Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality

Kelsea Schulenberg
Clemson University
South Carolina, USA
kelseas@g.clemson.edu

Lingyuan Li
Clemson University
South Carolina, USA
lingyu2@g.clemson.edu

Guo Freeman
Clemson University
Clemson, South Carolina, USA
guof@clemson.edu

Samaneh Zamanifard
Clemson University
South Carolina, USA
szamani@g.clemson.edu

Nathan J. McNeese
Clemson University
South Carolina, USA
mcneese@clemson.edu

ABSTRACT

Extensive HCI research has investigated how to prevent and mitigate harassment in virtual spaces, particularly by leveraging human-based and Artificial Intelligence (AI)-based moderation. However, social Virtual Reality (VR) constitutes a novel social space that faces both intensified harassment challenges and a lack of consensus on how moderation should be approached to address such harassment. Drawing on 39 interviews with social VR users with diverse backgrounds, we investigate the perceived opportunities and limitations for leveraging AI-based moderation to address emergent harassment in social VR, and how future AI moderators can be designed to enhance such opportunities and address limitations. We provide the first empirical investigation into re-envisioning AI's new roles in innovating content moderation approaches to better combat harassment in social VR. We also highlight important principles for designing future AI-based moderation incorporating user-human-AI collaboration to achieve safer and more nuanced online spaces.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Empirical studies in collaborative and social computing**;

KEYWORDS

artificial intelligence, content moderation, online harassment, social VR

ACM Reference Format:
Kelsea Schulenberg, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, and Nathan J. McNeese. 2023. Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581090

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581090

1 INTRODUCTION

The opportunity to remain anonymous and pseudonymous in various online social spaces has increased the possibility of spreading harmful and offensive content [20] and has led to a wide variety of misbehaving, including trolling, bullying, and online harassment [44]. In response, content moderation mechanisms have become crucial approaches to mitigate and prevent online harassment on social media [14, 30, 61, 84, 85], text-based online forums (e.g., Reddit) [15, 21, 33, 34], and live streaming platforms (e.g., Twitch) [8–11, 69, 98], including human-based moderation [9–11, 68, 98], community-driven moderation [27, 32, 67, 72], and a growing new trend of Artificial Intelligence (AI)-based moderation [31, 59, 91, 92].

However, social Virtual Reality (VR) platforms (e.g., Meta's Horizon Worlds, VRChat, and AltspaceVR), where multiple users can interact with one another through VR head-mounted displays in 3D virtual spaces [25, 51], seem to lead to more severe forms of harassment (e.g., embodied physicalized sexual assault [74, 75]) and challenges to mitigate and prevent such harassment. Social VR's unique incorporation of embodiment and body tracking, sense of presence within an all-encompassing space, and synchronous voice communication may afford harassers the opportunity to "grope", "touch", and verbally harass others in a way that can be felt as more severe than in traditional online environments [4–6, 26, 71]. Yet, it is still unclear how, if at all, traditional harassment mitigation methods that have been widely used in other online contexts, such as content moderation, can be leveraged to effectively combat these intensified forms of harassment and achieve safer social VR spaces.

In fact, current moderation practices on major social VR platforms often show arguably ambivalent success [55, 56, 58, 86–89, 94, 95]; prior social VR research also reveals somewhat contradicting findings about how traditional human/community-based moderation would be perceived and accepted by social VR users [5, 26]; and little to no research specifically explores the potential of how, if at all, new moderation mechanisms, especially the growing new trend of AI-based moderation, can be designed and used to manage emergent harassment in social VR. Therefore, as social VR becomes increasingly prevalent within the public sphere, it is imperative to understand and empirically investigate how best to mitigate harassment burdens via new moderation methods that

take into account the uniqueness of social VR, as extant methods are likely not enough to prevent these intensified forms of harassment.

In this paper, we thus focus on the ways in which the most recent technological advances in content moderation (i.e., AI-based moderation) for managing online harassment are currently perceived within social VR communities, and how these communities envision the design and use of future AI-based moderation to combat harassment in these spaces. By conducting 39 interviews with social VR users who have diverse backgrounds and perspectives, we investigate the following research questions:

RQ1: What are the perceived opportunities and limitations for AI-based moderation to address emergent harassment in social VR, especially in comparison to traditional human-based moderation?

RQ2: How can we design future AI moderators to enhance such opportunities and address limitations to better prevent emergent harassment in social VR?

We contribute to existing HCI research on content moderation and social VR in three ways. First, we offer the first empirical investigation into how social VR users view AI-based moderation as having unique advantages and limitations for mitigating new forms of harassment, and how they envision ways in which the AI-based moderation system itself, especially taken in combination with human-based and/or community-based moderation, should be designed to provide a sense of comfort and safety depending on individual needs. It is important to note that AI-based moderation has yet to be implemented in social VR. Therefore, our study is pioneering in its proactive approach to envisioning the future of moderation systems incorporating human moderators, AI, and actual users to better protect people from intensified harassment in novel online social spaces such as social VR. Second, using social VR as a unique online context, we expand the rapidly evolving body of literature on content moderation and AI by pointing towards AI's new and envisioned roles for innovating traditional moderation mechanisms. However, the potential risks of AI also playing a role in creating new and possibly unfair power dynamics in social VR must be addressed. Grounded in these insights, lastly, we propose three vital principles aimed at informing the designing of future AI-based moderation incorporating *user-human-AI collaboration* to achieve safer and more inclusive online social spaces and interaction dynamics.

2 RELATED WORK

2.1 Content Moderation for Managing Online Harassment

Online harassment can lead to serious and negative effects on the target individuals' well-being. Therefore, a large body of HCI research has investigated various strategies, mechanisms, and technical features to protect users from online harassment in diverse online contexts such as social media [14, 30, 61, 84, 85], text-based online forums (e.g., Reddit) [15, 21, 33, 34] and live streaming platforms (e.g., Twitch) [8–11, 69, 98], including the extensive exploration of content moderation as an effective mechanism. Content moderation can be broadly defined as *"the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse"* [29], and can often be characterized as a series of trade-offs between actions, styles, philosophies, and values based on the context and facilitators of moderation [38]. In particular, prior research has highlighted two main approaches of content moderation for managing online harassment: (1) *human-based moderation*, including human moderators and community-driven moderation; and (2) the incorporation of AI into moderation practices, i.e. *AI-based moderation*.

Human-Based Moderation for Mitigating Online Harassment. Human-based content moderation has been widely considered crucial for preventing and mitigating online harassment by providing a deep understanding of the specific context of the harassing behavior, effectively removing inappropriate and toxic content, and banning harassers from attacking more people in nuanced ways [10, 17, 65, 70, 76, 98]. Through this approach, professionally hired or contracted (i.e., centralized corporate [67]), or voluntary (i.e., user-driven [67]) human moderators go through posts and comments to remove abusive or harassing content manually, which often includes removing and/or disciplining offenders [8, 33, 36, 40, 65]. More recently, human moderators have had to undertake significantly more complicated efforts to moderate interactions happening in real-time and are performative and social in nature rather than simply in text-based, asynchronous online spaces, e.g., voluntary human moderators managing a live streaming channel in real-time [9–11, 37, 99].

There are, however, fundamental issues that hamper the long-term viability and sustainability of human-based moderation. First, intrinsic characteristics of human moderators - including their demographic and social identity, personality, and belief systems - undeniably shape their views on moderation [68]. Just as moderation at large is often criticized for perpetuating harmful social biases (e.g., Twitch moderation policies disproportionately targeting and sexualizing women streamers [100]), the biases of individual human moderators often creep into their policies and actions, which in turn can affect trust in human moderators. Second, the levels of trust and transparency attributed to human moderators largely depend on factors that are out of their direct control, such as specific platform communication features [39, 81]. Third, this moderation model requires significant emotional and mental labor from human moderators, rendering their efforts not scalable or sustainable when they have to monitor thousands of comments or real-time messages over extended hours [17, 24, 62, 97, 98]. In this sense, moderation becomes a notoriously laborious and emotionally draining, even traumatizing endeavor for many human moderators [21, 77, 98], especially when they belong to marginalized communities (e.g., Asian American and Pacific Islander moderators on Reddit) [21].

As a result, many online platforms leverage at least some type of community-driven moderation features in hopes of mitigating these issues (e.g., Reddit users' community efforts to "flag" offensive or harassing content [18, 43, 49]). This community-driven moderation approach has often been shown to be effective in promoting more civil political discourse [27] and to weed out toxicity in online communities by pushing toxic members out [32], which allows communities to shape their guiding principles and experiences [67]. However, community-driven moderation still seems to fall short of addressing the issue of scalability, such that smaller communities are more able to moderate themselves than large communities can [72]. Therefore, as a way to innovate traditional human- or

community-based content moderation, automated or algorithmic content moderation (otherwise known as AI-based moderation) has become a growing new trend to prevent and address online harassment.

AI-Based Moderation as a Growing New Trend to Mitigate Online Harassment. The definition of "AI-based moderation" in HCI research tends to be broad [28, 29]. At a high level, AI-based moderation is characterized by the use of machine learning and decision making to monitor online spaces for violations and incidents of harassment [31, 35].

Currently, two main methods for using AI-based moderation to mitigate online harassment exist. The first method focuses on automatically filtering certain keywords to block posts or comments that include specific harassing terms and phrases, such as the AutoModerator bot on Reddit [7, 19, 20, 35, 45, 63] and flagging systems used in gaming [42, 78]. The second method mainly leverages Natural Language Processing techniques to automatically detect cyberbullying content, such as parsing the language patterns used by bullies and their victims to automatically delete posts and comments and ban the user from future activities [3, 64]. This method can also be used to detect harassment in real-time voice communication online, where machine learning-trained AI can conduct voice analysis to detect sexual harassment online by searching for clues of fear, anger, and disgust emotions in women's voices [37, 66, 78]. In both methods, AI acts in some ways to enhance the capabilities of human moderators by leveraging its computational abilities to perform menial moderation tasks, thereby preserving human moderators' time and energy to focus on more complex decision-making and social practices [7, 31, 52].

