators to enforce these rules [51, 90]. Moderation involves both "visible" and "invisible" work [30], leading researchers to alternatively characterize it as "civic labor" [58], "emotional labor" [21], and "care work" [29]. The HCI community has a long history of addressing the needs of moderators [42], which suggests that research can contribute design solutions specific to AIGC. While existing research

has focused on handling harmful content such as misinformation [4], the specific challenges of moderating AIGC have not been studied.

One perpetual challenge for moderators is maintaining alignment between themselves and their community members [46]. Given the power dynamic between moderators and regular members, perceptions of unfair or biased moderation decisions can strain community relationships. Perceptions of bias are not unfounded: existing work has shown that individuals with marginalized identities are more likely to be the target of moderation actions [33, 83]. We suspect this will be a problem when moderators enforce rules about AIGC, as research on perception of AIGC has found that humans cannot reliably distinguish between AI and human-generated text [15] or images [63], and use flawed heuristics to make such determinations [39]. Though little is known about moderators' AI detection practices, recent work has shown that general determinations about AI use are biased according to race and gender stereotypes [62]. Our study builds on this work by exploring the real-world AI detection practices of community moderators.

The CSCW research community has researched and developed computational tools meant to aid moderators in their work [11, 42, 46, 78, 94]. However, detecting AIGC with computational techniques has, so far, proven challenging [93]. While commercial classifiers for AI-generated text claim to be effective [17], recent research on their practical application questioned their theoretical best-case performance [76]. Other work has found these classifiers to perform no better than random chance [87], or even worse, to be biased against the writing of non-native English speakers [52]. This research questions the feasibility of a reliable, general-purpose detection tool that will work on any snippet of text. Still, some studies have had success identifying AIGC in specific types of writing, such as news articles [36], conference peer reviews [53], and crowd-worker output [85], suggesting that such techniques might be adapted to other contexts. Our study will investigate how moderators think about these tools and currently incorporate them into their workflows.

3 METHODS: INTERVIEW STUDY

We chose to conduct semi-structured interviews with Reddit moderators, approaching the research questions qualitatively, to deeply "explore and understand the meaning individuals or groups ascribe to a social or human problem" [16], specifically the newly introduced challenges of AIGC. We identified target communities by looking at their public stances towards AIGC, and recruited participants via outreach and snowball methods.

3.1 Recruitment

As our research goal was to explore moderator concerns, we sought to speak with moderators of communities who had restricted the use of AIGC. To identify such communities, we turned to public data about subreddits' community rules. We performed a crawl of the Reddit website in the summer of 2023 and extracted the text-based rules from all subreddits indexed by Reddit's "Top Communities" list[1], which includes public subreddits with more than ten subscribers. We crawled $n = 337,399$ subreddits and found $n = 87,596$ with non-empty rules and English as a primary language. We then performed a text search on the rules to identify subreddits with rules related to AIGC. While this method may be imprecise, it was sufficient for our purpose, which was to understand broad patterns to guide our recruitment strategy.

In the data that we gathered, 4% of all subreddits had rules about AI, but these rules were much more common in larger subreddits: when we restricted our analysis to top 1% of subreddits according to subscriber count, 20% had rules governing the use of AI. We thus designed a recruitment strategy focused on the largest subreddits, which appear to be the most concerned with the potential impact

[1]https://www.reddit.com/best/communities/1/

of AIGC. Many moderators of larger subreddits also moderate small forums and were able to speak to the differences between these types of communities.

We recruited and interviewed participants during the summer of 2023. We manually examined the rules of subreddits in descending order of subscriber count. When we found a subreddit with either an explicit rule governing the use of AI or a rule that might implicitly cover AI use we added it to our sample. We also looked for moderator discussions about AIGC in other public forums, such as the r/ModSupport subreddit, and added those moderators' subreddits to our sample. We then messaged the moderators of each subreddit in our sample via Reddit's Modmail[2] feature, sending Modmail recruitment messages to 60 subreddits. We used two additional techniques to recruit participants beyond the largest subreddits. First, we conducted snowball sampling and asked moderators to forward our recruitment message to other potential participants. Second, we posted our recruitment call on several social media platforms. To ensure respondents were actual moderators, we asked for their Reddit username and the subreddit that they moderate, then verified that the users were indeed listed as moderators via their public Reddit user profile. Participants were offered $20 gift-cards as compensation. We conducted interviews on a rolling basis and continued recruiting until we had enough rich and complex data to adequately address our research questions, as per Braun and Clarke's saturation criteria for Reflexive Thematic Analysis [7]. Overall, we interviewed fifteen participants, meeting or even exceeding the sample size norms of the HCI research community [9]. These participants collectively moderated over 100 different subreddits with sizes ranging from less than ten members to more than 32 million (see participant information in Table 1).

3.2 Interviews

We designed an initial interview protocol based on our research questions and updated it as themes emerged during early interviews. Since pseudonymity is an important feature of the Reddit platform, we followed best practices in ethical online communities research [27] and took special considerations to protect the privacy of our participants. We did not ask for or record participants' names, demographic information, or any offline identifiers. As text is the primary mode of interaction on Reddit, we gave participants the opportunity to interview via text-based mediums, which is a common practice in Reddit and Social Computing research [19, 30, 42]. Accordingly, participants were asked to choose a medium that was comfortable for them for a 60-minute, synchronous interview: video call, audio call, email exchange, or message exchange via the Reddit platform. Regardless of channel, we posed our research questions one at a time, in real time, and followed up in a semi-structured way. For example, for text-based interviews, we were online at the same time as our participants, exchanging messages in a back-and-forth manner, which allowed us to probe participants' responses as necessary for deeper insight or clarity. The study was approved by our institution's IRB. All participants read and accepted a consent agreement permitting their responses to be reproduced in research reports along with the participants' associated subreddits. To ensure that we reliably represented participants' views, we contacted participants for feedback and approval on their included quotes before publication, following CSCW community recommendations [60]. We were able to integrate all participants' feedback in the final manuscript.

The first author conducted the interviews using these key guiding questions:

(1) What types of AIGC are moderators seeing in their communities?
(2) What do moderators think are the motives of AIGC posters?

[2]https://support.reddithelp.com/hc/en-us/articles/210896606-What-is-Modmail-

(3) How are moderator and non-moderator community members responding to this new phenomenon?
(4) What concerns do moderators have about AIGC and how do they compare to other moderation concerns?
(5) How do moderators identify AIGC?
(6) What could help moderators address these concerns?

For interviews conducted via audio and video calls, interview transcripts were generated automatically from recordings. The transcripts were manually reviewed and corrected by the first author. For the four interviews conducted via email and the four conducted via Reddit message, the text was captured in similarly formatted transcript files for analysis. Interviews lasted between 26 minutes and 147 minutes with a mean and median of 71 and 62 minutes, respectively. While we scheduled 60 minutes for the interviews, some oral interviews ended early and some text-based interviews went longer, as participants took their time to respond to our questions: oral interviews had a mean duration of 47 minutes ($SD = 15$), compared to a mean duration of 91 minutes ($SD = 38$) for written interviews. Despite the semi-synchronous nature of the longer written interviews, all yielded meaningful conversations with at least eight back-and-forth question and answer exchanges. Written interviews served as a valuable complement to our oral conversations by providing rich data that was not shared via oral conversation, such as links to example AIGC posts from participants' communities. Despite the differences between the mediums, both yielded deep and meaningful data that combine to create a more nuanced picture than any single medium could provide.

