Challenges of Moderating Social Virtual Reality

Nazanin Sabri, Bella Chen, Annabelle Teoh, Steven P. Dow, Kristen Vaccaro, Mai ElSherief
University of California San Diego
USA
{nsabri,zhc028,c1teoh,spdow,kv,melsherief}@ucsd.edu

ABSTRACT

Recent years have seen a rise in social virtual reality (VR) platforms that allow people to interact in real-time through voice and gestures. The ephemeral nature of communication on these platforms can enable new forms of harmful behavior and new challenges for moderators. We performed virtual field research on three VR environments (AltspaceVR, Horizon Worlds, Rec Room). Based on observing 100 scheduled events, our analysis uncovered 13 distinct types of potentially harmful behaviors enabled by real-time voice, embodied interactions, and platform affordances. We witnessed potential harm at 45% of our observed events; only 24% of these incidents were addressed by moderators. To understand moderation practices, we conducted interviews with 11 moderators to investigate how they assess real-time interactions and how they operate within the current state of moderation tools. Our work sheds light on how moderation tools and practices must evolve to meet the new challenges of social VR.

CCS CONCEPTS

• **Human-centered computing** → **Social networking sites**; **Empirical studies in collaborative and social computing**; **Ethnographic studies**.

KEYWORDS

Content Moderation, Ephemeral Social Spaces, Virtual Reality, Virtual Ethnography, Interview

ACM Reference Format:
Nazanin Sabri, Bella Chen, Annabelle Teoh, Steven P. Dow, Kristen Vaccaro, Mai ElSherief. 2023. Challenges of Moderating 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, 20 pages. https://doi.org/10.1145/3544548.3581329

1 INTRODUCTION

Content moderation is a crucial, but challenging, responsibility for social media [59, 63, 118]. Traditional social media platforms feature post-based content, where a user posts text, images, or videos. In that paradigm, all material related to the post (e.g., comments, shares, retweets) remains available for ex-post moderation. Thus, on post-based platforms, moderators or algorithms act on a clear record of the content exactly as it was uploaded. However, with the recent rise of ephemeral social spaces, like voice-based platforms and virtual reality (VR) social spaces, this reactive approach will no longer suffice [70].

Social VR introduces fundamental changes to social media: the ephemeral nature of interactions, the additional modality of voice and gestures, and the added spatial dimensions of VR. And in turn, these changes usher in new challenges for moderation. For example, recent work identified that in voice-based social networks, interruptions or talking over another person could be considered a harmful behavior [70]. However, little is known about what new harmful behaviors ephemeral social VR has enabled or how moderators have responded.

In this work, we seek to understand the kinds of harmful behavior that have emerged within these nascent ephemeral social spaces, the strategies and mindsets around moderating, and the opportunities to improve moderation. To do so, we conducted two qualitative studies: a virtual field study based on participant observation and a set of interviews with current moderators of VR communities. For the first study, we observed a total of 100 scheduled events across three platforms (AltspaceVR, Horizon Worlds, and Rec Room). In the second study, we interviewed 11 moderators about their moderation experiences. Both our observations in VR and our interview data revealed insights into users' harmful behaviors, how moderators responded, and the challenges moderators faced.

Through our analysis of observation data, we identified emerging and existing types of harmful behavior. These novel forms of potential harm were a product of (a) voice, (b) virtual embodiment, and (c) platform affordances: disruptive movement, stalking of avatars, and enactment of physical and sexual violence, self-harm or suicide, to name a few of the observed behaviors. Harmful behaviors were observed in 45% of the attended events. Moderators were present in 51% of events where harms occurred and yet, only 60% of harms in actively moderated spaces were addressed. We also discuss our observations of moderation responses.

In the second study, through interviews with moderators, we present deeper insights into emergent moderation practices, different notions of what it means to be a moderator in these spaces, and perspectives on the impact of factors such as platform affordances on moderation practices. We found that difficulties of de-escalation in real-time encourage moderators to employ proactive measures to prevent harm. During events, the lack of persistent records results in the use of moderation strategies that attempt to achieve complete visibility of participants: world design, strategic avatar placement, and the use of teamwork, to name a few of the discussed methods.