However, some other research has also shown that existing AI moderation tools in some online spaces may lack the ability to differentiate genuine harassing behaviors from non-harassing behaviors (e.g., the censoring of the "mock impoliteness" utilized by LGBTQ+ communities to cope with hostility) [82]. Some platforms even perpetuate a lack of transparency (i.e., clear communication) regarding how AI moderation works and the reasons behind its decision-making, making users feel a decreased sense of agency over their online experiences [13, 13, 22, 22, 28, 46, 83]. Additionally, some current iterations of AI-based moderation have come under fire for disproportionately targeting marginalized individuals (e.g., women, people with mental illness, and Black individuals) in its moderation and punishment practices [2, 23, 30].

Despite limitations, AI-based moderation's promise to be a computationally powerful and expansive facet of online content moderation ultimately still makes it a valuable technological advancement for combating online harassment. Importantly, it is the responsibility of HCI researchers and practitioners alike to investigate how to mitigate existing issues in AI moderation while simultaneously elevating its inherent usefulness, especially as concerns over how best to address harassment and safety in new and embodied social spaces such as social VR are becoming more widely and critically discussed in both popular media [74, 75] and HCI research [4–6, 26].

2.2 Challenges of Mitigating Harassment in Social VR

Social VR platforms (e.g., VRchat, Rec Room, Bigscreen, AltspaceVR, and Meta Horizon Worlds) have increasingly grown in popularity over the recent years, as they provide new online social spaces where people can meet, interact, and socialize in more embodied (i.e., experiencing a virtual body representation as our own body within a virtual environment [73]) and immersive ways compared to traditional online contexts such as social media and gaming. As such, social VR users can enjoy offline-like social activities (e.g., walking in public spaces, playing a game, watching a movie, participating in a concert, and having a party) in a highly realistic and immersive simulated 3D virtual environment in a way that is similar to offline face-to-face communication through the predominant use of real-time voice chat, partial or full-body tracked avatars, and more customized avatar design.

A growing concern, however, is that social VR may also lead to intensified and more severe forms of harassment compared to other online contexts. These incidents have been frequently reported in mass media, such as the virtual "groping" behaviors [75] and the most recent "rape" in the metaverse [74]. Therefore, there is an emerging research agenda in HCI and CSCW that focuses on understanding and mitigating new forms of harassment in novel social VR spaces [5, 6, 26, 71]. This body of work has warned that social VR's focus on embodiment, sense of presence, body tracking, and synchronous voice conversation may allow people to verbally assault and virtually "touch" (e.g., grabbing and groping) others without their permission [5, 6, 26], the latter of which seems to simulate types of physical harassment and assault that often happen in the offline world [5, 26]. As a result, it may be felt as more realistic and disruptive compared to harassment in traditional online gaming and virtual worlds [26].

Overall, prior work points to three main challenges for preventing and mitigating emergent harassment in social VR. **Challenge 1**: an apparent lack of consensus amongst social VR users on what social norms/behaviors are harassing rather than simply inappropriate or "fun/play" creates barriers to effectively define and identify harassment, as a diverse array of individuals and communities may have different understandings [5, 26]. **Challenge 2**: although existing social VR platforms equip users with various harassment prevention tools (see 2.3), social VR users have pointed out their various limitations [5, 26]. For example, it is difficult to document harassment in social VR for reporting because incidents often happen within real-time synchronous interactions, which can be ephemeral and not recorded or archived (e.g., as verbal attacks or physical touch) [26]. **Challenge 3**: while human-based moderation (e.g., formal moderators or dedicated community members as volunteers) has been proposed as a potential solution to help prevent and mitigate harassment [5], many social VR users are concerned that a human moderator's subjective bias might affect their abilities to moderate spaces equitably [26].

Taken in sum, these challenges point towards an urgent need for research on more nuanced methods to address social VR harassment. Yet, it is unclear how, if at all, AI-based moderation - a relatively successful method for mitigating harassment in other online social spaces - can be leveraged towards this aim. Still, in recognition of all

of these challenges, major social VR platforms, including VRChat, AltspaceVR, and Meta Horizon Worlds, have instituted various practices and features for moderation and harassment prevention with arguably ambivalent success, which we detail in the next section.

2.3 Existing Moderation Efforts in Social VR

Major social VR platforms such as VRChat, AltspaceVR, and Meta Horizon Worlds, which are also most used by our participants (see Methods), have made various efforts to moderate their virtual spaces and mitigate harassment with ambivalent success. These efforts include community guidelines, penalty enforcement policies, and moderation pipelines as stated on these platforms' official websites.

Community Guidelines & Punishments in Social VR. How social VR platforms define their community guidelines and punishments directly determines what types of behaviors and content creation are considered by these platforms to be inappropriate/harassing and what moderation actions will be taken in response to violations. On their community guideline pages, all three major social VR platforms (i.e., VRChat, AltspaceVR, and Meta Horizon Worlds) list certain behaviors and content types as violations requiring moderation, including: defamation/intolerance/hate speech; discovery and disclosure of personal information, or doxxing; violating other users' personal space repeatedly; creating worlds and events that promote and/or display overt violence and hate; creating worlds and events that are either sexually suggestive without an 18+ restriction or sexually explicit (e.g., displaying pornography) regardless of age restriction; and impersonating a VRChat/AltspaceVR/Meta employee [54, 57, 86, 89, 95]. All three platforms also mention punishments for violations, including account suspension and banning/termination, although Meta Horizon Worlds and AltspaceVR have an additional punishment tier of a "warning" to violators to cease their behavior [54, 89].

Existing Moderation Pipelines in Social VR. For the context of this paper, we define a *moderation pipeline in social VR* as the process through which a behavior or content creation in social VR first becomes flagged or noticed as inappropriate/harassing based on the above-mentioned community guidelines, then the flag or notice is reviewed by a party, and finally, punishment is executed by a party. As of November 2022, all three major social VR platformsonly utilize the human-based moderation model within their pipelines, and often make it unclear when and in what situations these company-employed moderation teams intervene [57, 86, 95]. In this sense, the vast majority of the moderation burden seems to be placed upon individual moderators and user community leaders/hosts to manage inappropriate or harassing behavior while the companies running the platforms play a rather ambiguous and, in some cases, unresponsive role in the process [55, 56, 58, 86–89, 94, 95].

For example, VRChat's community guidelines [86] and terms of service [88] explain that individual users are responsible for reporting behavior and/or worlds deemed inappropriate or harassing via an online form on VRChat's website [87], after which *"VRChat Moderation will take action based on its discretion in gray areas"* [86]. However, crucial information is obscured, as what the platform considers to be a "gray area" is unclear. AltspaceVR's moderation pipeline focuses their platform's content exclusively on user-created worlds, parties, and events [1]. It thus requires the hosts of said user-created spaces to be responsible for moderating their own spaces through various platform-provided tools, including: kicking users out of events and spaces; delivering warnings to users; and assigning other users as moderators, amongst others [94–96]. Likewise, Meta Horizon Worlds' moderation pipeline mainly emphasizes individual users' and world creators' responsibility for managing their own experiences [54]. One example is the "Poll to Remove" feature, which allows users to anonymously start a poll within a group to vote on whether a group member should be removed for being disruptive [57]. A majority "yes" vote will automatically take the offending person out of the world and transport them to their personal space, with everyone in the world notified of the action and resolution [57]. Despite these existing moderation efforts in social VR, social VR users are often made to feel that they are on their own and cannot rely on the platforms to keep them safe, a message that is reinforced by these platforms' heavy emphasis on individual actions and responsibility for moderation [54, 57, 86, 89, 94–96]. This is especially burdensome for platforms whose content is entirely user-driven [47].

Given the often inadequate and problematic nature of social VR platforms' existing moderation practices as detailed above, it becomes necessary to explore how other moderation approaches that are currently not being utilized by social VR platforms can serve to better protect users from harassment without creating unequal power dynamics and burdens between users, community leaders, and the platform itself. Therefore, in this paper we especially focus on how AI-based moderation may be perceived as a more nuanced approach, along with both opportunities and risks, to address unique harassment in social VR compared to the traditional human-based moderation approaches that have been utilized in existing social VR spaces and other online environments (**RQ1**). We also aim to explicate social VR users' own opinions and recommendations to enhance these opportunities and remedy risks to inform the future design of AI moderators to effectively prevent emergent harassment in social VR (**RQ2**).

3 METHODS

Recruitment and Participants. The University's Institutional Review Board (IRB) approved this study for research ethics prior to the recruitment of participants. We posted recruitment messages on various popular online forums for social VR users (e.g., r/SocialVR, r/VRchat, r/OculusQuest, r/Recroom, and r/gaymers in Reddit) and social media platforms (e.g., Facebook and Twitter) to recruit participants who engage in various social VR platforms. We then interviewed all individuals who responded to our recruitment message and were willing to participate in March and April of 2022 (N=39). We acknowledge that our recruitment methods may have led to potential self-selection bias, e.g., only social VR users who are also active social media users may have responded. However, the individuals recruited through these methods provide unique insights on the moderation needs of social VR users, which are much needed for HCI and social VR research. For example, although the vast majority of participants are currently living in the U.S. (N=32), these views are still valuable as they represent a large user base of social VR. The remaining participants are located in Germany (N=2),

France (N=1), Canada (N=1), and Guatemala (N=1), with two N/A responses. Our participants contain a nearly even split between Cis Women (N=17) and Cis Men (N=16), with the remainder of the participants divided between Trans Woman (N=1), Trans Man (N=1), Trans Unspecified (N=1), Non-Binary (N=1), Genderqueer Feminine Presenting (N=1), and Genderqueer (N=1). Around half of our participants are Black (N=18) and 12 identify as White. Participants also identify as Biracial (N=8), Hispanic (2), Middle Eastern (N=1), and Asian (N=1). A wide variety of sexual identities are represented, including Straight (N=14), Lesbian (N=8), Gay (N=4), Asexual (N=4), Bisexual (N=4), Queer (N=2), No Answer (N=2), Pansexual (N=1), and either gay or bisexual/pansexual (N=2).

The average age of our participants is 25.62 (excluding 2 No Answer responses), with a range of 18 - 44 years old. The majority of our participants are users of VR Chat (N=23), AltspaceVR (N=12), Meta Horizon (N=5), and Rec Room (N=4). Additionally, each of the following platforms were represented with less than three participants engaging in them: Spatial, Decentraland, Immersed, Bigscreen, Mozilla Hub, and Spatio VR. On average, our participants have been engaging in social VR for 2 years and 3 months, with variations from 2-3 times in total to 6 years.