The first two authors then collaborated to iteratively code the interview content into categories, using an inductive thematic analysis method akin to Braun and Clarke's Reflexive Thematic Analysis [5, 6]. The two authors began by discussing themes that the first author observed while conducting the interviews and correcting the transcripts. From this conversation, an initial high-level set of codes emerged. Next, the first author open-coded all of the interview transcripts at the sentence-level, applying the initial codes and creating new ones where existing codes were insufficient. The two authors then reviewed the set of codes and refined them by merging similar codes and nesting those with a clear hierarchy. This round of coding did not produce new themes, only new sub-codes, suggesting that the themes were fairly stable. Both authors then independently applied these codes to the same three interview transcripts and compared their results. Neither application had created any new codes and the few cases of disagreement were resolved by further merging similar codes. This process yielded approximately 400 tagged phrases across all transcripts, which were grouped into the three high-level themes that we discuss below.

4 FINDINGS

Three main themes emerged from our interviews. With regards to RQ1, our interviews surfaced insights into moderators' (1) concerns about AIGC in their communities; answering RQ2, our interviews elicited details on (2) how moderators' communities are responding to AIGC and (3) the challenges of enforcing AIGC rules.

Our interviews did not reveal a single definition of "AIGC". Instead, our participants mentioned three different categories of content related to our inquiry: images produced by generative AI, text produced (entirely or partially) by generative AI, and posts made by automated accounts (bots). Communities sometimes banned one of these categories while allowing others.

Participants generally used the term "AI-generated" in a pejorative way, but some acknowledged that AI-use falls along a spectrum, from benign to problematic, and shared the subjective factors

Table 1. Details for each interview. For each participant, we note the largest subreddit (according to subscriber count at the time of the interview) and the number of individual subreddits that they moderated.

grayheight**Interviewee**	**Largest Moderated Subreddit (Subscriber Count)**	**Subreddits Moderated**	**Interview Medium**	**Primary Threat**
1	r/todayilearned (32M)	25	Audio Call	Amplified Attacks
2	r/food (23M)	25	Reddit Message Exchange	Lowered Quality
3	r/explainlikeimfive (23M)	3	Email Exchange	Less Accurate Information
4	r/explainlikeimfive (22M)	2	Audio Call	Less Accurate Information
5	r/WritingPrompts (17M)	3	Video Call	Fewer Opportunities for Skill Development
6	r/politics (8.3M)	25	Reddit Message Exchange	Strained Community Relationships
7	r/CryptoCurrency (6.7M)	8	Reddit Message Exchange	Amplified Attacks
8	r/itookapicture (5.1M)	2	Email Exchange	Fewer Opportunities for Skill Development
9	r/NintendoSwitch (5.0M)	23	Email Exchange	Lowered Quality
10	r/Fantasy (3.4M)	1	Email Exchange	Lowered Quality
11	r/changemyview (3.4M)	2	Video Call	Less Human Connection
12	r/changemyview (3.4M)	2	Audio Call	Less Human Connection
13	r/AskHistorians (1.8M)	4	Video Call	Less Accurate Information
14	r/AskHistorians (1.8M)	1	Reddit Message Exchange	Less Accurate Information
15	r/GCTrading (22.7K)	3	Audio Call	Amplified Attacks

that they considered when judging AI use. Some participants spoke of the *intention* of a poster[3], like a moderator of r/itookapicture[4] who said: *"I don't mind people using AI to better state their thoughts nor create art as long as it is not being used deceptively."* Others referenced the *degree* to which a poster relied on AI, like a moderator of r/ChangeMyView who reported being OK with a post made with AI, *"if it was sufficiently [a poster's own] words, just sort of juiced up by ChatGPT."* We heard something similar from a participant who cared that a post was a poster's *"own intellectual contribution or content"*:

> ***r/AskHistorians moderator:*** *So say, for example, we have an expert… perhaps they're an expert on German history, but they don't speak English all that well. So they write their answer in German, and then use ChatGPT to try to translate it… it is their own intellectual contribution or content, they've created this themselves, but then they just use this as a tool in a perfectly viable way.*

Other participants took a less nuanced view, like a moderator of r/WritingPrompts, who relayed their community's official stance: *"Let's be absolutely clear: you are not allowed to use AI in this subreddit, you will be banned"*. Despite hearing arguments that, *"ChatGPT is another writing tool that authors are using and banning ChatGPT will be the same as banning a spellchecker,"* this participant disagreed and categorically rejected all uses of AI, going so far as to say that, *"[AIGC] is antithetical to everything that [our] subreddit is about."* The simplicity of this stance has its appeal, as more subjective definitions of AIGC introduce the possibility that community members will disagree about specific examples—an implication which we will revisit in later sections.

4.1 Findings: Moderator Concerns About AIGC (RQ1)

Our recruitment focused on moderators of communities that are actively restricting the use of AIGC. Accordingly, while our participants offered differing definitions of AIGC, they generally opposed its use in their communities. We group our participants' reported AIGC concerns into three, often overlapping, categories: content quality concerns, social dynamics concerns, and governance

[3]We use the term *poster* to refer to both those posting top-level subreddit submissions as well as those posting comments in response to submissions.

[4]To provide additional context, we introduce quotes with the name of the subreddit that the participant moderates. In the case where the participant moderates several subreddits, we include either the subreddit that the quote is referring to or, for more general quotes, the largest subreddit that the participant moderates.

Table 2. Summary of our participants' main concerns about AIGC in their communities.

Concern	Examples
Low Quality Content	Poor prose, factual inaccuracy, off topic
Impacted Social Dynamics	Fewer authentic connections, strained relationships, ideological opposition
Difficult to Govern	Increased volume of community attacks, harder to detect deception

concerns. Despite general opposition, participants also acknowledged some potential benefits of AIGC, detailed below.

Table 2 summarizes the three main categories of concerns that our participants shared. Content quality concerns take issue with the quality of AIGC and worry that it decreases the utility of an online community. Social dynamics concerns focus on how AIGC reduces the social value that an online community provides its members. Governance concerns emphasize the ways in which AIGC makes the job of moderating online communities more difficult. Naturally, these categories overlap and interact, and our participants often described considerations that span multiple categories.

Moderators shared various perspectives on the magnitude of these concerns. When asked if AIGC was one of their top concerns, a moderator for r/CryptoCurrency noted that, *"Early 2023, I would've said yes, but right now it has calmed down quite a bit."* On the other hand, some participants did regard AIGC as a major concern:

> ***r/news moderator:*** *The concern is the more weaponized stuff… the folks who are utilizing these tools for very focused purposes, often for political reasons, or for commercial reasons, those are the ones that I think concern us the most. Whether or not somebody manages to get a post up there that doesn't belong for a few hours, most of us don't care… But what we do care about is creating a veneer of legitimacy for bad actors.*

We will explore the challenges of detecting AIGC in Section 4.3, but first we dig deeper into the types of concerns that our participants shared.

4.1.1 Content Quality Concerns. First and foremost, moderators were concerned about the quality of AIGC. Nine participants reported that AIGC did not meet the quality standards of their subreddits, like a moderator of r/AskHistorians who said that AIGC, *"tries to meet the substance and depth of a typical post… however, there are frequent glaring errors in both style and content."* Participants shared several reasons that they considered AIGC low quality, such as their issues with its style, its tendency to be factually inaccurate, and the belief that it was off-topic.

Several participants reported that there were certain stylistic issues with AIGC that made it low-quality content. We heard from one moderator that AI-generated answers to questions in their community often, *"do not address the question being asked… tend to be very general and 'hedge' more than a real human."* Other moderators put it more bluntly, like a moderator of r/CryptoCurrency who noted that, *"[AIGC] is not enjoyable to read"*, or a moderator of r/Eldenring who told us: *"[AIGC] is content that provides no value/discussion at all, it's low effort content basically."* This moderator considered effort to be a signal of value or quality, which we will revisit in the discussion. They went on to share that this sort of low effort content can, *"[discourage] users who want to post their own original content"*.