Through this mixed methods approach, our work captures a broad view of the challenges of moderating social VR and the strategies currently used to address these challenges. We discuss how moderation tools and practices must evolve to meet the challenges of ephemeral social media platforms.

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

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581329

2 BACKGROUND

We situate our work within three areas of prior research. First, we discuss prior work that documents harm in online spaces and the potential impacts on individuals and communities. Next, we review prior research on content moderation including existing frameworks for mitigating harmful behaviors and the role of volunteer moderators. Finally, we focus on recent work related to harm and moderation strategies on ephemeral social media platforms.

2.1 Harms in Online Spaces

Prior work in the HCI research community has documented four types of harm that often occur in online social spaces: (1) technical [20, 41, 44, 49, 106], (2) privacy and surveillance [104, 134], (3) content-based [43, 123], and (4) targeted interpersonal harms [45, 89, 95, 156]. Researchers have defined *technical harms* as "concerns about account integrity" [123] which include phishing scams [77] and hacking [129]. *Privacy and surveillance harms* include any "abuse of information entrusted to organizations" [24]. Privacy-related concerns include identification, secondary use, exposure, increased accessibility, blackmail, appropriation, and many more [139]. *Content-based harms* are those "caused to individuals who view undesirable content on a social media platform" [123]. Spread of misinformation [51, 109], and content related to violence, self-harm, and/or suicide are all examples of content-based harms that can create negative impacts for a larger audience [43, 122]. For example, suicide-related content might include suicidal ideation [76], or visuals [21] that could be harmful to those exposed to the content [22]. Incitement to violence, "when a speaker causes others to engage in violence or illegal acts" [137] is another example of this type of harm that is discussed in the space of hate speech literature [46]. *Targeted interpersonal harms* include any attack towards specific individuals in online spaces. Examples include hate speech [45, 89, 95, 108], bullying [156], and harassment [80, 107, 153, 158]. Interpersonal harms have been increasingly targeting specific racial or gender identities [64, 88, 102, 124].

Our work builds upon this literature by identifying different harms that occur in ephemeral spaces. Given the rising popularity of these spaces, this work identifies new harms and examines previously identified harms (in non-ephemeral contexts) and their unique nature in ephemeral spaces as well as the challenges they introduce due to ephemerality and platform design.

These types of risks, attacks, and malicious content continue to persist online and create potential harm for communities unless they are mitigated through content moderation strategies. As a result, we now explore current moderation practices in traditional social media and if these strategies apply to new forms of social interaction.

2.2 Content Moderation

Content moderation is the "practice of screening user-generated content posted to Internet sites, social media, and other online outlets, in order to determine the appropriateness [...] for a given site, locality, or jurisdiction" [118] which is conducted by all online platforms to some extent [60]. Moderation can occur through human review [23], algorithmic decision making [25, 61, 62], or a combination of the two [78, 86]. Any piece of content that is uploaded online might be flagged by humans or automated flagging systems. If flagging occurs automatically, it could be done before the content is posted. In these cases, the user could be informed and prompted to review the content for violations of policies. Automatic detection systems can also take down the content directly or redirect it to human moderators for further review. If the post is seen and reported by other users, it could be sent to human moderators or available algorithms for a more thorough review. However, flagging systems can be flawed if toxic behaviors are normalized [28, 75].

If it is determined that a piece of content violates platform policies or community guidelines, actions with regard to the content could be taken. These actions could include any of the following techniques: excluding, pricing, organizing, or norm-setting [63], some of which would allow the extension of responses to account-level ones depending on the severity of the infraction (e.g., adopting an account striking policy). Our findings reveal that a large portion of these responses has not been implemented in VR moderation toolkits.