It is also important to note that, although some of our participants have occupations or are in schoolwork related to a technology sector (N=5) (e.g., software developer, P3; student in computer science, P7; Blender model creator, P9; student in 3D art and animation/VFX, P11; and IT coordinator, P27), most participants do not have any specific experience building or developing AI technology. Given this, our participants' perceptions are primarily based upon their experiences with AI-based moderation in other online contexts (e.g., social media, Discord, and gaming) rather than technical building experience. Indeed, most of our participants were recruited from sites that either use AI-based moderation in some form (e.g., Reddit and their AutoModerator bot system) or human-based moderation (e.g., Facebook), so it is reasonable to assume that participants have sufficient experience to speak upon their perceptions of both AI- and human-based moderation in nuanced and informed ways. As such, the design recommendations that our participants put forth are not focused on technical elements of building an AI-based moderation system in social VR, but rather reflect how their personal perceptions of and experiences in social VR combine with prior moderation experiences. As social VR continues to attract diverse users, it is indeed expected that most users will not be AI experts. Therefore, our sample represents how *actual* social VR users perceive, envision, and approach the future of moderation systems to prevent emergent harassment in social VR.

Interviews. We conducted 39 semi-structured in-depth interviews via text/voice chat over Discord or Zoom per the participants' personal preference as one-on-one sessions to protect their identity and privacy. Prior to the interviews, we provided an informed consent document to all participants based on their communication preferences (e.g., email or Discord message). We did not collect names or identifiable information from participants. Interview questions were crafted using dialogic techniques designed to encourage participants to engage deeply with their responses [93]. These questions as detailed further below drew inspiration from prior literature on social VR and harassment in social VR, particularly from the works of Blackwell et al. [4–6] and Freeman et al. [26], as well as from our own prior experiences with social VR as both researchers and users. Interviews first began with introductions, basic demographic questions, and questions in regard to their level of experience in social VR as well as experiences with harassment in social VR to orient the conversation towards harassment moderation. Participants were then asked to describe any new strategies for mitigating harassment that they might find to be beneficial, focusing particularly on human-based moderation (e.g., *"How would you feel about having more moderators in public spaces? What are the benefits and drawbacks?"*). Next, and most relevant to this study, interview questions turned towards the potential for AI moderation in social VR. We first provided participants with a brief explanation of AI:

"In short, we can define AI as 'the ability of a machine or a computer program to think and learn. The concept of AI is based on the idea of building machines capable of thinking, acting, and learning like humans.' Some very common examples of AI would be Siri or Google Assistant, a computer-controlled opponent in games such as a NPC or a 'boss,' or an enemy in League of Legends."

Participants were then asked to generally describe how they feel about the idea of using AI to prevent social VR harassment and to reflect on AI-based moderation in various dimensions, including invasivness (*"What about having an AI be a moderator? Would that be more or less invasive than a human moderator?"*), trust (*"Would you trust an AI more than another human in social VR to moderate the environment and stop harassment? And why?"*), fairness (*"Do you think an AI would be more fair compared to a human moderator when handling harassment? And Why?"*), effectiveness (*"Do you think an AI would be more effective compared to a human moderator when handling harassment or not? Why?"*), and empathy (*"Do you think an AI would be more empathetic compared to a human moderator when handling harassment or not? Why?"*). Finally, participants were asked to describe in detail how they would design an AI moderation system in social VR to effectively prevent emergent harassment in social VR. Interviews lasted 102 minutes on average and participants received a $50 Amazon digital gift card after they completed the interviews.

Data Analysis. After interviews were complete, recordings were first transcribed and organized within spreadsheets for clarity during data analysis. We then used empirical, in-depth qualitative analysis to analyze the data [16, 79]. A qualitative approach is appropriate for this study because qualitative methodologies are well-suited for investigating questions about "how people interpret their experiences, how they construct their worlds, and what meaning they attribute to their experiences" [53]. As outlined by McDonald et al.'s [50] guidelines for qualitative analysis in CSCW and HCI practice, data analysis procedures were not focused on obtaining inter-rater reliability between coders, but instead aimed to uncover categories of interest and to find relationships amongst categories to unveil connections and formulate them into groups of greater complexity and breadth.

First, all authors carefully read through the collected data line by line to obtain a holistic sense of participants' perceptions, expectations, and recommendations for leveraging AI-based moderation to prevent harassment in social VR. Second, the first two authors independently conducted open coding [16] of each transcript, categorized participants' responses into thematic topics related to

our research questions, and developed sub-themes emerging in participants' descriptions of their perceptions for further analysis. Third, all authors discussed and refined themes and sub-themes in a collaborative and iterative axial coding process [16] to streamline participants' perceptions of AI-based moderation in social VR and group these themes and subthemes by each research question. Then, the same two authors involved in step two extracted quotes based on themes and sub-themes refined in the third step through focused coding [16]. Lastly, all authors further discussed and refined themes and sub-themes and used the quotes to generate a rich description synthesizing answers to the research questions.

4 FINDINGS

In this section, we first explain how an interwoven blend of novel opportunities and urgent challenges arise when envisioning AI-based moderation for dealing with harassment in social VR, especially in comparison to the traditional human-based moderation approach (**RQ1**). Drawing on social VR users' own suggestions and recommendations, we then identify three potential design directions to enhance opportunities and remedy new challenges associated with this new moderation approach in social VR (**RQ2**).

4.1 Perceived Opportunities and Limitations of AI-Based Moderation to Manage Harassment in Social VR vs. Human-Based Moderation

When reflecting upon the possibility of leveraging AI-based moderation to manage harassment, an increasingly severe and urgent issue in social VR, the majority of our participants enthusiastically welcomed the novelty of this idea. However, they specifically highlight three ways in which AI-based moderation could simultaneously provide opportunities and limitations to this management.

4.1.1 AI-based moderation helps make consistent judgements regarding harassment in social VR but can show interpretation limitations if designed without proper consideration. Prior research has highlighted how social VR's unique technical features have led to various new forms of online harassment, ranging from violations of physical and personal space to physical touch without consent [5, 6, 26]. These varied forms of harassment thus makes defining and identifying harassing behaviors in social VR challenging. Many participants thus share a common concern that human-based moderation (i.e., social VR platforms' current strategy [57, 86, 89]) can perpetuate inconsistencies in moderation as a result of variations between individual human moderators' personal definitions of and experiences with harassment in social VR. In contrast, AI-based moderation, a computer program that will "*work on how you've programmed it to work*" (P1, 25, Cis Woman, Black, Lesbian, U.S.), is viewed as inherently more consistent given that each instance of this program would repeatedly follow a set of pre-defined codes and algorithms, rather than relying on individual and varied human judgements in the moment.

As P30 (30, Cis Man, Mixed Race, Genderqueer, U.S.) further explains, "*A human being is unpredictable. A person can assess and define harassment in one way or the other, or he or she can say, probably this word it is not regulatory, or it's not appropriate. But for AI, those words are already registered and they're constant. So, you expect the same results at every given instance.*" For P30, a human's unpredictability as a moderator stems from a recognition that humans can and do *"assess and define harassment"* differently, making it difficult to predict when moderation action will be taken. AI's programmatic nature, however, engenders an expectation of consistency and predictability across *"every given instance,"*. In this sense, so long as an AI-based moderation system is designed to encompass a wide variety of harassment incidents, the execution of moderation practices becomes routine and predictable. As such, AI-based moderation also has comparatively greater potential to set up platform-wide standards for detecting and handling emergent harassment in social VR than human-based moderation.

Interpretation Limitation: Diminished interpretation of sociocultural context. While participants find AI's greater consistency a comparative advantage in most harassment cases, they also see the potential risk of adopting a "one-size-fits-all" AI-based moderation approach, as some harassment incidents might require a high degree of sociocultural contextual understanding to interpret and adjudicate with nuance, something that humans have a comparative advantage on.

Indeed, P4 (24, Cis Woman, Black, Lesbian, U.S.) and P14 (20, Cis Woman, Biracial Black and Italian, Asexual, U.S.) both describe how AI-based moderation may fall short in identifying harassment in social VR if it is based on a de-contextualized, limited pre-programmed harassment judgment criteria (e.g., language, text, and pictures). For example, an AI trained without sociocutlural context might misinterpret joking between friends as harassment (*"At times two friends may meet on the social VR and may use the terms in which they love using and the AI could simply block them."* - P4), or might fail to understand the differences between lewd nudity and nudity displayed in a virtual art museum (*"some art has naked women in it. Would they (AI moderators) flag that?!"*, P14). For both P4 and P14, managing harassment is not simply a matter of applying a formula, but rather is interpersonally and socially constructed. In this sense, the types of AI-based moderation they have encountered in other online contexts such as Discord may not be able to tease out the subtle discrepancies between two objects (e.g., nudity without consent for harassment purposes vs. artistic nudity) or the nuances of different situations (verbally attacking others vs. joking between friends). These issues can arguably, in turn, lead to arbitrary, unreliable decision-making and judgements to identify and address harassment in social VR if not properly managed.

P32 (30, Queer, Hispanic, Bisexual, U.S.) further points out that AI moderation from what she understands in other online contexts lacks the cognitive ability and judgment to differentiate unintentional from malicious harassment, *"I think if someone in any way harass another person, they may not mean to do so. If the moderators are human, the moderators can be able to probably judge, and warn the person, give the person a second chance based on the situations that surround it. But an automated machine, AI would just give out the punishment. I think that kind of beats the humanity in us."* In this sense, based on P32's prior experiences, AI's focus on setting up a consistent standard to detect and address harassment can miss the very fact that, as P9 (20, Cis Man, White, Bisexual, Germany) states, *"harassment is situational in social VR and AI could have a hard time making the right decisions"*.

4.1.2 AI-based moderation effectively manages social VR harassment in real time and at a large scale but still shows some technical limitations to address new forms of harassment. Given the large-scale, multi-world/event, immersive nature of social VR, our participants often view AI-based moderation and its exponentially superior and expansive computational abilities as having a comparative advantage over human-based moderation in the context of social VR. Indeed, according to P19 (25, Cis Woman, White/Russian, Pansexual, Germany), *"a human is a finite resource and they can only do so much,"* and their ability to act in-the-moment on a large scale is intrinsically limited by their inability to be monitoring multiple spaces at once in real time. AI, on the other hand, *"is not limited to the brain of a human so it's going to be faster to collect and recognize information and data and act right away"* (P20, 24, Cis Woman, Black, Bisexual, U.S.). For participants like P19 and P20, AI-based moderation is *"not limited"* as humans are, and thus has the comparative advantage over humans to rapidly detect and act upon harassment incidents in multiple spaces simultaneously.