Four participants who moderate communities that emphasize the accuracy of information (e.g. r/AskHistorians and r/explainlikeimfive) specifically mentioned their concern about inaccurate AI hallucinations. This concern is captured by a moderator of r/explainlikeimfive who objected to AIGC on the grounds that ChatGPT was a *"bullshit generator"*, or by another participant who shared that AIGC's inaccuracy was unacceptable in their communities:

> ***r/AskHistorians moderator:*** *History and the work thereof is something that I'd say AI will continue to find impossible for the foreseeable future, until such time as it can gain a sense of truth, or at least coherence.*

This participant described AIGC as *"truthy-seeming"* because of its tendency to include fake citations and felt that the appearance of being thoroughly researched increased the chances that readers would believe it, which, they feared, would *"undermine the trust that people have [in the community]."*

Finally, we heard from several moderators that AIGC was simply off-topic in their communities. For example, a moderator of r/itookapicture shared that the goal of their community was, *"sharing great photography and photography techniques,"* and accordingly, *"preventing the submission of AI generated images and human-created compositions alike are a top priority."*

4.1.2 Social Dynamics Concerns. Of course, online communities are not only a source of high-quality content, but also of meaningful social interactions. We heard from ten participants who were concerned that AIGC would negatively impact the social dynamics in their communities, reducing the social value that their communities provide to members. Participants mentioned several ways that this could happen, such as by decreasing opportunities for human connection, straining relationships, and violating their communities' values.

Ten participants reported that AI went against their communities' purpose, which was to share posts created by humans. These participants mentioned the inherently inhuman nature of AIGC and expressed concern that this would lead to less authentic connection in their communities. These were not necessarily objections to the quality of AIGC, but rather to the *process* through which it is created:

> ***r/explainlikeimfive moderator:*** *In our case, we do want someone to write these explanations. And we don't want it done by some AI that has no clue what it's talking about. And even if it does give an accurate answer, the purpose of this site is for people to write in their own words … there's already Google, there's already Wikipedia.*

These participants believed people come to their community to connect with others through dialogue and that AIGC does not meaningfully contribute to—or even detracts from—such conversations. A moderator from r/changemyview summarized this sentiment well: *"How can we change your view when you haven't actually stated your view at all? This is ChatGPT's view. And if I wanted to, I could go argue with ChatGPT. But that's not helping here."* We heard something similar from six moderators of communities of practice (e.g r/itookapicture, which is devoted to photography) who felt AIGC did not contribute to the collaborative learning that the community aimed to encourage. These are communities where members seek to develop particular skills by doing, sharing, and receiving feedback from other practitioners. Accordingly, posts produced *without using* their practice did not provide educational value to the community. This concern was summarized by a moderator of r/WritingPrompts who shared that their community existed to, *"[give] writers a chance to practice their craft and practice their skill. And if you are taking away the practice element of that, because all you're doing is feeding a prompt into ChatGPT… The server is no longer serving its purpose."*

Additionally, we heard concern from participants that AIGC could strain relationships in their community. For example, a moderator of a political subreddit noted that: *"the possibility [of AIGC] fuels Redditors accusing other comments of being written by AI (which is a form of incivility)."* Beyond incivility, the controversial nature of AIGC could cause discussions to devolve into off-topic arguments:

> ***r/Zelda moderator:*** *Many comment sections would fill with the same discussion points about the controversies of AI art… We would often see comments that attacked people for using AIGC, as well as attacks on people for gatekeeping art.*

Finally, some participants took an ideological stance and objected to the use of AI and, by extension, AIGC, because they saw it as incompatible with their communities' values. For example, we heard from a moderator of a creative subreddit who objected specifically to the way that generative AI models are trained:

> ***r/<removed>*[5] *moderator:*** *Perhaps most importantly, as stewards of a creative space, we feel it's our duty to support the real human artists, authors and other creatives whose work has been exploited to train the vast majority of these models without their knowledge and without credit or compensation.*

4.1.3 Governance Concerns. Participants also objected to AIGC because it made the job of moderating their communities more difficult. We heard from eight participants who saw AIGC being used for malicious purposes and were concerned that existing moderation techniques would struggle to handle it. These moderators spoke of pre-existing problems, such as spam and harassment, that are more difficult to manage when they are amplified in scale or sophistication by AIGC. While AIGC may be less common than other moderation challenges, it can still be quite disruptive:

> ***r/explainlikeimfive moderator:*** *It's not our most common removal, by far. But personally, I would rate it as the most threatening concern…It's often hard to detect and we do see it as very disruptive to the actual running of the site.*

Five participants, who moderated the largest subreddits, spoke of AIGC being used in broad attacks on their communities. These attacks often involved automated accounts posting large amounts of spam. These attacks may aim to disrupt the functioning of a community through sheer volume, as we heard from one moderator who experienced a ChatGPT-powered spam attack that required the help of Reddit staff to resolve:

> ***r/AskHistorians moderator:*** *A few months ago, we had an instance where we were subject to a bot attack that was using ChatGPT… we ended up reporting it to the admins who were able to take care of it through whatever they do on their back end.*

Other times these attacks aim to deceive users as a part of influence operations, as we heard from a moderator of r/news, one of the platform's largest subreddits, who shared their experience dealing with, *"a highly sophisticated deployment, an attempt to sway a narrative by utilizing accounts that had had their karma bolstered by automated commenting."* This participant references Reddit *karma*[6], a reputation points system that communities can use to limit who can post, noting that AIGC can be used to bypass this system through a practice known as *karma-farming*, in which automated accounts make frequent posts in order to rapidly accumulate karma. Karma-farming is not a new practice, but this moderator thought that AIGC made it harder to detect, and thus more effective. Beyond karma-farming, we also heard about AIGC being used in other forms of deception:

> ***r/CryptoCurrency moderator:*** *We've seen some 'shills' (aka companies or group of people) coming with AI content to promote their product…And of course promoting is detrimental, as it is often disguised as a 'research' post about that specific coin (for example), making it seem like a legit user researching a subject and that might influence buyers in comparison to a regular ad.*

[5]We use *<removed>* to disassociate subreddits from participant quotes either at the participant's request or when we find the quote together with the subreddit name potentially too revealing.
[6]https://support.reddithelp.com/hc/en-us/articles/204511829-What-is-karma

In addition to broad attacks on a community, we also heard from participants who were concerned about AIGC being used for targeted attacks on individual members. One moderator of r/politics reported that they associated AIGC with, *"usually some form of trolling, or 'outsourcing' writing a comment the user wanted to make."* Another participant also mentioned trolling, though they said that existing moderation solutions for this problem would be sufficient:

> ***r/changemyview moderator**: The one thing that I am worried about AI is that it does make trolling easier. You can generate larger amounts of trolling text and make it seem reasonable or seem like somebody wrote it for a lot longer…But…it's not that hard to tell when somebody's trolling on our subreddit anyway. So it makes trolling easier [to do], but it's something that I don't think we're going to be too stressed handling.*

Finally, we heard concerns about AI making it easier to post objectionable content. One participant who moderated r/Zelda shared that they had seen AI used to produce specific types of objectionable content that their community had already banned, such as commercial or NSFW content. They noted that while this type of content was not new, *"it seemed to be new territory for people to expedite these things via AI."*

4.1.4 Potential Benefits of AIGC. While our recruiting approach and research questions primarily surfaced an overall attitude of opposition, several moderators mentioned ways that AIGC could potentially *benefit* their communities by increasing engagement. These moderators perceived a legitimate, or at least good-faith, motivation behind posters who used AIGC to complement their knowledge and writing ability:

> ***r/AskHistorians moderator:** We're a very popular subreddit, and posts tend to be held up as examples of excellent content on Reddit. Some people may want to be a part of it, but not have the specific knowledge or resources to be able to contribute 'properly'. Hence they turn to AI tools, hoping to provide a decent answer.*

Three moderators mentioned that AI could help non-native English-speakers improve their posts. We heard from a moderator of r/changemyview who thought that, *"AI as a tool is going to be useful…it's useful for people who are poor writers, useful for people who aren't strong English speakers. And when they need help, AI can fill in those gaps,"* though they noted that it was important that such a contribution was still, *"sufficiently their words"*. Of course, the definition of "sufficient" will be context dependent and may change over time as norms around AI use become more established.