Because there is potential for mistakes in the content moderation process (e.g. inconsistencies and unfairness of decisions [36, 60, 94, 119, 128, 142] that could particularly impact marginalized communities [26, 32, 52]), platforms might also make use of an appeals process. Appeals could also be used as signals for false automatic removal, where content is only reviewed by human moderators if the user contests the moderation decision. Thus, designing for contestability (i.e., "mechanisms for users to understand, construct, shape and challenge model predictions" [74]) is another focus in the literature [142]. Social VR platforms have currently implemented appeal procedures through the submission of tickets on their websites and offer no options to do so within the VR platforms [10].

While these processes for content moderation are well established and studied, they rely on the existence of clear and persistent logs of activity online. Additionally, the majority of existing practices address harmful content after they have been posted. Our work will examine these frameworks in the context of ephemeral content identifying challenges and limitations. Moreover, the research community has historically critiqued the fact that moderation only included punitive responses [103, 143]. Over the years studies have found different practices utilized by moderators that extend existing punitive measures. Grimmelmann classified moderation into ex-ante (i.e., allowing or prohibiting certain behavior before anything is posted) and ex-post (i.e., punitive responses to posted content) [63]. Seering et. al. also identified non-punitive views moderators have on their role, by asking moderators about their social roles [127]. While the paper does not focus on functional roles, they find facilitation as one of the ways moderators view their role [127]. In this study, we combine these existing structures and explore VR moderation responses in the framework of (1) preemptive (an extension of ex-ante), (2) facilitatory, and (3) punitive moderation responses.

Content moderation is performed by commercial content moderators (CCM) or volunteers. As CCMs are in charge of platform-level governance, they are often subject to disturbing material such as witnessing beheadings or child abuse which could result in harm to the moderators [116, 117, 132, 133]. On the other hand, volunteer moderators are oftentimes community members who want

to fulfill certain roles within the online platform. An analysis of how volunteer moderators view their roles on platforms with non-ephemeral content was conducted by [127]. They report five high-level categories where metaphors for moderation fit into, namely: (1) Nurturing and supporting communities, (2) overseeing and facilitating communities, (3) governing and regulating communities, (4) managing communities, and (5) fighting for communities. Further research has found that volunteer moderators on video streaming platforms often collaborate with one another on tasks such as preparing before live streams and working together in real-time during streams [37]. Over time, these volunteer moderators end up contributing to community guidelines and social norms in the space. Our findings reveal that due to the ephemeral nature of content in social VR and the need for immediate responses, the collaboration that was previously observed in non-ephemeral spaces is much more challenging to implement on these platforms.

2.3 Social Interaction on Ephemeral Social Media Platforms

Ephemeral social media content refers to "any communication that is explicitly designed to disappear after a short, finite period" [27]. Instagram stories are examples of ephemeral content that have been studied in prior work. In this study, however, we consider ephemerality to refer to any communication that is not recorded and can not be retrieved at any time after execution (e.g., voice or movement in VR). In this section, we provide a brief overview of studies of ephemeral content, including stories, voice, and VR.

Stories. Instagram and Snapchat are among the social networks that allow users to share content with a 24-hour lifespan. The media shared through stories is often visual accompanied by brief textual or sticker elements [48, 149]. The short lifetime of these material encourages users to share minimally edited experiences [146] or out-of-the-ordinary moments without concern for not maintaining a consistent self-presentation [27, 98, 140]. Different elements of storytelling utilized in stories [19], as well as their use for journalism and political expression, have been explored in the literature [79, 145].