Above all, AI is in many ways an infinitely employable resource while human labor by nature is a rather limited resource due to factors such as emotional burn out and comparatively reduced computational abilities. P33 (25, Cis Man, Black, Straight, U.S.) describes the unrealistic expectations placed upon humans to monitor and moderate harassment in social VR around the clock, which can be comparatively feasible for AI to achieve, "*it (AI) would be there all the time, humans go to talk, they go on break*." This means that human moderators more easily and readily reach their limits for how long/much they can moderate within a multi-user online environment compared to AI-based moderation systems, as AI does not emotionally burn out and has far more computational capacity to handle more moderation situations at once.

As a result, many participants expect that the presence of AI-moderation alone would effectively deter potential harassers from taking actions to harass others, as these harassers would understand that, unlike human moderators who may be on and off, AI-based moderation is able to always "watch" at a large scale and then take actions right away. For example, P29 (31, Cis Woman, Middle Eastern, Gay, U.S.) compares this phenomenon to cameras in a store preventing theft, *"After adding the cameras they (customers) won't [steal] because they may think they can get away from the people (store employees), but they're pretty sure they can't get away from the cameras."* In this sense, being aware of the presence of an AI moderation system that can detect the large-scale embodied and immersive multi-user virtual environments and act in-the-moment is, according to many of our participants, an effective method to prevent harassment in social VR from even happening in the first place.

Technical Limitation: Reduced technical capabilities to address new forms of harassment. Despite most participants' belief in AI computational sufficiency, some still express concerns that the technical limitations often seen in AI-based moderation used in other online contexts would significantly hinder detection and moderation of more unique forms of harassment in social VR. For instance, P19 (25, Cis Woman, White/Russian, Pansexual, Germany) does not believe that existing AI-based moderation technologies are advanced enough yet to accurately detect and act upon voice-based harassment in the highly dynamic and synchronous voice communication space of social VR, *"I don't like to trust an AI with actions involving things like voice recognition in real time and punishing users. I still feel it's a little too unpredictable in many cases. Can AI detect language other than English? Can AI detect heavy accents? How about background noise? Can AI understand hints?"* For participants like P19, to what degree AI can accurately and effectively moderate more unique forms of harassment in social VR (e.g., harassment in real time and rich voice communication) is still questionable compared to human-based moderation, as prior experience with AI-based moderation has shown unsatisfactory results in comparable scenarios (e.g., real-time voice chat in Discord [37] for P19). Relying on AI, then, to moderate arguably more complex voice-based interactions in social VR would be *"too unpredictable in many cases."*

Perceived technical insufficiency of current iterations of AI-based moderation also extend to uniquely embodied harassment in social VR , as P17 (18, Cis Man, White, Gay, U.S.) describes, *"If out of nowhere you have some super loud person with a gigantic avatar who's obstructing things...it'd be very easy for AI to moderate it cuz that guy is obviously a troll. But in other instances, maybe like horror maps...it could definitely be a lot harder because you are supposed to disrupt people almost."* According to P17, it may be simple for AI-based moderation to detect sudden and unusual physical movements to mitigate embodied harassment (e.g., using a gigantic avatar to obstruct others). However, together with the Interpretation Limitation mentioned in the previous section, AI may not posses the technical cognitive capabilities to unpack why such physical movements happened and what motivated a given user to do so (e.g., as part of a gameplay), hampering its usefulness in complex situations.

4.1.3 AI-based moderation overcomes potential subjective biases of individual human moderators but may introduce new equality limitation. Most of our participants actually perceive AI-based moderation as a fairer system to deal with harassment in social VR than human-based moderation, a surprising and hopeful contrast from prior AI moderation literature, which often focuses on the ways in which AI moderation can be more systemically problematic than human moderators [2, 23]. P2 (25, Cis Woman, Black, Lesbian, U.S.) explains, "*It (AI) certainly doesn't have emotions. It doesn't pick sides. So, it's going to just dispense justice the way it is.*" For participants like P2, by virtue of not being human, AI must not have characteristics associated more with humans than machines (e.g., emotions), and are therefore unable to be influenced either by its own emotional biases (*"doesn't have emotions"*) or external emotional pressures (e.g., the pressure to *"pick sides"*).

Human-based moderation, on the other hand, is viewed as a comparatively more emotionally biased process that can be particularly harmful for minority communities in social VR. P35 (40, Cis Woman, Native and Hispanic, Lesbian, U.S.) thus emphasizes, "*you could imagine a human being moderator in AltspaceVR, assuming an African American is doing harassment above and beyond a white person, where a white person could do the same thing and not be deemed as a harasser.*" Here P35 powerfully highlights how human moderators bring in their own preconceived notions of what harassment looks like and who perpetrates harassment. This subjectivity can be particularly harmful and discriminatory against social VR users who already face marginalization at a large scale in the offline

world (e.g., African Americans). Thus, for the majority of our participants, AI's programmatic and non-emotional nature creates the perception that it is going to be less biased on an individual basis than human moderators, particularly for or against marginalized individuals.

As a result, the majority of our participants also express more trust in AI moderation than in humans to maintain privacy and security when dealing with harassment in social VR. P2 (25, Cis Woman, Black, Lesbian, U.S.) and P20 (24, Cis Woman, Black, Bisexual, U.S.) explain,

"*Maybe an AI wouldn't leak information from meetings. But humans have the tendency of saying stuffs of these conversations outside VR and that's very unprofessional.*" (P2)

"*if it's a human being taking the place of a robot and gets to hear something, he could actually use those things against me or use it to also harass me and tamper with my emotions too.*" (P20)

Both P2 and P20 identify offline as Black women working in tech and education, respectively, who often have to use social VR for meetings related to their offline work. For them, dealing with harassment in social VR is a matter that requires serious protection of their privacy and security, as it is often intrinsically tied to their offline identities and could have offline repercussions. Indeed, in P2's case, because the social-VR *"meetings"* she is referring to are company meetings for her offline job, a privacy leak via an *"unprofessional"* human moderator could actually have tangible effects on her offline professional life and could open her up to even more harassment within her workplace. Given many social VR platforms' policies on moderating user-created worlds and events [86, 94], P2 and P20's concerns that a human moderator could obtain and weaponize private information are not unfounded, as the human moderator in a social VR room created specifically for a workplace meeting is likely to be an offline coworker with significant moderation power. For both participants, AI is more trustworthy as it does not have the capacity to work beyond its programming (i.e., will not leak information when it is not programmed to do so) and/or the same desire - indeed an emotion - as humans might to use people's personal information against them. This resulting sense of increased privacy and security is especially vital for marginalized individuals and communities that are most likely to be targeted in their online and offline lives.

Equality Limitation: Creating the potential for new unfair and unequal power dynamics. However, some participants also raise a critical question: *Is AI-based moderation truly fair and without subjective bias?* Indeed, P11 (18, Trans Man, White, Queer, U.S.) points out, "*AI is only capable of having the biases it's programmed with.*" Here P11 is referring to how AI systems can exhibit biases that stem specifically from their programming and data sources, e.g., how data is obtained, how algorithms are designed, how AI outputs are interpreted, and most importantly, whose value is reflected in the design/development process.

According to some participants, this concern is particularly reasonable if AI is built by certain people who are "privileged" to have a voice/role in designing and deciding how social VR should be moderated and who can moderate. This may create unfair and unequal power dynamics if not explicitly addressed and mitigated during the AI development process. As P31 (44, Cis Woman, White, Straight, U.S.) argues, *"Some developers say, 'This is stupid. You can't be harassed in VR.' Because to them, it's only harassment if you're in physical danger."* According to P31, some social VR developers may not be equipped to empathize with victims of harassment if they themselves have not experienced harassment, either within or outside of social VR, or if they personally define harassment in a way that does not account for all types of harassment experiences. As an example, P31 later went on to describe how an AI moderator built specifically by men may not account for types of harassment commonly experienced more by women, such as sexual assault or sexist aggressive language, simply because they do not share the same level of empathetic experiences with women. As a result, they may not be able to create effective AI systems to moderate harassment incidents in social VR if intervention efforts are not made to proactively ensure that various views and experiences are accounted for during the AI development process. Thus, our participants envision several ways future AI-based moderation systems can be crafted to maximize the benefits and minimize any limitations associated with their use.

4.2 Envisionings for Overcoming Limitations of AI-based Moderation to Address Social VR Harassment

While acknowledging and outlining the **Interpretation Limitation, Technical Limitation,** and **Equality Limitation** in existing AI moderation systems, participants also pointed towards many promising opportunities for leveraging AI-based moderation to address harassment in social VR. As mentioned earlier in this paper, none of our participants have the sufficient experience and background to comment on specific technical suggestions to address AI's current *Technical Limitation*. Yet, they envision three essential recommendations for designing future AI-based moderation systems to address all three limitations while maximizing the identified promising opportunities for its use.

4.2.1 User-human-AI collaboration as a Comprehensive Approach for Improving AI-Based Moderation to Address Social VR Harassment. The vast majority of participants indicate *user-human-AI collaboration* as one of the most important and comprehensive approaches to design future AI-based moderation to handle emergent harassment in social VR. In this novel moderation system, social VR users, human moderators, and AI moderators work together as a collective team with each party occupying a distinct role on the team to achieve a common goal (i.e., addressing social VR harassment). Our participants especially envision three foundational considerations to design this *user-human-AI collaboration* moderation system for managing social VR harassment in a way that addresses the *Interpretation*, *Technical*, and *Equality* Limitations described in 4.1.

Users as Vital Collaborators for Moderating Social VR Harassment. First, the majority of our participants feel that a fundamental principle of leveraging user-human-AI collaboration to address social VR harassment is to intentionally view social VR users as a vital collaborator in moderation efforts rather than merely having human- or AI-based moderation imposed upon them. For instance, P34 (21, Cis Man, White, Straight, U.S.) describes how

AI can learn from the views of actual users to make the moderation system user-focused and and to keep the power to set up community standards in the hands of users themselves,

"It would make it a lot easier when you give the community an ability to be like, 'This person, not good' and then if a lot of people say, 'This person not good,' the AI will pick that up and represent that data in some way and then if somebody wishes to take action based on that data. Once a person's gone too far, they [human moderators] can then make that decision based on the evidence being shown."