One last benefit that we heard was that AIGC could increase engagement in an inactive conversation. For example, one moderator speculated about the potential benefits of AIGC for platforms and provided a hypothetical example:

> ***r/<removed> moderator:** So a subreddit that needs a little bit of life breathed into it, [Reddit] deploys some AI there and make a variety of posts. And then people comment and they see a whole bunch of ads. And everyone's happy, right? A weekly automated thread isn't necessarily a bad thing.*

This moderator was not saying that their community had done this or that it aligned with their values, but rather that they believed this approach could be useful in certain contexts.

On the whole, our evaluation of the interview transcripts suggests that despite these potential benefits, all of our participants felt that, at least for now, the negative aspects of AIGC outweigh the good.

4.2 Findings: Online Communities' Responses to AIGC (RQ2)

To counter the perceived threats of AIGC, moderators have enacted specific changes in their communities. First, our participants emphasized collaborating with their community members to

clarify norms about the use of AIGC. Once clarified, these norms could be codified into rules, which moderators were then tasked with enforcing.

4.2.1 Clarifying Community Norms. Most participants spoke of establishing a community stance towards AIGC through a collaborative and iterative process involving community members and moderators. Often it was the moderators who started the discussion, but other times it was the community:

> ***r/museum moderator:*** *It was the users who protested. Mildly: it's a fairly polite subreddit. They said, we don't believe that this belongs here. And so we had a blanket ban on it for a little while. And then we slowly allowed it to be integrated. And then there was a counter protest [against the integration].*

Most participants reported they wanted to ensure that their stance on AIGC was aligned with their communities' interests. One participant said that their community initially allowed AIGC before eventually voting to ban it:

> ***r/CryptoCurrency moderator:*** *At first (late 2022 early 2023) we let people use AI to write partial content in their posts *if they mentioned it*. But lots of abuse, people not mentioning it, and just overall quality decreasing, so [the community] decided to "ban" AI use.*

This participant went on to describe their community's process for discussing issues, surfacing proposals for solutions, and soliciting proposal votes from all community members—a process that they used here to enact a rule banning AIGC.

Almost all participants mentioned eventually codifying community norms into public subreddit rules. Some communities used rules that explicitly banned AIGC or bots, while others used more general rules (such as rules against plagiarism) as effective AIGC bans. Participants shared mixed results about the efficacy of these rules. We heard from seven participants that the volume of AIGC had decreased in response to new rules, including from a moderator of r/AskHistorians who said, *"People are using [AIGC] a little bit less now that they know what the rules are. They know that people don't necessarily want it."* However, a moderator of r/explainlikeimfive expressed skepticism about the efficacy of rules: *"I think it may stop some of the average users (to the extent that they actually read the rules, which is a reddit-wide problem.) Bots obviously don't read rules."* As this participant suggested, explicit rules can guide the behavior of good-faith users, but might not deter bad actors.

4.2.2 Enforcing Community Rules. All of our participants' communities devised—and tried to enforce—rules about AIGC. However, the difficulty of detecting AIGC meant that enforcement of such rules was often challenging. Three moderators spoke of AIGC increasing their workload:

> ***r/AskHistorians moderator:*** *It's not fun to deal with having to assess whether or not something is [AIGC]. Reading through it like that, it's a lot of extra work, right? Because you have to read through it all the way. It's so fast for people to input this kind of content and a lot slower for people to assess it.*

This imbalance between the effort to produce AIGC and to read it, given AIGC tends to be verbose, puts moderators at a disadvantage. Also, given how alienating a false accusation can be, moderators must take extra care with their decisions; a moderator of r/explainlikeimfive shared, *"You don't want to be banning someone for using GPT when they don't actually use it. So we might have to watch someone for a bit of time."*

Our participants used a variety of strategies to try to make enforcement of these rules less burdensome. Seven moderators mentioned sharing the work with community members, like a moderator of r/news who noted, *"The first line of defense is the users themselves."* However, this runs

the risk of causing incivility between members in the case of false accusations, as mentioned in Section 4.1.2.

Four moderators spoke of their reliance on platform tools, especially *automoderator*[7], a programmable bot that scans and performs actions on all new posts. Such automated tools lessened the manual effort required by moderators, but they are only as effective as the detection heuristics programmed into them, which, as we discuss in the next section, can be flawed. We heard from moderators that such tools are not enough to catch all AIGC, and in cases of extreme post volume, such as during coordinated bot attacks, moderators sometimes sought help from Reddit platform admins:

> ***r/news moderator:*** *There's the stuff that's publicly available to every user, then you have the stuff that's available to Moderator accounts, within the subreddit that the moderators are working on, and then you have the admin stuff—which is much more sophisticated than what we have.*

All of these enforcement techniques ultimately require making decisions about what is and is not AIGC. Next we discuss our participants' reported methods for making these difficult decisions.

4.3 Findings: Methods for Detecting AIGC (RQ2)

Our participants shared both custom heuristics and technical solutions that they use to detect AIGC. Even if participants thought their detection techniques mostly work now, they worry that their tools will be less effective as generative AI improves:

> ***r/news moderator:*** *We see the obvious stuff and we prune the obvious stuff… but there has to be a lot that we're missing. And I imagine that it's gonna get more and more sophisticated over time.*

4.3.1 Detection Heuristics. There are a number of signals that moderators believe help them identify AIGC. To avoid aiding bad-faith actors, we do not list specific tells, but they broadly fell into the following categories: recognizable language patterns in posted content, details of users' accounts and behavior, deviation from a community's style norms, and inaccurate information.

Content Signals. Twelve participants mentioned that certain patterns in a post's content, such as keywords, phrases, or distinct forms, may make moderators suspect that a post is AIGC. A moderator of r/Eldenring noted that AI-generated text is, *"very repetitive i.e it tries to justify a point but instead of doing that, it ends up repeating the same point over and over to the point of ad nauseum."* A moderator of r/itookapicture shared that with AI-generated images, *"there is often something that feels a bit unnatural… The images look like the average of something, rather than a unique individual with all of the flaws that come with that."*

User Signals. When confronted with suspect content, six participants reported that they often consider the details of a poster's account, including their past posting behavior. One red flag is a dramatic change in the style of a poster's prose. For example, a moderator of r/changemyview shared that suspicious posts were, *"often very different in terms of writing from what the user posted in response to other users questions or comments."* Additionally, four moderators, like a moderator of r/Zelda, noted that signals that identify more traditional bots are also relevant for identifying AIGC: *"accounts used for spam… there are other indicators to find those (account age, history,…) Even before ChatGPT, we would find copy-pasted or markov-generated comments with these same account indicators."*

[7] https://www.reddit.com/wiki/automoderator/

Deviation From Subbreddit Style Norms. Four participants reported using their knowledge of their communities' particular communication norms to identify atypical posts as AIGC:

> ***r/AskHistorians moderator:*** *It feels very formal, and it feels very different from the normal kinds of comments. There's definitely sort of a genre to answering a question on r/AskHistorians that people who have been around for a while tend to all follow, just because there are certain ways to communicate historical knowledge in an open online space to a group of non experts.*

Inaccurate Information. Four participants reported using their domain knowledge to identify posts as false or incoherent, which to them suggests AIGC:

> ***r/museum moderator:*** *In r/museum… there is a Canon, so to speak, of material. So it's not like there's suddenly going to be a brand new, you know, 16th century French academic painter that nobody's ever heard of. And so something that seems incongruent to that situation would obviously raise an eyebrow.*

4.3.2 Technical Tools. Though they acknowledged such tools' unreliability, most participants used automated approaches to detect AIGC, including both third-party, black-box detection tools and community-developed solutions. Participants spoke of programming the detection heuristics mentioned above into tools such as *automoderator*, which apply the heuristics to every new post and flag anything suspect. Six participants mentioned manually copying posts into third-party detection tools to see if they were AIGC. A moderator of r/writingprompts reported that they had tested, but never deployed, a homespun bot that would apply third-party detectors automatically to all new posts.

At the same time, a moderator of r/changemyview mentioned that because they are aware of the unreliability of third-party detection tools, they rarely defer to them absolutely: *"I don't necessarily trust the tool–that's why we have several layers.… We always try to talk to the user."* Two participants mentioned that the tools were least effective on shorter text, like a moderator of r/writingprompts who noted that one tool, *"was very bad with short comments. If a comment was like 100 words or a really short story–every poem got flagged as AI. And like, No, it's not. It's just short."*

5 DISCUSSION

Our qualitative investigation offers two main research contributions: (1) a description of moderators' attitudes towards AIGC in subreddits that restrict its use and (2) insights into how these moderators and their communities have responded to the challenges of AIGC. On the whole, our findings align with those of recent studies into the impact of generative AI on online communities [8, 88], which emphasize that the *social value* of online communities is more important than ever in the age of generative AI. Platform designers and community stewards should focus on ways to increase this social value, and one way to do this is by empowering communities to take their own stance on AIGC.

As generative AI usage grows, we place our work in dialog with Seering's call to the CSCW research community to center community self-moderation [79], especially their second "guiding question": *"What are the processes of context-sensitivity in online community moderation, and how might they be better supported?"* We believe our findings demonstrate that allowing communities to make context-sensitive decisions about AIGC is central to their self-determination and should be an important topic for CSCW research. To offer implications for design, we expand below on the values that shape community attitudes towards AIGC as well the importance of clear and enforceable community norms.

5.1 Community Values Shape Attitudes Towards AIGC

Our study reveals a spectrum of moderator attitudes that vary with the purpose and size of communities. We expect certain objections to AIGC will lessen over time, while others will be more persistent.

5.1.1 Community Types and AIGC Attitudes. Our participants' concerns about AIGC were largely based on the *types* of communities they moderated. Accordingly, it is useful to consider our results through the lens of several online community types from the Social Computing literature, such as *knowledge communities* [40], *empathic communities* [68], *communities of practice* [49], and the five *Community Archetypes* presented by Prinster et al. [69]: Topical Q&A, Learning & Perspective Broadening, Social Support, Content Generation, and Affiliation with an Entity. We suspect that AIGC's threats to the social value of a community (Section 4.1.2) would be the most concerning to communities that emphasize authenticity [82], like *empathic communities* [68], or communities that fit the Social Support archetype. Moderators of *knowledge communities* [40] like r/AskHistorians and r/explainlikeimfive, which also fit the Topical Q&A archetype, also shared this concern, though they additionally took issue with AIGC's inaccuracy (Section 4.1.1). Arguably the clearest objections came from moderators of *communities of practice* [49] like r/itookapicture and r/WritingPrompts, which fit the Content Generation archetype. These communities shared both content quality and social dynamics concerns about AIGC, though we expect new communities of practice will emerge specifically devoted to AI use[8]. Moderators of larger, more impersonal communities that fit the Learning & Perspective Broadening archetype, such as r/news, r/CryptoCurrency, and r/politics, primarily voiced governance concerns (Section 4.1.3). We suspect this is because the size of these communities make them a target of bad actors. Finally, we primarily heard ideological objections from the communities in our sample that fit the Affiliation with an Entity archetype. We suspect this is because in our sample these are Content Generation communities (r/WoW, r/Zelda), and it remains to be seen if communities affiliated with non-media entities, like geographical places or organizations, would share similar concerns.

On the whole, our participants rejected AIGC because they felt the main purpose of their communities was to share posts created by humans—this was true across all of the community types that we encountered (of course, remember that our sampling strategy focused on communities with concerns, not representativeness). Future work should further explore the relationship between community type and attitude towards AIGC. For example, it would be interesting to see, quantitatively, whether different types of communities tend to create different types of rules about AIGC.

5.1.2 Attitudes May Shift Over Time. Our findings indicate that community norms about AIGC are still in flux. Perhaps over time people will come to treat generative AI the same as tools like auto-complete or spellcheck (a possibility we discussed in Section 4). If so, process-based objections may fade away. Of course, the opposite could happen as well: as generative AI becomes more pervasive people may seek out communities of collective refusal. In order to support community self-determination, we should design platforms that allow for both possibilities.

Additionally, generative AI will continue to change, and in time the content-based objections that we heard may no longer be relevant. For example, we consider objections based on AIGC's accuracy and prose quality, as well as those based on unethical model training processes, to be potentially addressable by technological advances. If online communities are to remain relevant, moderators will need to stay abreast of their members' stance on this evolving technology; ultimately, if moderator and member perspectives on AIGC differ too much, Reddit's *distributed moderation* system allows

[8]https://www.reddit.com/r/aiArt/

users to "agree to disagree" by leaving and finding other, more aligned communities [32]. At the time of writing, some platforms have begun incorporating generative AI into their products [54], likely in an effort to increase engagement. Though this might be aligned with platforms' profit motives, our interviews suggest it is not aligned with the preferences of all users or communities, as the low effort required to use these features will likely increase the amount of AIGC on the platforms. As we mention in Section 4.1.4, generative AI might bring certain benefits, but given the strong objections that we've seen, attempts by platforms to integrate generative AI should be done via a collaborative design process involving users [30, 57, 59]. We suggest that platforms that host a variety of community types, such as Reddit, give individual communities the choice to opt-out of generative AI features. If communities choose to opt out, this decision could be communicated to posters via the UI, perhaps by displaying the tools in a disabled state with an informative tooltip. Signaling a community's stance on AIGC in this way can have the added benefit of additionally discouraging posting AIGC produced off-platform. Additionally, for platforms designed for just one of the community types that we discussed, we recommend considering this community type's specific values when deciding whether to add generative AI features. Otherwise, platforms run the risk of alienating users [82].

5.2 The Challenges of Establishing and Enforcing Norms

Our findings relate to and reflect on research about the power of norms to shape behavior in online communities [10, 12, 45, 50]. We explore the specific challenges of establishing and enforcing community norms about AIGC and discuss ways in which platform designers might be able to aid community moderators with these tasks.

5.2.1 Clarifying Norms in a Moment of Flux. Our interviews show that self-governing online communities are adapting to AIGC by clarifying norms and codifying these into rules. Prior research has shown that online communities often make rules in response to sudden changes [80] and that this response can be an effective way to navigate challenging transitions [44]. Moderators sometimes struggle to identify AIGC, which makes enforcing rules about it a challenge; even so, discussing and clarifying rules can help inform the behavior of well-intentioned community members.