Voice-based. Voice-based social media platforms have become increasingly popular in recent years [13]. As studies have shown that lonely individuals prefer voice communications over text-based ones [115], the rise of platforms such as Clubhouse and Discord can be partially attributed to the need for connectivity and intimacy during the COVID-19 pandemic [15]. Discord, released in 2015 [4], is one of the older platforms offering voice communications. Initially, it was used by those in the esports and gaming communities [5], but has since grown to be used for other purposes such as educational classes [148, 155]. Clubhouse is a newer addition to voice-based social media platforms. Released in 2020 [2], the platform reached immense popularity in 2021 with 10 million weekly active users in February 2021 [3]. A number of studies looked into the popularity of Clubhouse and how it is being used by people, especially during the COVID-19 pandemic [68, 92, 160]. Through surveys of users on Clubhouse, Zhu reported that this application "breaks the boundaries between the private and the public domain and creates a small society" [160]. In another study consisting of interviews with 26 regular users of the platform, [72], Jung found that the ephemerality of voice, and the interactivity and intimacy it brings, can help establish social relationships. They also give suggestions for ways to increase privacy controls and methods for giving users more ways to contribute other than chat and voice [72].

Similar to social VR, the ephemerality of new communication methods offered by voice-based platforms creates a need for re-examination of the possible harmful behavior and applicable moderation strategies in these spaces. Closely related to our work is the study of moderation of voice-based communications on Discord [70]. Jiang et. al. reported a number of harmful behaviors that are introduced in these spaces that are unique to the use of voice communications. Some of the harmful behaviors discussed in [70] are: (a) slurs and hate speech, (b) disruptive noise, (c) music queue disruptions, and (d) raids.

VR and Gaming. Social VR refers to "3D virtual social spaces where multiple users can interact with one another through VR head-mounted displays and engage in 360-degree immersive content" [55, 99]. These platforms are described as mediums for authentic social experiences [54]. Affordances of social VR platforms (e.g., non-verbal interactions) for interpersonal relationships could be influencing these perceptions [54, 90, 91]. Feelings of authenticity, however, are accompanied by skepticism and privacy concerns [138]. In addition to VR technology, world design within VR games also plays a major role in interpersonal relationship building. Digital proxemics is the paradigm of how space is used in virtual environments and how the presence of others in our virtual space can affect our behavior [152]. Research into digital proxemics has found that the introduction of social signals, personal space, and background noise in virtual environments affects how players behave in social VR. Williamson et. al. also found that considering the size of the world, smaller spaces tend to foster cohesive groups more easily compared to larger groups. However, larger groups allowed for more personal space than smaller groups [151]. This knowledge of how user behavior is affected by world design can help us identify and examine the aspects that afford harmful behaviors.

A few studies have discussed the need for VR/gaming technology to accurately reflect real-world experience. In one study of VR dancers, interviewees mentioned that engaging in social VR platforms has generated feelings of freedom, community, and genuine connection [112]. However, many in the VR dance community find issues with the lack of avatar customization that allows users to adjust features like running speed, jumping height, or flexibility [110, 112]. Without features that can mimic the way we perform ourselves in the real world, users lose the ability to express themselves in VR and user agency is lost [71], which can ultimately affect how these communities and interpersonal relationships form in VR spaces. There are recent studies innovating new interactions to make the VR world more real [29, 100, 125, 141, 159]. User experimentation with and perceptions of virtual body representation have also been explored in prior work [55, 56, 58, 84].

Finally, prior work has explored harassment in the context of social virtual reality [31, 57] and online games [69]. These studies explore harm from the perspective of users and investigate how they experience and are affected by harm in this immersive environment. Our work provides a different perspective by focusing on moderators' views of harm and the strategies that are used by them.

Our work also enables the formation of a full view of harms in social VR by pairing the perspective of users (offered by prior work), with first-hand experiences (through our virtual ethnography), and the moderator's perspective (through interviews with moderators). Additionally, while all previous work focuses on harassment, we explore a broader view of online harm.

3 STUDY 1: VIRTUAL FIELD RESEARCH

Given the rise of social VR and the potential challenges of moderating behavior in ephemeral social spaces, our virtual field research explores the types of potential harm being done and how they differ from non-ephemeral spaces. The goal of this study is to observe and document novel forms of harm and to investigate moderation practices that have emerged in response.