In this case, repeated reports from individual users (i.e., *"the community"*) help the AI moderator to learn and incorporate general norms of what behavior is considered harassing in social VR, which in turn streamlines the decision-making of human moderators (i.e., "human" in user-human-AI collaboration) and codifying user-focused community standards. P6 (27, Trans Woman, Non-Binary Fluctuating, Biracial White Canadian and Indigenous Canadian, Asexual, Canada) further underscores the importance of establishing collective user awareness of harassment within said user-human-AI collaboration moderation system,

"Ideally it's a balance between both, crowd sourced accountability being the forefront, supported by human compassion and AI analysis behind the scenes to determine resolution. You need the people in all situations to be aware of what is and isn't acceptable before you can expect an AI to understand any of it. If people hold themselves accountable they then expect others to do the same, if there's a clearer understanding of how boundaries change depending on who's talking among people then eventually that translates to the AI learning these things either by way of pre-programming or machine learning."

According to P6, before AI and even human moderators can be expected to understand the intricacies of harassment in ever-changing and complex interactions, social VR users and communities themselves need to understand their own boundaries and *"how boundaries change."* This creates a loop of user-human-AI collaboration, wherein users define harassment over time across many interactions through existing mechanisms such as reporting and muting. By codifying users as a vital and necessary collaborator in the loop, this user-human-AI collaboration moderation system overcomes *Interpretation Limitation* through incorporation of community needs and *Equality Limitation* by giving users more comparative power in the moderation process than traditional human-based or AI-based moderation systems.

Multi-Level Decision-Making to Balance Advantages of Human Moderators and AI Moderators. Second, our participants view user-human-AI collaboration as a system that utilizes multi-level decision-making to leverage each actor's comparative advantages and balance their comparative shortcomings to address *Interpretation Limitation* and *Technical Limitation*. Here, participants specifically refer to each "level" as the decision-making of the *human moderator(s)* (e.g., Event Hosts and company-employed human-based moderation teams), the decision-making of the *AI moderator(s)*, and the decision making of *actual social VR users*. P28 (22, Cis Man, Hispanic, Bisexual, Guatemala) explains why such multi-level decision-making is logical, *"Human and AI [working together]... let the AI do all the things that us humans literally cannot do, right? I've always thought of computers as extensions of what we can do."* Inherent in P28's statement is a sense that user-human-AI collaboration works to not only reduce limitations of each "level" or actor, but actually *extends* possibilities beyond any one actor's abilities.

Manifestations of this multi-level decision-making process varied by participant. For participants like P6 (27, Trans Woman, Non-Binary Fluctuating, Biracial White Canadian and Indigenous Canadian, Asexual, Canada), social VR users themselves should act as the first line of defense against harassing behavior, monitoring situations and alerting their human and AI moderator collaborators for further review, *"crowd sourced accountability...supported by human compassion and AI analysis."* In contrast, given the speed and breadth with which AI can monitor spaces compared to humans, other participants believe AI to be better suited for basic and consistent platform-wide monitoring of suspicious behavior, such as to *"identify the potential risks or people who are of risk"* (P34, 21, Cis Man, White, Straight, U.S.), while higher-order, contextual decision-making could be delegated to their human moderator collaborators with constant feedback from the actual social VR users. The role the human moderator plays from there can vary, including: choosing a punishment suggested by the AI (P32, 30, Queer, Hispanic, Bisexual, U.S.); acting as the *"decisive"* (P34, 21, Cis Man, White, Straight, U.S.) actor in ambiguous cases; and/or being pulled in to adjudicate ban appeal cases (P31, 44, Cis Woman, White, Straight, U.S.)

Regardless of their specific role, though, this multi-level decision-making process simultaneously increases the ability of the user-human-AI moderation system to monitor harassment while also reducing the risk of human moderators becoming *"worn out by stress"* (P25, N/A, Cis Man, Black, Straight, U.S.) from constant moderation vigilance. This in turn saves valuable human resources for more complex decision-making, such as determining if a "second chance" decision should be made (e.g., sending a warning signal to a harasser instead of immediate punishment). P31 (44, Cis Woman, White, Straight, U.S.) details this potential process, "*you block someone, it sends a report off to the moderation AI and it says, 'All right, well, this is the first report we've ever gotten from this person. No big deal.' But if that person gets 20 blocks, maybe kick them out. Then if they appeal, then get a real person in.*" According to P31's vision, AI simultaneously flags potential harassment incidents for human collaborator review and provides a warning to the user to halt their behavior, giving human collaborators more time to engage in complex decision-making. In this way, concerns regarding AI's inadequate interpretation of sociocultural contexts (*Interpretation Limitation*) and its technical limitations to handle new forms of harassment (*Technical Limitation*) can both be addressed by designing a system that does not require AI to be *better than humans* at a task that humans have a comparative advantage in. Instead, AI, human moderators, and users will *work together* in this system to achieve essential goals of sociocultural nuances when dealing with harassment in social VR.

Diversifying the Human Moderators in the Loop. Third, the *Equality Limitation* lies in the issue of whose perspectives are being incorporated into the initial design and building process of an AI moderation system. This concern prompts some participants to advocate for demographic and experiential diversity within the human moderator's part of the user-human-AI collaboration system to mitigate bias against vulnerable populations in social VR. P31 (44, Cis Woman, White, Straight, U.S.) explains,

"I think if you're going to have a team of people who are there to deal with harassment, that team needs to be primarily composed of people who experience harassment. [...] So it should be minorities, and women, and LGBTQIA community. Because they know what it looks like. They experience harassment from the receiving end, which can be vague, and squishy, and subtle, and maybe seem like not a big deal if you're not the one experiencing it."

According to P31, AI can be prone to inheriting or learning bias from its creators, as well as the users and human collaborators it works with towards moderation goals. Deliberately diversifying the human moderator element of the larger collaboration, especially to create a balance of power between minority and majority voices, is thus important and could look like having the human moderator teams consist of women, LGBTQ, minorities, and other individuals most likely to face harassment, as well as people who are not typically considered marginalized. Such a balance would not only address the potential problem of human moderators catering too much to the views and preferences of the majority, but would also establish and reinforce more inclusive views on handling harassment into the AI's programming through iterative learning.

4.2.2 Leveraging Code Source Transparency and User-Controlled Creative Customization of AI Moderators to Address AI's Equality Limitation. For many of our participants, the potential *Equality Limitation* of using an AI-based moderation system to manage harassment in social VR is fundamentally a trust issue that reflects the power imbalances between social VR platforms (e.g., VRChat, Microsoft's AltspaceVR, Meta's Horizon Worlds) and their users. In other words, whether/how social VR users will perceive an AI-based moderation system as fair and equitable is dependent upon if users trust the social VR platform to design and use the AI-based moderation system in a way that meets their needs. To address this limitation, our participants highlight the importance of (1) *code source transparency* to build user trust in the AI-based moderation system; and (2) *user-controlled creative creation of AI moderators* to customize personal experiences of moderation in social VR to meet individual needs.

Regarding **(1)**, some participants argue that knowing and having access to read-only documentation that details the source code used to build and train the AI moderation system would be necessary for some social VR users to feel informed on how their data is being collected and used to detect, determine, and handle harassment in social VR. P37 (25, Cis Man, Asian, Mostly Straight, U.S.), himself an active voluntary community moderator in VRChat, elaborates,

"I'm fine with it [AI moderation] as long as they're open to how they're using it. Because a lot of people in VRChat are coders. A lot of them are techy people. If VRChat is open to how the tools are being used, like what's the programming code to it, what is the database it's being put into, what process is being used for it, then I'm totally open for that. But once you put the word, 'We're using AI to record your voice', VRChat's going to die, like instant die, because I know for AI to work, you need to gather a lot of data, a lot of listening into people's conversations, and once that happens, VRChat's dead. The best way [to avoid that] is just have the code open. If you have that source code just open to the public that people can read, and if VRChat's open to criticisms and changes to the source code, then people are fine with it."

P37's extensive reasoning provides insight on the benefits of code source transparency for trust and acceptance in two ways. First, for social VR users who feel they have the technical know-how to interpret how lines of code translate into AI moderation decisions, a platform such as VRChat releasing a read-only version of the AI moderation source code would help build a sense of agency and awareness about how the AI is being used to monitor their activities within the social VR space. These users, including P37 (25, Cis Man, Asian, Mostly Straight, U.S.) and P11 (18, Trans Man, White, Queer, U.S.), are people who build user-generated content in VRChat (e.g., custom-made avatar designs, worlds, and events), or are *"coders"* and *"techy"* people, and thus presumably have a level of technical experience with reading and understanding source code that makes this a feasible venture.

Second, even for users who lack the technical know-how, the act itself of releasing a readable version of the source code is a demonstration of *transparency* on the part of a social VR platform's company, i.e., a sign of good faith that the company intends to keep its users as informed as possible about how this system is being used and designed. P37 contends that, if companies go a step further by also being *responsive* to their users' concerns, *"if VRChat's open to criticisms and changes,"* then users will be far more likely to accept AI-based moderation. Adopting a code source transparency approach, therefore, may act as a way to balance out the otherwise unequal power dynamics existing between the platform and developers who design and develop the AI moderation system and its users (i.e., *Equality Limitation*) by 1) putting the power to know and understand how their information is being used and how moderation regarding harassment happens into the hands of users who can interpret that information, and 2) sending signals of transparency and responsiveness to users who do not have the requisite skills to interpret a code source.

Regarding **(2)**, the issue at the core of *Equality Limitation* in social VR lies in the imbalances of power disproportionately favoring the views and design choices of companies and their developers in regard to how AI-based moderation should work. This thus means that platform-wide implementation of a new social VR feature from a top-down approach (e.g., implementing platform-wide AI-based moderation) may invariably affect individual users' abilities to control their own experiences to meet their needs. To mitigate how this loss of control may drive users away, platform designers and developers often implement *customization* of features, as *customization* has been heavily linked by HCI and other disciplines to greater feelings of self-efficacy, agency, and control over one's experiences [48, 80]. Therefore, on the whole, our participants advocate for customization to accommodate many varied envisionings of what an AI moderator would or should look like to elicit comfort and trust and to grant agency to users to craft their own experiences based on their specific needs, thus addressing the *Equality Limitation* of AI-based moderation in social VR.