Community norms around AIGC are in flux and likely will be for some time. Community members and moderators may have different understandings of whether or not AIGC is acceptable, and this misalignment [46] can result in well-intentioned users inadvertently violating community norms. Communities can lessen the chance of misalignment by making norms more explicit and known to their members, which has been shown to be effective with other moderation challenges [56]. This approach is supported by participants' reports that AIGC has decreased since they enacted rules banning it, although these reports are based on the moderators' perceptions, and have not been independently verified. Platform designers can help by updating UIs to ensure that well-meaning posters are aware of community rules. For example, platforms could update their authoring UIs to display a community's rules, perhaps with a checkbox that users must mark to indicate that they have read and will abide by the rules, similar to how privacy policies are commonly presented. Another option would be to display the rules as placeholder text in the UI element where posts are written, which may remind users of norms at the time of posting [44]. That said, displaying rules in a creative way may not be enough, as past research has demonstrated that privacy policies are rarely read [65] and many posters simply do not read a community's rules before posting [41]. Communities may also consider more thorough "onboarding processes" to ensure new members are familiar with community rules, such as an interview with an existing member. However, this

approach adds an additional time burden and might deter potential members who are wary of such a commitment.

5.2.2 Equitably Enforcing AIGC Bans. Our findings detail some of the challenges that moderators currently experience enforcing AIGC restrictions and bans. Even before the AIGC challenge, moderators were stretched thin trying to protect their communities from problematic content (e.g. mis- and disinformation) and behavior (e.g. harassment, incivility, and influence campaigns). With detection of AIGC being an unsolved problem [22], moderators rely on heuristics to enforce restrictions or bans. Our findings show that these heuristics can be time-intensive. Even worse, they may be biased, as other work has shown that people's intuitions about AIGC are wrong in predictable ways [39] and that certain social groups are more often suspected of AI use [43]. Such consistent bias could disproportionately affect minorities and individuals seeking legitimate participation [52], such as non-native speakers, newcomers, and gender or racial minorities. It is concerning that most of our participants were confident in the efficacy of their heuristics and unaware of any potential bias. Our participants' reliance on these heuristics suggests that moderators should be educated about potential bias and further studies are needed to evaluate the efficacy of detection heuristics in content moderation. Platforms may be able to help moderators by adding tools for AIGC detection. Earlier work demonstrated that "synthesized social signals," or "social signals computationally derived from an account's history," help community members evaluate the credibility of other accounts [37]. Similar tools may be useful for detecting AIGC. For example, showing information about a poster's interactions with the system, such as time spent editing a post, could be helpful, though this also has user privacy implications. Surfacing a poster's past behavior and other community memberships could also be relevant, as these have been shown to be useful in identifying misinformation posters [4, 18].

Given the combination of concerns about AIGC and the lack of tools to detect it, we expect to see communities develop their own, non-technical, means for demonstrating authenticity. For example, past research into anonymous online communities [3] found several such practices, such as *triforcing* (posting complex Unicode symbols only meaningful to the community) and *timestamping* (posting photos of oneself holding a piece of paper with the current date and time on it). These techniques signal in-group membership and authenticity by presenting something that only an authentic poster can produce. These forms of *effortful communication* [95] signal the poster's effort, which makes them seem more valuable, given the relationship between effort and value mentioned by a participant in Section 4.1.1. In the age of low-effort AIGC we expect to see online communities place an even greater premium on effortful communication, which platform design can encourage. While these social innovations may emerge naturally from within a community, platforms can provide technical affordances to support their adoption, and the HCI research community can help.

6 LIMITATIONS AND FUTURE WORK

Our qualitative study is based on interviews with moderators of communities that are concerned with AIGC use and are taking steps to curb it. This recruiting criteria supports our goal of understanding some of the attitudes and practical considerations of Reddit moderators who are adapting to AIGC. We note, however, that our study and results are *not* meant to provide a generalizable view of how all Reddit moderators and communities consider AIGC.

Our study purposefully omitted those communities that have fully embraced AIGC, as well as those that have not yet created rules governing its use. Our recruitment additionally focused on popular subreddits in order to target those that are currently most affected by AIGC. However, our participants often also moderated smaller subreddits and were able to share how their experiences

and concerns are relevant to communities of various sizes. Additional studies could include moderators from a broader set of communities to provide a more complete picture of the range of ways that online communities are responding to AIGC, and to further explore the relationship between a community's values [69] and its stance on AIGC. We hope that future quantitative work will complement our own, by investigating a wider range of communities and making stronger inferences about the platform as a whole. For example, future work could use surveys or computational content analysis to test how our qualitative findings generalize across subreddits of different size and type. Further, since AIGC is an internet-scale phenomenon, future studies should explore how moderators are responding on other platforms, which may have different purposes and norms.

7 CONCLUSION

We performed a qualitative investigation into volunteer Reddit moderators' experiences with and attitudes towards AI-generated content. To better understand how CSCW researchers can support community self-determination we focused on communities that have restricted the use of AIGC. We heard from moderators that they are concerned about AIGC impacting the content quality, social dynamics, and governance processes in their communities, despite also seeing its potential to make it easier for some community members to contribute. In response, communities have enacted bans, which moderators enforce using time-intensive and imperfect heuristics. Despite enforcement challenges, our moderators report that, for now, their rules suffice, as they serve to clarify community norms about AIGC use.

As generative AI usage grows, the CSCW community needs to pay attention to how it affects the dynamics, practices, and health of online communities. There is much that the research community can do to develop tools and practices that may help community moderators guide their communities through this period of technological change. For example, new systems for effortful communication may help community members demonstrate authenticity, and new methods of communicating norms may help communities stay aligned. As a starting point, this paper documents the perspectives of moderators at this critical juncture, where, as one moderator commented, "there has to be a lot that we're missing."

ACKNOWLEDGMENTS

This material is based upon work partially supported by the National Science Foundation under Grant No. CHS 1901151/1901329.

REFERENCES

[1] Open AI. 2022. Introducing chatgpt. *OpenAI Blog*, (Nov. 2022). Retrieved 9/10/2023. https://openai.com/blog/chatgpt.

[2] Isabelle Augenstein et al. 2023. Factuality challenges in the era of large language models. (2023). arXiv: 2310.05189 [cs.CL].

[3] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: an analysis of anonymity and ephemerality in a large online community. In *Proceedings of the international AAAI conference on web and social media* number 1. Vol. 5, 50–57.

[4] 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, (Apr. 2023), 33 pages. DOI: 10.1145/3579631.

[5] Virginia Braun and Victoria Clarke. 2021. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18, 3, 328–352. eprint: https://doi.org/10.1080/14780887.2020.1769238. DOI: 10.1080/14780887.2020.1769238.

[6] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11, 4, 589–597. eprint: https://doi.org/10.1080/2159676X.2019.1628806. DOI: 10.1080/2159676X.2019.1628806.

[7] Virginia Braun and Victoria Clarke. 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, 13, 2, (Mar. 2021), 201–216. Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1704846. DOI: 10.1080/2159676X.2019.1704846.

[8] Gordon Burtch, Dokyun Lee, and Zhichen Chen. 2024. The consequences of generative AI for online knowledge communities. en. *Scientific Reports*, 14, 1, (May 2024), 10413. Publisher: Nature Publishing Group. DOI: 10.1038/s41598-024-61221-0.

[9] Kelly Caine. 2016. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). Association for Computing Machinery, San Jose, California, USA, 981–992. ISBN: 9781450333627. DOI: 10.1145/2858036.2858498.

[10] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). Association for Computing Machinery, Montreal QC, Canada, 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174240.