3.1 Observed Platforms with Ephemeral Interactions

In this section, we will provide a description of each platform, how they work, and the powers available to moderators. We chose to observe some of the most popular and frequently visited platforms for social VR: (1) AltspaceVR, (2) Horizon Worlds, and (3) Rec Room. While most VR platforms are only accessible through special VR headsets, AltspaceVR provides access through desktop machines as well. Due to this need for expensive equipment, VR platforms are less heavily used compared to other social media platforms, but the three selected were among the most popular. As of February 2022, Horizon Worlds reached 300,000 monthly users [8] and 10,000 created worlds [12]. Investigating three VR platforms also allowed us to study social spaces that cater to different age groups and demographics, where Horizon Worlds and Rec Room skew younger than AltspaceVR.

AltspaceVR. AltspaceVR is a VR-based social space owned by Microsoft available to users worldwide over the age of 13. There are two types of gatherings on AltspaceVR: events and worlds. Worlds are environments built either by official developers or users. Public worlds are accessible at any time of day; neither owners nor moderators need to be present. Users joining a world see a prompt that the space is unmoderated. Events are time-bound gatherings that are often organized with a specific goal in mind. Events are usually one hour long and have hosts present for the duration of the event. Events can be held in any user-designed or publicly available world.

Users can see a list of scheduled events and a description that includes a list of rules and general goals of the event, although posting rules is not a requirement. Rules (or general messages) can also be shown as pop-ups when the user joins the space, or added as virtual objects near the event space. Users can interact with each other through verbal communication, text-based messages, or emojis. The default setting of all spaces is spatial audio, where users can only hear the people nearby. Muting others prevents a user from hearing audio from them while blocking removes both audio and avatar from view. However, since blocking does not remove a user from the space, the blocked user can still manipulate objects and interact with other users in the shared space. AltspaceVR employs commercial content moderators, however, their presence is not guaranteed. Many events are moderated by individuals not affiliated with the platform. Moderators have the following capabilities [1]: muting users, messaging individuals in the space, and removing users from an event temporarily or permanently. Hosts have additional capabilities and can also: amplify voices, mute all, message all, and toggle stage blockers (i.e., mechanisms that restrict access to specific sections of an environment unless users are given explicit permissions).

Horizon Worlds. Horizon Worlds is another social gaming VR platform owned by Meta. The minimum age of users is 13 [11]. Horizon World shares many features with AltspaceVR, including events and worlds and many user actions. However, there are a few important differences. First, Horizon Worlds allows mature content, including content that is sexual in nature or that depicts regulated substances, as long as the world is tagged as mature. Second, users have more available actions, including the ability to create opinion polls to remove users. Polls are designed as a self-moderation strategy; if there are no moderators present, users can remove a user from a space if the majority of the users agree that they did something harmful. As in AltspaceVR, moderators can mute, give warnings to, or remove users from the space. However, unlike AltspaceVR, muted users cannot unmute themselves and removed users cannot come back.

Rec Room. Rec Room (owned by Rec Room Inc.) is a social hangout where users can create or explore spaces alongside others. Unlike other VR platforms, however, Rec Room does not provide minimum age limits. For children younger than 13, Rec Room requires the creation of junior accounts [14]. With the goal of protecting children, junior accounts are linked to and moderated by a parent or guardian account. They also can't send or receive private messages or verbally communicate with other avatars [9]. Much of the features of Rec Room are similar to those discussed for AltspaceVR and Horizon Worlds. There are small differences, however, for instance in this platform, users can be blocked or friended using menu options and hand gestures. Moderators can remove users from spaces, in which case they would be banned from the space and can not return.

3.2 Method

To understand new potential forms of harm enabled by VR platforms, a virtual ethnography [33] of each of the selected platforms was conducted. A total of 100 events were observed between June and August 2022 across AltspaceVR, Horizon Worlds, and Rec Room. Table 1 shows a summary of observations. The study was approved by our institution's IRB.