First, for some of our participants, customizing *humanoid physicalized AI moderators* would provide experiences of comfort and acceptance when interacting with AI-based moderation. For them, the familiarity of a humanoid appearance will help them to feel that their interactions with AI moderators are more comfortable, natural, and intuitive, especially when contacting a moderator to initiate an intervention or to seek recourse and support. Many participants

additionally expect that they would be able to customize such humanoid AI moderators to have an approachable and considerate personality (*"very, very nice and not mean"* - P10 (N/A, Genderqueer Feminine Presenting, Biracial White and Black, N/A, U.S.) and have human anthropomorphized elements to increase that familiarity (e.g., *"tend to act like a human moderator, maybe the voicing and other stuff"* - P17, 18, Cis Man, White, Gay, U.S.). Interestingly, some participants, including P18 (29, Cis Male, Mixed Race, N/A, U.S.) and P23 (29, Cis Woman, Black, Lesbian, U.S.), also indicate that male-presenting AI moderators would be best, often because of the gender stereotypes associated with other "protector" roles within their sociocultural context (e.g., offline-world police or soldiers). While gendered differences are not the focus of this study, it is valuable to note how offline-world sociocultural biases on gender stereotypes creep their way into social VR spaces. Nevertheless, for social VR users who feel that humanoid, physicalized AI moderators would help them to feel safer in social VR, the option to customize how they see an AI moderator to fit such a description would give them control over their own experiences of how moderation should operate.

Second, most participants who want a physicalized AI moderator prefer customization to create *non-humanoid physicalized AI moderators* - e.g., *"a genderless robot"* (P11, 18, Trans Man, White, Queer, U.S) - as it would offer a sense of transparency when using AI-based moderation. Indeed, specific visions of a non-humanoid AI moderator by participants like P27 and P33 (Both 25, Cis Man, Black, Straight, U.S.) and P39 (25, N/A, Black, Straight, U.S.) reflect an expectation that an AI moderator's lack of human characteristics should be reflected in their appearance, i.e., being transparent about its non-human nature. For instance, P19 (25, Cis Woman, White/Russian, Pansexual, Germany) feels that AI masquerading as a human would not be true to its nature as a non-human technology, *"I never really liked AI trying to be human looking, it's a robot and I want it to embrace it."* P16 (25, Cis Man, Black, Straight, U.S.) expresses similar sentiments, stating that having the AI appear as non-human would serve as a visual indicator that provides knowledge about what a user is interacting with, thus providing more confidence and a sense of control over their experiences within social VR. In this sense, non-humanoid presenting AI moderators in social VR would create a greater level of transparency regarding who people are interacting with, even if the AI moderator still retained some personable traits that help comfort levels and approachability (e.g., friendliness).

Third, many participants would prefer customization to allow for *non-physicalized AI moderators* (i.e., no physical presence in social VR, operates behind the scenes) to provide users the ability to avoid an uncomfortable sense of restriction and "being watched" when using AI-based moderation. The primary concern of our participants is, as stated by P16 (25, Cis Man, Black, Straight, U.S.), that *"people should be free to interact and not feel that they're being overwatched by someone."* When an AI moderator is given a physicalized virtual body, no matter the nature of said body, it creates a secondary presence that can often feel *"highly creepy, like Terminator creepy"* (P31, 44, Cis Woman, White, Straight, U.S.). Additionally, P32's (30, Queer, Hispanic, Bisexual, U.S.) account summarizes well the fundamental trade-off between efficiency and comfort when AI moderators are physicalized,

"In the short term, we might be having an efficient moderation and reduction of harassment in the visual world. But let's be clear that people wouldn't want to be followed around, people wouldn't want to be treated as machines. So eventually, there might be a form of protest against it directly, indirectly. It will affect both the companies that provide those services for their financial output and the users experience."

As P32 reveals, physicalized AI moderators may actually *remove* feelings of agency and comfort rather than protecting social VR users by making them feel like they are subject to the moderation system and are not free to interact in social VR. Rather than feeling as though they are constantly being watched, social VR users need to be able to interact without such discomfort while still retaining the safety associated with AI moderation, and thus strike a balance between feeling secure in the system's ability to address harassment and feeling able to interact freely. Given these varied needs, then, customization of an AI moderator's (non)appearance appears to be an effective approach to balance all needs while giving users a sense of control, thus addressing the *Equality Limitation* in a nuanced way.

5 DISCUSSION

In answering **RQ1**, on the one hand, our participants feel that AI-based moderation brings unique and computationally powerful opportunities to better manage emergent harassment in novel social VR spaces based on its ability to make consistent judgements and decisions, to monitor and intervene overcome potential subjective biases inherent in human-based moderation. On the other hand, they also feel that AI-based moderation as seen in other online contexts may show certain *Interpretation Limitation*, *Technical Limitation*, and *Equality Limitation*. Grounded in these complex understandings, our participants thus overwhelmingly advocate for the use of user-human-AI collaboration to address all above-mentioned limitations, specifically by: emphasizing users as vital collaborators to reduce AI's *Equality Limitation*; encouraging multi-level decision-making to make up for AI's *Interpretation Limitation* and *Technical Limitation*; and by diversifying the human moderator element of the collective to address the *Equality Limitation*. Our participants additionally envision improving the trust, transparency, and sense of agency in AI moderation through code source transparency and customized AI moderator design to address the *Equality Limitation* (**RQ2**).

In this section, we further discuss how our findings help re-envision AI's new roles in addressing the complicated, sometimes even contradictory, needs of individual users and communities while maintaining the integrity of the social VR experience, thus offering new insights to innovate existing content moderation approaches to better combat harassment in emerging online social spaces. We also highlight potential future directions on how AI-based moderation, especially in the form of user-human-AI collaboration, can and should be approached and designed to achieve safer and more nuanced online experiences in the future.

5.1 Re-Envisioning AI's New Roles in Achieving Nuanced Moderation Mechanisms to Combat Online Harassment

One key insight from our study lies in how AI is re-envisioned by many social VR users to play a more dynamic, collaborative role in achieving nuanced moderation. Indeed, participants view AI as needing to interact more at the level of human expectation and to be more proactive in its adaptation to the unique environmental considerations of social VR rather than reactive as is seen in other moderation modalities. This insight is urgently needed for advancing existing literature on AI-based moderation and social VR – not only have little to no studies explicitly explored AI-based moderation in this context, but prior empirical research also does not depict a consistent image regarding the role *any* existing moderation practices may play in managing said harassment [5, 26]. In fact, existing moderation practices on major social VR platforms often place a tremendous burden on individual human moderators and users rather than on the platforms themselves [55, 56, 58, 86–89, 94, 95]. Our findings directly fill this gap by proactively envisioning AI's new roles in innovating existing moderation approaches, an especially important task to guide future AI system design for combating emergent social VR harassment as AI-based moderation has yet to be implemented in these spaces.

AI as a Potential New Form of Empowerment for Addressing Harassment Challenges in Social VR. As detailed earlier in this paper, there are at least three main challenges to preventing and mitigating emergent harassment in social VR, including the lack of consensus on how to define harassment [5, 26] (*Challenge 1*), the limitations of existing tools to manage new forms of harassment [5, 26] (*Challenge 2*), and the concern about subjective bias associated with human moderators to effectively identify and manage such harassment [26] (*Challenge 3*). In general, AI is positively envisioned by many of our participants as a promising mechanism to better address all these challenges in social VR. This perspective is particularly hopeful in comparison to prior research that often vilifies AI-based moderation as a mechanism that allows social issues such as sexism and racism to proliferate [2, 23, 30].

Instead, AI is envisioned by our participants as a potential new form of empowerment to better protect social VR users, especially marginalized communities, from emergent harassment, often because of its perceived consistency in judgments (addressing *Challenge 1*) and computational power, as well as its overall ability to be everywhere all at once (addressing *Challenge 2*). While human moderators may not catch or acknowledge a harassing incident in social VR - either because they do not have the proper life experiences to recognize harassment [68] or because of logistical limitations [17, 24, 62, 97, 98], AI-based moderation's use means a harassing incident can be recognized and dealt with nearly instantaneously. Importantly, when built and trained carefully and with cultural sensitivity in mind as recommended by our participants, AI is not subject to the same kinds of biased decision-making that individual human moderators can perpetuate. In other words, marginalized social VR users feel that harassers are less likely to get away with their behavior if an AI is adjudicating rather than a human who could be swayed to give a lighter punishment or who may even side with the harasser if the harasser is someone with a position of inherent power (e.g., white, male, heteronormative, and U.S.-based). Thus, the incorporation of AI moderation into social VR can potentially empower marginalized users to feel that they are not going to be unfairly targeted because of their inherent characteristics (addressing *Challenge 3*).

AI as a New Ally for Combating Harassment in Social VR. While prior research has explored several approaches for leveraging AI for advancing content moderation within contexts such as social media [14, 30, 61, 84, 85], text-based online forums (e.g., Reddit) [15, 21, 33, 34], and live streaming platforms (e.g., Twitch) [8–11, 69, 98], many studies tend to depict AI as a passive tool for human moderators to deploy or a supplemental component in human-based moderation systems [31, 35]. Examples include using AI as a flagging and filtering tool within text chat on social media [7, 19, 20, 35, 45, 63] and gaming [42, 78], or using AI to monitor and flag voice-based social spaces based on Natural Language Processing [3, 37, 64, 66, 78]. In this sense, traditional AI-based moderation is either perceived as a simple (if powerful) tool that performs menial (e.g., flagging specific words in a text post) or even complex tasks (e.g., determining emotion from voice to detect harassment) to assist humans' moderation efforts, or as a standalone mechanism with little oversight and cooperation with human moderators, actual users, or the unique moderation needs of specific communities [41].

In contrast, our participants' visions of user-human-AI collaboration as a promising future model to manage emergent harassment in social VR significantly differ from the above-mentioned traditional AI-based moderation. In our findings, AI is envisioned as an *ally* with human moderators and community members/users alike, creating a cooperation link to address harassment on all fronts. This vision constitutes a specific type of interaction whereby collaborators (i.e. human users, human moderators, and AI moderators) work together via *"mutual goal understanding, preemptive task co-management and shared progress tracking"* [90]. This vision is reflective of a concept within HCI known as human-AI collaboration, more specifically human-AI teaming (HAT), which is characterized as *"interdependence in activity and outcomes involving one or more humans and one or more autonomous agents, wherein each human and autonomous agent is recognized as a unique team member occupying a distinct role on the team, and in which the members strive to achieve a common goal as a collective"* [60]. Building upon these two concepts of collaboration and HAT, user-human-AI collaboration/teams for managing social VR harassment can be characterized by a level of interdependence between one or more humans (i.e., human users and moderators) and one or more AI agents (i.e., AI moderators), with each member occupying a clear role on the teamand who are working toward achieving a shared goal as a collective unit [60].