[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, Article 174, (Nov. 2019), 30 pages. DOI: 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. *Proc. ACM Hum.-Comput. Interact.*, 2, CSCW, Article 32, (Nov. 2018), 25 pages. DOI: 10.1145/3274301.

[13] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17). Association for Computing Machinery, Portland, Oregon, USA, 1217–1230. ISBN: 9781450343350. DOI: 10.1145/2998181.2998213.

[14] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17). Association for Computing Machinery, New York, NY, USA, (Feb. 2017), 1217–1230. ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998213.

[15] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: evaluating human evaluation of generated text. (2021). arXiv: 2107.00061 [cs.CL].

[16] John W. Creswell and J. David Creswell. 2017. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. en. Google-Books-ID: KGNADwAAQBAJ. SAGE Publications, (Dec. 2017). ISBN: 978-1-5063-8671-3.

[17] Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: a comprehensive survey of threat models and detection methods. *IEEE Access*, 11, 70977–71002. DOI: 10.1109/ACCESS.2023.3294090.

[18] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. (Oct. 2010), 115–122. DOI: 10.1109/VAST.2010.5652922.

[19] Jill P. Dimond, Casey Fiesler, Betsy DiSalvo, Jon Pelc, and Amy S. Bruckman. 2012. Qualitative data collection technologies: a comparison of instant messaging, email, and phone. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work* (GROUP '12). Association for Computing Machinery, Sanibel Island, Florida, USA, 277–280. ISBN: 9781450314862. DOI: 10.1145/2389176.2389218.

[20] Renee DiResta and Josh A. Goldstein. 2024. How spammers and scammers leverage ai-generated images on facebook for audience growth. (2024). arXiv: 2403.12838 [cs.CY].

[21] 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* (CHI '19). Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300372.

[22] Benj Edwards. 2023. OpenAI discontinues its AI writing detector due to "low rate of accuracy". en-us. (July 2023). Retrieved Sept. 12, 2023 from https://arstechnica.com/information-technology/2023/07/openai-discontinues-its-ai-writing-detector-due-to-low-rate-of-accuracy/.

[23] Hany Farid. 2022. Creating, Using, Misusing, and Detecting Deep Fakes. en. *Journal of Online Trust and Safety*, 1, 4, (Sept. 2022). Number: 4. DOI: 10.54501/jots.v1i4.56.

[24] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM*, 59, 7, (June 2016), 96–104. DOI: 10.1145/2818717.

[25] Casey Fiesler. 2024. AI chatbots are intruding into online communities where people are trying to connect with other humans. en-US. (May 2024). Retrieved June 11, 2024 from http://theconversation.com/ai-chatbots-are-intruding-into-online-communities-where-people-are-trying-to-connect-with-other-humans-229473.

[26] Casey Fiesler, Jialun Jiang, Joshua Mccann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12, 1, (June 15, 2018). DOI: 10.1609/icwsm.v12i1.15033.

[27] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. 2024. Remember the human: a systematic review of ethical considerations in reddit research. *Proc. ACM Hum.-Comput. Interact.*, 8, GROUP, Article 5, (Feb. 2024), 33 pages. DOI: 10.1145/3633070.

[28] Jon Gertner. 2023. Wikipedia's Moment of Truth. en-US. *The New York Times*, (July 2023). Retrieved July 2, 2024 from https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html.

[29] Sarah Gilbert. 2023. Towards intersectional moderation: an alternative model of moderation built on care and power. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 256, (Oct. 2023), 32 pages. DOI: 10.1145/3610047.

[30] Sarah A. Gilbert. 2020. "i run the world's largest historical outreach project and it's on a cesspool of a website." moderating a public scholarship site on reddit: a case study of r/askhistorians. *Proc. ACM Hum.-Comput. Interact.*, 4, CSCW1, Article 19, (May 2020), 27 pages. DOI: 10.1145/3392822.

[31] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: emerging threats and potential mitigations. (2023). arXiv: 2301.04246 [cs.CY].

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

[33] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 466, (Oct. 2021), 35 pages. DOI: 10.1145/3479610.

[34] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction*, 7, CSCW1, (Apr. 2023), 133:1–133:28. DOI: 10.1145/3579609.

[35] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25, 1, (Mar. 2020), 89–100. DOI: 10.1093/jcmc/zmz022.

[36] Hans W. A. Hanley and Zakir Durumeric. 2024. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. arXiv:2305.09820 [cs]. (Mar. 2024). DOI: 10.48550/arXiv.2305.09820.

[37] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, Honolulu, HI, USA, 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376383.

[38] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. Ai-mediated communication: how the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19). Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300469.

[39] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120, 11, (Mar. 2023). DOI: 10.1073/pnas.2208839120.

[40] Henry Jenkins. 2006. Introduction: "Worship at the Altar of Convergence": A New Paradigm for Understanding Media Change. In *Convergence Culture: Where Old and New Media Collide*. NYU Press, 20. ISBN: 978-0-8147-4281-5. Retrieved June 19, 2024 from http://www.jstor.org/stable/j.ctt9qffwr.4.

[41] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "did you suspect the post would be removed?": understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3, CSCW, Article 192, (Nov. 2019), 33 pages. DOI: 10.1145/3359294.

[42] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: the case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26, 5, Article 31, (July 2019), 35 pages. DOI: 10.1145/3338243.

[43] Kowe Kadoma, Danaë Metaxa, and Mor Naaman. 2024. Generative ai and perceptual harms: who's suspected of using llms? (2024). https://arxiv.org/abs/2410.00906 arXiv: 2410.00906 [cs.HC].

[44] 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 (CHI '16). Association for Computing Machinery, San Jose, California, USA, 1152–1156. ISBN: 9781450333627. DOI: 10.1145/2858036.2858356.

[45] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2011. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, 1, 125–179.

[46] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring user-moderator alignment on r/changemyview. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 286, (Oct. 2023), 36 pages. DOI: 10.1145/3610077.

[47] Robert E Kraut and Paul Resnick. 2011. Encouraging contribution to online communities. *Building successful online communities: Evidence-based social design*, 21–76.

[48] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to participate in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). Association for Computing Machinery, Atlanta, Georgia, USA, 1927–1936. ISBN: 9781605589299. DOI: 10.1145/1753326.1753616.

[49] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

[50] Lawrence Lessig. 1999. *Code and Other Laws of Cyberspace*. Basic Books, Inc., USA. ISBN: 046503912X.

[51] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. en. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, (May 2022), 584–595. DOI: 10.1609/icwsm.v16i1.19317.

[52] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. (2023). arXiv: 2304.02819 [cs.CL].

[53] Weixin Liang et al. 2024. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. arXiv:2403.07183 [cs]. (Mar. 2024). DOI: 10.48550/arXiv.2403.07183.

[54] Ingrid Lunden. 2023. LinkedIn, now at 1B users, turns on OpenAI-powered reading and writing tools. en-US. (Nov. 2023). Retrieved June 28, 2024 from https://techcrunch.com/2023/11/01/linkedin-now-at-1b-users-turns-on-openai-powered-reading-and-writing-tools/.

[55] Makyen. 2023. Temporary policy: Generative AI (e.g., ChatGPT) is banned. Forum post. (Aug. 2023). Retrieved Sept. 12, 2023 from https://meta.stackoverflow.com/q/421831/9737437.

[56] 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, 9785–9789.

[57] 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* (CHI '16). Association for Computing Machinery, San Jose, California, USA, 1138–1151. ISBN: 9781450333627. DOI: 10.1145/2858036.2858391.

[58] J. Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media + Society*, 5, 2, 2056305119836778. eprint: https://doi.org/10.1177/2056305119836778. DOI: 10.1177/2056305119836778.