3.2.1 Selecting Events. Observations were conducted only in public spaces (no private events or worlds). Even though each platform offered thousands of worlds, few users were present in each space at any given time. In order to observe a critical mass of users, we focused our observations on scheduled events. The promotion and limited time frame of events made them more likely to attract larger groups of individuals. A list of the type of events that were observed is included in the supplementary material. AltspaceVR was also observed more often, as it allowed access via a desktop machine as well as a headset. Our research team selected events with high participant counts and unique topics (i.e., among the most highly attended options at the time of observation, we would select spaces

Platform	Observed Events	Events With Harmful Behaviors	Harmful behaviors (moderator addresses problem)	Harmful behaviors (no moderator)	Harmful behaviors (moderator present but no action)
AltspaceVR	45	9	8	1	0
Horizon Worlds	25	12	2	8	2
Rec Room	30	24	4	13	7
Total	**100**	**45** (45%)	**14**	**22** (49%)	**9**

Table 1: Observations performed for each platform. We observed 100 total events. In 45 events, one or more harmful behaviors occurred (the number of instances of harmful behavior is reported in Table 2). 49% of all events where a harmful behavior occurred had no moderators present.

that were least similar to previously attended ones). Some events were revisited throughout our observation period.

3.2.2 Observations. Observations were conducted individually by three members of the research team (conducting 52, 32, and 16 observations each). During each observation, the researcher would immerse themselves in order to watch, listen, and move around the space. If approached during the observation, the researchers would interact with users, however, they made every attempt not to get involved while harmful behaviors were taking place. The researchers paid particular attention to moderator actions and to any user behaviors that could potentially do harm, deliberate or not. VR events were recorded, using the functionality provided by the headset or video recording applications. After attending each event, the researcher would take notes on what they observed including any harmful behavior or moderation activities.

3.2.3 Analysis. Once the observations were concluded, one member of the research team reviewed all notes and extracted all mentioned harmful behaviors. Behaviors were analyzed at the level of individual interactions. The listed behaviors would then be labeled using an inductive and iterative open coding approach [97]. After the labels were identified, we performed multiple rounds of grouping, trying to identify the underlying reason or capability that enabled these forms of harmful behaviors.

3.3 Findings

Table 1 provides high-level statistics on the presence of harmful behaviors and moderators. Harmful behaviors were observed in 45% of the attended events. Moderators were only present in 51% of the events where harmful behaviors were observed. However, even in events where moderators were present, not all harms were addressed. This could have been due to limitations in sight and hearing due to spatial constraints or preoccupation with other concurrent actions by other users. We will begin the discussion of the results of this study by exploring the types of harm we were able to observe through our virtual ethnography. Next, we will discuss observed moderation practices.

3.3.1 Types of Potential Harm Observed in Social VR.
In this section, we report harms that have novel forms in social VR and connect them to previous forms identified on social media, discussing how VR could exacerbate specific aspects of the harms. We group observed harms into those enabled by (a) voice, (b) virtual embodiment, and (c) platform affordances. Table 2 documents the 13 new types of harmful behaviors and the number of times each harm has been observed. As multiple harms can take place concurrently, the total instances of observed harm are larger than the number of events in which harms took place. We can see that only 24% of observed harms were moderated. This was often because moderators were not present in the spaces where the harms were taking place or they were not spatially present (e.g., not able to hear or see the harm due to limitations in sight or spatial audio). We describe the 13 behaviors, separated based on the factor that enables them, and ordered based on observation frequency within each group. Observational anecdotes will be identified using *('O' + Observation Number, Platform Name)*.