Our participants view this model as a promising way to address all potential limitations of both AI-based moderation *and* human-based moderation identified in this paper. Such collaboration would be able to (1) avoid perpetuating any new unfair and unjust power imbalances through purposefully designing human collaborator moderation teams to include marginalized individuals (addressing the Interpretation Limitation of AI *and* human bias issues [67, 100]); (2) technically improve AI to be less biased at the beginning stages of formation (addressing the Technical Limitation of AI *and* boosting agency through community involvement); and (3) ultimately

improve AI's ability to be sensitive to marginalized individuals' specific experiences within their sociocultural context (addressing the Equality Limitation of AI *and* incorporating human understanding of situational contexts [10, 17, 65, 70, 76, 98]). In this model, human collaborators iteratively work with the AI as it makes decisions to correct its path when they see a lack of sociocultural and technical nuance, while AI leverages its superior powers in consistency and range to account for human moderators' deficiencies. Additionally, user/community input becomes incorporated into this process in the ways human-AI teams are alerted by users to the presence of harassing persons or behaviors when AI misses it (e.g., flagging [42] and community input at the time of algorithmic building [41]).

AI as a Reflection of New and Possibly Unfair Power Dynamics Involved in Social VR. Although using AI-based moderation to prevent emergent harassment in social VR is envisioned in an overall optimistic way, some participants still share reservations about its use, particularly in light of their mistrust in certain social VR companies (e.g., Meta). In fact, similar to existing concerns about AI's general lack of transparency [13, 22, 28, 46, 83], AI can be envisioned as a reflection of new and possibly unfair power dynamics involved in social VR. Prior research has doubted the goodness of the intentions behind *any* moderation efforts under a centralized, profit-driven system [67], especially when these suspicions are compounded by increased fears of AI moderation perpetuating unfairness and marginalization at a global level [28]. This fear and hesitancy is also reflected in the statements of the few participants who envision that AI-based moderation, if not constructed carefully, may reinforce or introduce unfair power dynamics.

Therefore, of the participants who expressed such reservations, one important strategy to AI-based moderation acceptance in social VR would be if a read-only version of the source code that built and drives the AI moderator were to be made publicly available by social VR companies, and if these companies would be amendable to feedback from users on the AI's inner workings. This code source transparency policy would be a way to simultaneously serve as a signal of good faith that social VR companies are willing to be honest and collaborative about their practices while also increasing user awareness and knowledge of AI moderators' operations and decisions. Both are important in establishing trust in a moderation system and can be severely negatively impactful when missing [13, 22, 28, 46, 83].

5.2 Future Directions for Designing AI-Based Moderation for Safe Novel Online Spaces

Grounded in our findings and reflections, we propose three vital high-level principles aimed at rethinking how HCI researchers and practitioners approach the challenges inherent in mitigating and addressing harassment in unique online social spaces, particularly when including AI and its associated new power dynamics in the mix. We view these principles as facets of a novel moderation system in which individual needs, community needs, and platform-wide needs are accounted for, with the purpose of creating a user-focused AI-based moderation system that avoids falling into the assumption that a one-size-fits-all AI moderation system is either appropriate or desirable.

Principle 1 - Designing AI moderators for the individual: The importance of appearance customization. One insight from our findings is how highly varied social VR users' conceptualizations of what an AI moderator should look like really are - i.e., humanoid, non-humanoid, or non-physicalized. This variety thus leads to a fundamental principle for designing future AI-based moderation systems in novel online spaces like social VR: customization of an AI moderator's appearance (or lack thereof) is the first and most straightforward way to ensure that individuals feel a sense of agency, as they will be able to control how they receive the type of comfort and support they need most from an AI moderator. Such agency is sorely lacking in current iterations of social VR moderation pipelines in two ways. First, human-based moderation on social VR platforms is made almost invisible due to the distant and often unresponsive reporting systems [87, 89]. Second, even when such moderation is made visible, it is often considered unhelpful because these companies either make it unclear as to when moderators are deployed in social VR (e.g., AltspaceVR's Concierge system [95]) or by in-VR platform representatives having limited power to adjudicate harassment incidents (e.g., Meta Horizon's Community Guides [57]).

We thus suggest that social VR designers should consider building a customization feature for how users wish to see (or not see) an AI's presence when incorporating AI-based moderation in platforms. While we recognize the technical difficulties inherent in this suggestion, we posit that this customization could take the form of a set of prearranged avatars representing some variety (e.g., humans of various genders and races, a few types of animals, and a few types of robots), rather than a complete customization kit. Additionally, such a kit could provide some pre-set personality options, such as "harsh" or "nurturing", to cater to the emotional needs of individual users. In much the same way that Siri's voice can be customized within a range of possibilities, we believe approaching future AI-based moderation in social VR with that same simple but impactful customization can go a long way in improving the individual's feelings of comfort and safety when leveraging moderation to manage harassment in novel social spaces.

Principle 2 - Designing AI moderators for specific communities: The importance of achieving sociocultural awareness. Overall, our findings demonstrate that one of the core anxieties surrounding AI-based moderation - particularly for marginalized communities - is that existing AI systems may lack the necessary complex skills to navigate and understand intricate sociocultural contexts. For example, communities and individuals whose interactions - particularly amongst minority groups - could be misinterpreted as harassing by the majority culture (e.g., words and phrases used for joking amongst friends) would be disproportionately harmed by an AI moderator built with a one-size-fits-all (i.e., a majority-rules) mentality. Thus, these groups might prefer a system that employs warnings and second chances to ensure that they are not indiscriminately targeted for punishment. Both AltspaceVR and Meta Horizon indeed explicitly mention giving warnings before taking more severe action as a common practice in their moderation pipelines [57, 89]. What is currently lacking, however, is clear communication of when something warrants a "warning" or, in VRChat's case, an escalation starting at suspension [86], versus more severe action. Therefore, another important principle for designing

future AI-based moderation should focus on achieving a sense of sociocultural awareness in the AI design process.

In doing so, we suggest allowing for collective communities and individuals within those communities to have the ability to both 1) define and train an AI system on what is and is not considered to be harassment; and 2) adjust how strict and autonomous an AI agent is when detecting harassment and carrying out sanctions. Picturing how this principle might potentially be realized in design will naturally depend upon the nature of interactions as facilitated by the sociocultural context embedded in the specific virtual space (e.g., user-created private spaces vs. public worlds), as the ability to gauge and define community needs in regard to what is harassment and how strict an AI should be become necessarily more difficult in the later.

For user-created private spaces/worlds on a social VR platform, this sociocultural contextualization principle could potentially be achieved through a focus on giving users and creators of a given space/world (e.g., Event Hosts in AltspaceVR [94, 96]) the power to iteratively work together with an AI to form a feedback system that operates within this particular space (e.g., give a warning first, automatically kick out/ban, consult the Event Host before taking action, etc.). For public spaces run by the platform itself, they might find a better balance between protection and context by leaning more towards an AI-based moderation system that provides warnings to harassers before taking direct action as is already common practice in AltspaceVR [89, 95] and Meta Horizon [57]. Existing platforms should, however, go a step further by providing some mechanisms that would allow two or more users who are engaging in a consensual exchange that is being flagged as harassment by an AI system to inform the system that the interaction is consensual and further action should not be taken. Regardless of the actual manifestation of such a feature, it is important that the AI only ceases further moderation escalation if *all parties* within the interaction mutually and separately agree that the interaction is consensual in this specific context.

Principle 3 - Designing AI moderators for all: The importance of user-human-AI collaboration. Arguably the most impactful and interesting revelation from our findings is our participants' immediate and strong gravitation towards using *user-human-AI collaboration* for addressing and mitigating harassment for all users engaging in novel and immersive online social environments. As previously detailed in 5.1, this system of user-human-AI collaboration is a fundamentally more interdependent and iterative way to approach moderation than in more traditional AI-based moderation in online social spaces, as it necessitates multiple team members that serve distinct yet cohesive roles to achieve a shared goal [60, 90]. Additionally, rather than perpetuating the problems inherent in the human-based moderation systems currently employed by social VR companies - i.e., overemphasis on individuals managing their own experiences without transparent or responsive access to the human moderation team [57, 86, 95] - we uniquely propose that actions taken and (in)direct input given by users should be integrated into the moderation pipeline, such that *human and AI moderators are equally as dependent on users to continuously redefine harassment and shape their community as they are on each other.*

The specific construction and design of such a future *user-human-AI collaboration* moderation system is likely to be contextually determined - based on community and site-wide size, platform affordances and resources, and specific harassment needs and styles. Thus, rather than focusing on delivering specific technical recommendations in a one-size-must-fit-all manner, we instead highlight essential elements that must be included for building such new moderation systems in the future, particularly for the protection of marginalized populations.

First, social VR designers should seek to understand the comparative advantages of both human- and AI-based moderation in order to craft the specifics of the user-human-AI collaboration moderation pipeline for their platform. For example, AI's perceived lack of human emotions and biases can be used to create a sense of fairness and impartiality, acting as a check on human biases in moderation decision-making. Humans, on the other hand, can use their emotional capabilities to make more nuanced decisions that an AI is far more likely to miss, thus ensuring rather than decreasing the sense of human agency in a moderation system [13, 13, 22, 22, 28, 46, 83]. Second, the human elements of these teams will also need to be comprised of diverse individuals and perspectives to ensure that moderation practices are not dictated only by those in powerful and privileged positions, and who are therefor less likely to recognize and possibly empathize with the types of harassment minority populations face. This is not to say that these user-human-AI moderation teams should not have *any* individuals who might fall into the majority culture (e.g., cis-gendered, white, and heterosexual men), as a lack of representation of *any* type of user necessarily reduces the sociocultral nuance of the system. Rather, social VR platforms should strive to be intentional about crafting a team of individuals that creates a balance of power between minority and majority voices, which could inherently mean having a larger ratio of individuals of varied identities (e.g., in gender, sexuality, race, etc.) compared to the majority identity to ensure equitable teams. Finally, beyond just the human-AI moderation team, the communities within these novel online social spaces need to act as another collaborator through direct (e.g., soliciting written opinions) and indirect (e.g., blocking and reporting) feedback processes to shape more specific and detailed community guidelines that reflect the varied needs of their specific user population, thus creating a user-human-AI collaboration system that is multi-layered to better combat harassment in novel online spaces as a collective effort.