[59] J. Nathan Matias and Merry Mou. 2018. Civilservant: community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). Association for Computing Machinery, Montreal QC, Canada, 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3173583.

[60] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3, CSCW, Article 72, (Nov. 2019), 23 pages. DOI: 10.1145/3359174.

[61] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. 2023. Addressing the harms of AI-generated inauthentic content. en. *Nature Machine Intelligence*, 5, 7, (July 2023), 679–680. Number: 7 Publisher: Nature Publishing Group. DOI: 10.1038/s42256-023-00690-w.

[62] Jaron Mink, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, and Gang Wang. 2024. It's trying too hard to look real: deepfake moderation mistakes and identity-based bias. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24) Article 778. Association for Computing Machinery, New York, NY, USA, 20 pages. DOI: 10.1145/3613904.3641999.

[63] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119, 8, (Feb. 2022), e2120481119. Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.2120481119.

[64] Rajvardhan Oak and Zubair Shafiq. 2024. Understanding underground incentivized review services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24) Article 950. Association for Computing Machinery, New York, NY, USA, 18 pages. DOI: 10.1145/3613904.3642342.

[65] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23, 1, (Jan. 2020), 128–147. DOI: 10.1080/1369118X.2018.1486870.

[66] OpenAI. 2023. Gpt-4 technical report. (2023). arXiv: 2303.08774 [cs.CL].

[67] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. en. *Patterns*, 5, 5, (May 2024), 100988. DOI: 10.1016/j.patter.2024.100988.

[68] Jenny Preece. 1999. Empathic communities: balancing emotional and factual communication. *Interacting with Computers*, 12, 1, (Sept. 1999), 63–77. DOI: 10.1016/S0953-5438(98)00056-3.

[69] Gale H. Prinster, C. Estelle Smith, Chenhao Tan, and Brian C. Keegan. 2024. Community archetypes: an empirical framework for guiding research methodologies to reflect user experiences of sense of virtual community on reddit. *Proc. ACM Hum.-Comput. Interact.*, 8, CSCW1, Article 33, (Apr. 2024), 33 pages. DOI: 10.1145/3637310.

[70] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7, 2, 20563051211019004. eprint: https://doi.org/10.1177/20563051211019004. DOI: 10.1177/20563051211019004.

[71] Irene Rae. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 2024), 1–14. DOI: 10.1145/3613904.3642076.

[72] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. (2021). arXiv: 2102.12092 [cs.CV].

[73] Yuqing Ren, Robert Kraut, Sara Kiesler, and Paul Resnick. 2012. Encouraging commitment in online communities. *Building successful online communities: Evidence-based social design*, 77–124. Publisher: MIT Press Cambridge. Retrieved June 20, 2024 from https://books.google.com/books?hl=en&lr=lang_en&id=lIvBMYVxWJYC&oi=fnd&pg=PA77&dq=encouraging+commitment+to+online+communities&ots=z-A1dlk7HB&sig=M6PKeXnUlho3YIVLarHZNHOcckY.

[74] Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in internet transactions: empirical analysis of ebay's reputation system. In *The Economics of the Internet and E-commerce*. Emerald Group Publishing Limited, 127–157.

[75] Regina Rini and Leah Cohen. 2022. Deepfakes, Deep Harms. eng. *Journal of Ethics and Social Philosophy*, 22, 2, 143–161. Retrieved May 16, 2024 from https://heinonline.org/HOL/P?h=hein.journals/jetshy22&i=150.

[76] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? (2023). arXiv: 2303.11156 [cs.CL].

[77] Aida Sanatizadeh, Yingda Lu, Keran Zhao, and Yuheng Hu. 2023. Exploring the Effect of Large Language Models on Knowledge Seeking and Contribution in Online Knowledge Exchange Platforms. en. SSRN Scholarly Paper. Rochester, NY, (May 2023). DOI: 10.2139/ssrn.4459729.

[78] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive moderation of online discussions: existing practices and the potential for algorithmic support. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, Article 370, (Nov. 2022), 27 pages. DOI: 10.1145/3555095.

[79] Joseph Seering. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4, CSCW2, Article 107, (Nov. 2020), 28 pages. DOI: 10.1145/3415178.

[80] 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, (Jan. 11, 2019), 1417–1443. Retrieved Sept. 9, 2023 from http://dx.doi.org/10.1177/1461444818821316.

[81] C. Estelle Smith, Irfanul Alam, Chenhao Tan, Brian C. Keegan, and Anita L. Blanchard. 2022. The impact of governance bots on sense of virtual community: development and validation of the gov-bots scale. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, Article 462, (Nov. 2022), 30 pages. DOI: 10.1145/3555563.

[82] C. Estelle Smith, Hannah Miller Hillberg, and Zachary Levonian. 2023. "thoughts & prayers" or ":heart reaction: & :prayer reaction:": how the release of new reactions on caringbridge reshapes supportive communication in health crises. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 244, (Oct. 2023), 39 pages. DOI: 10.1145/3610035.

[83] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2024. (in)visible moderation: a digital ethnography of marginalized users and content moderation on twitch and reddit. *New Media & Society*, 26, 7, 4034–4055. eprint: https://doi.org/10.1177/14614448221109804. DOI: 10.1177/14614448221109804.

[84] Stuart A. Thompson. 2023. A.i.-generated content discovered on news sites, content farms and product reviews. *The New York Times*, (May 2023). Retrieved 9/10/2023. https://www.nytimes.com/2023/05/19/technology/ai-generated-content-discovered-on-news-sites-content-farms-and-product-reviews.html.

[85] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. arXiv:2310.15683 [cs]. (Oct. 2023). DOI: 10.48550/arXiv.2310.15683.

[86] Xiaoxiao (Shawn) Wang and Jinyang Zheng. 2024. Can Banning AI-generated Content Save User-Generated Q&A Platforms? en. SSRN Scholarly Paper. Rochester, NY, (Jan. 2024). DOI: 10.2139/ssrn.4750326.

[87] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. (2023). arXiv: 2306.15666 [cs.CL].

[88] Yiluo Wei and Gareth Tyson. 2024. Understanding the impact of ai generated content on social media: the pixiv case. (2024). https://arxiv.org/abs/2402.18463 arXiv: 2402.18463 [cs.CY].

[89] Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. en. *Proceedings of the International AAAI Conference on Web and Social Media*, 16, (May 2022), 1121–1132. DOI: 10.1609/icwsm.v16i1.19363.

[90] 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* (CHI '19). Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300390.

[91] Junzhi Xue, Lizheng Wang, Jinyang Zheng, Yongjun Li, and Yong Tan. 2023. Can ChatGPT Kill User-Generated Q&A Platforms? en. SSRN Scholarly Paper. Rochester, NY, (May 2023). DOI: 10.2139/ssrn.4448938.

[92] Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an ai-powered malicious social botnet. (2023). arXiv: 2307.16336 [cs.CY].

[93] Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A Survey on Detection of LLMs-Generated Content. arXiv:2310.15654 [cs]. (Oct. 2023). DOI: 10.48550/arXiv.2310.15654.

[94] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. Policykit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (UIST '20). Association for Computing Machinery, Virtual Event, USA, 365–378. ISBN: 9781450375146. DOI: 10.1145/3379337.3415858.

[95] Lei Zhang, Tianying Chen, Olivia Seow, Tim Chong, Sven Kratz, Yu Jiang Tham, Andrés Monroy-Hernández, Rajan Vaish, and Fannie Liu. 2022. Auggie: encouraging effortful communication through handcrafted digital experiences. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, Article 427, (Nov. 2022), 25 pages. DOI: 10.1145/3555152.