5.3 Limitations and Future Work

We acknowledge that our recruitment methods may lead to potential self-selection bias (e.g., active social media users who are also social VR users). Our sample is also overwhelmingly U.S.-centric (N=32). While this sample does represent a large user base of social VR, future work should focus on recruiting social VR users from different regions and cultures of the world to ensure a diversity of viewpoints on the merits and pitfalls of implementing AI-based moderation in social VR. In addition, although 39 participants for a qualitative study is considered healthy and exceeds the typical interview sample size for CHI of 12 [12], the results of this study should not be extrapolated to all social VR users. Our ongoing future work involves constructing and distributing a large-scale survey to a wider social VR audience using the results of this qualitative

study. Indeed, this study, being the first to our knowledge to explore perceptions of AI-based moderation within social VR, also leaves ample room for future directions in research. For example, most of our participants could not readily provide specific suggestions to improve the technical nature of AI to help combat the embodied and immersive forms of harassment in social VR, possibly because they lack technical AI expertise. Therefore, future research could seek to interview social VR users with technical experiences with AI development to gain more nuanced technical perspectives.

6 CONCLUSION

Social VR represents a unique online social space that is growing in prevalence and, in turn, creating higher risks for users to be harassed in far more embodied and immersed ways than seen in other online contexts. This concern has driven the need to understand how traditional moderation techniques should be re-envisioned, approached, and designed to reflect and accommodate for the complex harassment challenges faced by social VR users. Thus, our research unpacks the ways in which social VR users view AI-based moderation as presenting opportunities for mitigating harassment while also comprising inherent limitations, especially in comparison to human-based moderation. Our research additionally provides explicit insight into how social VR users envision the future design of an AI-based moderation system to meet the unique needs of social VR users. Our findings shed light on how social VR users envision users, human moderators, and AI working together in an iterative and holistic user-human-AI collaboration moderation system for optimal harassment mitigation, as well as the various individual-level, community-level, and platform-level needs that must be taken into consideration in the future. We hope that these insights guide HCI researchers' future efforts to design nuanced moderation systems to achieve safer and more inclusive novel online social spaces.

ACKNOWLEDGMENTS

We thank our participants and the anonymous reviewers. We also thank Catherine Barwulor for helping data collection. This work was supported by the National Science Foundation under award 2112878.

REFERENCES

[1] Team AltspaceVR. 2022. Making altspacevr a safer space. https://altvr.com/making-altspacevr-a-safer-space/
[2] Carolina Are. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies* 20, 5 (2020), 741–744.
[3] Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. 2021. CyberPolice: Classification of Cyber Sexual Harassment. In *EPIA Conference on Artificial Intelligence*. Springer, 701–714.
[4] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[5] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[6] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social VR: Implications for design. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 854–855.
[7] Hannah Bloch-Wehba. 2020. Automation in moderation. *Cornell Int'l LJ* 53 (2020), 41.
[8] Johanna Brewer, Morgan Romine, and TL Taylor. 2020. Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 757–769.
[9] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[10] Jie Cai and Donghee Yvette Wohn. 2022. Coordination and Collaboration: How do Volunteer Moderators Work as a Team in Live Streaming Communities?. In *CHI Conference on Human Factors in Computing Systems*. 1–14.
[11] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation visibility: Mapping the strategies of volunteer moderators in live streaming micro communities. In *ACM International Conference on Interactive Media Experiences*. 61–72.
[12] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
[13] Robyn Caplan and Tarleton Gillespie. 2020. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society* 6, 2 (2020), 2056305120936636.
[14] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.
[15] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.
[16] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
[17] Fatih Çömlekçi. 2019. Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. *Communication Today* 10, 1 (2019), 165–166.
[18] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
[19] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*. 121–126.
[20] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 11–17.
[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*. 1–13.
[22] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.
[23] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[24] Colin Ford, Dan Gardner, Leah Elaine Horgan, Calvin Liu, AM Tsaasan, Bonnie Nardi, and Jordan Rickman. 2017. Chat speed op pogchamp: Practices of coherence in massive twitch chat. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 858–871.
[25] Guo Freeman and Divine Maloney. 2021. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.
[26] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30.
[27] Dennis Friess, Marc Ziegele, and Dominique Heinbach. 2021. Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication* 38, 5 (2021), 624–646.
[28] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[29] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[30] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[31] Qinglai He, Yili Kevin Hong, and TS Raghu. 2022. The Effects of Machine-powered Content Modereation: An Empirical Study on Reddit. In *55th Hawaii International Conference on System Sciences (HICSS)*.

[32] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[33] 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. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[34] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[35] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[36] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

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

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

[39] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–35.

[40] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2.

[41] Yubo Kou and Xinning Gui. 2020. Mediating community-AI interaction through situated explanation: the case of AI-Led moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[42] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

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

[44] Saul Levmore and Martha Craven Nussbaum. 2012. *The offensive Internet: Speech, privacy, and reputation*. Harvard University Press.

[45] Emma Llansó, Joris Van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. Artificial intelligence, content moderation, and freedom of expression. (2020).

[46] Renkai Ma and Yubo Kou. 2021. " How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[47] Aaron Mak. 2022. I Was a Bouncer in the Metaverse. https://slate.com/technology/2022/05/metaverse-content-moderation-virtual-reality-bouncers.html

[48] Sampada Marathe and S Shyam Sundar. 2011. What drives customization? Control or identity?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 781–790.

[49] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society* 19, 3 (2017), 329–346.

[50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[51] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping pro-social interaction in VR: an emerging design framework. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[53] Sharan B Merriam and Elizabeth J Tisdell. 2015. *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

[54] Meta. 2022. Code of conduct for virtual experiences. https://www.meta.com/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/

[55] Meta. 2022. Notification of warning or suspension of your meta or oculus account. https://www.meta.com/help/quest/articles/accounts/account-settings-and-management/received-warning-or-suspension-notification-on-account/

[56] Meta. 2022. Reporting someone on Meta Quest and rift S. https://www.meta.com/help/quest/articles/accounts/privacy-information-and-settings/reporting-someone-on-oculus/

[57] Meta. 2022. Safety and privacy in Meta Horizon Worlds. https://www.meta.com/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/index-safety-and-privacy/

[58] Meta. 2022. Someone in my party is Bothering me. https://www.meta.com/help/quest/articles/in-vr-experiences/social-features-and-sharing/someone-in-my-party-is-bothering-me/

[59] Maria D Molina and S Shyam Sundar. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society* (2022), 14614448221103534.

[60] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors* 64, 5 (2022), 904–938.

[61] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.

[62] Jenny Preece and Diane Maloney-Krichmar. 2003. Online communities: focusing on sociability and usability. *Handbook of human-computer interaction* (2003), 596–620.

[63] Kim Renfro. 2016. For whom the troll trolls: A day in the life of a Reddit moderator. *Business Insider* (2016).

[64] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, Vol. 2. IEEE, 241–244.

[65] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).

[66] Shikhar Sakhuja and Robin Cohen. 2020. RideSafe: Detecting Sexual Harassment in Rideshares. In *Canadian Conference on Artificial Intelligence*. Springer, 464–469.

[67] Joseph Seering. 2020. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 107.

[68] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

[69] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.

[70] 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 (2019), 1417–1443.

[71] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*. IEEE, 225–226.

[72] Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna, Sami Safadi, Fouad Chebib, et al. 2020. Successful moderation in online patient communities: inductive case study. *Journal of medical Internet research* 22, 3 (2020), e15983.

[73] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria V Sanchez-Vives. 2009. Inducing illusory ownership of a virtual body. *Frontiers in neuroscience* (2009), 29.

[74] Weilun Soon. 2022. A researcher's avatar was sexually assaulted on a metaverse platform owned by Meta. https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5

[75] Hannah Sparks. 2021. Woman claims she was virtually 'groped' in Meta's VR metaverse. https://nypost.com/2021/12/17/woman-claims-she-was-virtually-groped-in-meta-vr-metaverse/

[76] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927.

[77] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[78] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*. 19–24.

[79] Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge university press.

[80] S Shyam Sundar and Sampada S Marathe. 2010. Personalization versus customization: The importance of agency, privacy, and power usage. *Human communication research* 36, 3 (2010), 298–322.

[81] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *new media & society* (2022), 14614448221109804.

[82] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.

[83] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[84] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.

[85] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.

[86] VRChat. 2022. Community guidelines. https://hello.vrchat.com/community-guidelines#:~:text=VRChat%20Moderation%20will%20take%20action,boards%2C%20and%20other%20online%20portals.

[87] VRChat. 2022. Submit a request – VRChat. https://help.vrchat.com/hc/en-us/requests/new

[88] VRChat. 2022. Terms of Service. https://hello.vrchat.com/legal

[89] Vryunji, Qian Wen, and Harrison Ferrone. 2022. Community standards - altspacevr. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/community/community-standards

[90] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381069

[91] Leijie Wang and Haiyi Zhu. 2022. How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 824–838.

[92] Sai Wang. 2021. Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital journalism* 9, 1 (2021), 64–83.

[93] Amy K Way, Robin Kanak Zwier, and Sarah J Tracy. 2015. Dialogic interviewing and flickers of transformation: An examination and delineation of interactional strategies that promote participant self-reflexivity. *Qualitative Inquiry* 21, 8 (2015), 720–731.

[94] Qian Wen. 2022. Event host guide - altspacevr. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/explore/host-events?source=recommendations

[95] Qian Wen. 2022. User safety and moderation - ALTSPACEVR. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/user-safety

[96] Qian Wen and Harrison Ferrone. 2022. Host tools overview - altspacevr. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/tutorials/host-tools-overview#safety-and-moderation-tools

[97] Ruth L Williams and Joseph Cothrel. 2000. Four smart ways to run online communities. *MIT Sloan Management Review* 41, 4 (2000), 81.

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

[99] Bingjie Yu, Joseph Seering, Katta Spiel, and Leon Watts. 2020. " Taking Care of a Fruit Tree": Nurturing as a Layer of Concern in Online Community Moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[100] Andrew Zolides. 2021. Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society* 23, 10 (2021), 2999–3015.