(a) Harms Enabled by Voice. The ability to communicate verbally, is not unique to social VR. Voice-based social media platforms (e.g., Discord and Clubhouse) use speech as their main form of communication as well. However, the addition of spatial audio, and the inability to hear every conversation taking place within a space, can exacerbate the harms introduced by the addition of voice. We discuss these challenges in the context of two observed harms.
(a.1) Disruptive Noise. The ability to use voice to interact with others within VR spaces, can introduce harm in the form of interrupting, yelling, or making any sounds that could irritate others. Jiang et. al. previously identified this harm in the context of a voice-based platform (Discord) [70]. While the general concept of "disruptive noise" as a harm is shared between social voice and VR platforms, VR could introduce additional challenges, specifically for participant identification. Social voice platforms such as Discord offer one audio stream that is shared among all participating users within the voice channel. As a result, anything that is said can be heard by all users and moderators, allowing for moderation to take place if need be. In VR however, due to the use of spatial audio in the majority of worlds, a user's voice can only be heard by those in their immediate spatial vicinity. This design choice can have both positive and negative consequences. On the positive side, it would be much harder for a single user to disrupt everyone else's experience since they would not be heard by everyone unless given voice-amplification access by moderators. On the negative side, this can stand in the way of moderation. Since moderators would not be able to hear all users, offenders could be left undetected if they stay away from moderators and/or stop making noises when a moderator approaches.

	Harmful Behaviors Observed in VR	Instances	Moderated	Prior Work
Enabled by Voice				
	Disruptive noise	17	6	[70]
	Harmful or harassing language	10	1	[70, 95]
Enabled by Virtual Embodiment				
	Disruptive movement	9	5	[31, 57]
	Enactment of physical harm	9	0	[154]
	Group acts of bullying	3	0	[38, 93]
	Not following social norms	3	2	[147]
	Enactment of sexual assault	2	0	[83]
	Following and stalking avatars	1	0	[18, 113, 135, 136]
	Enactment of self-harm or suicide	1	0	[81]
	Enactment of violence due to incitement	1	0	[82, 121]
Enabled by Platform Affordances				
	Misuse of moderation power	4	N/A	[157]
	Misuse of platform features	3	1	[123]
	Pornographic content	2	0	[39]
Total		**65**	**15** (24%)	

Table 2: The number of times specific harmful behaviors were observed in VR. The total number of incidents is larger than the number of events that had harmful behavior because one event could have multiple occurrences of harmful behavior. Prior work refers to studies that have identified similar harms in other (often non-ephemeral) social media platforms.

Given the current moderation tools, muting or removing offending parties (with or without warning) was the most common way of dealing with the harm *(O[7, 30], AltspaceVR)*. In *(O7, AltspaceVR)* the background noise coming from a user's microphone was disrupting the conversation. The user was muted by the moderator multiple times, unmuting themself each time. They were eventually warned that they would be removed if they continued to unmute themself, and since they proceeded to do so, they were removed from the space. Muting all users upon entering a space, was another method used by moderators to prevent this type of harm *(O[16, 19, 32, 34], AltspaceVR)*. This method was often used in interview or talent show settings where minimal interaction with or between the audience was expected. In these environments, users would only be able to speak if given permission by the moderator. *(O79, Rec Room)* was one event in which users were muted by default for the duration of the show. After the show was over, the moderator exited the event, while participants were able to remain in the space. With the moderator no longer present, the mute-all was also deactivated. This resulted in an extremely loud room where people were observed making random noises and repeatedly saying the N-word. Additionally, the small size of the event space compounded the problem, by causing everyone to be able to hear one another while they were all being very loud. Some hosts used this potential for chaos as part of their event. At the end of an interview in *(O82, Rec Room)*, the moderator gave an "ear rape warning" before lifting the mute all, resulting in everyone talking at once, creating an extremely loud experience.

(a.2) Harmful or Harassing Language. Hate speech and slurs as types of harmful language have been studied in both text [95] and voice-based [70] social media platforms. These types of language can make communities feel unsafe [53]. Social VR platforms are not an exception to this type of harm. While these platforms offer peer-to-peer messaging as a feature, the messages are private and the difficulty in writing them (as you have to write one character at a time by pointing your hand controller at keys on a keyboard shown to you in the space) makes them less likely to be used. Moreover, while Rec Room does offer a shared group messaging feature, allowing all users within a world to communicate using text, we only observed the feature being used in one observation. Consequently, voice is likely to be the medium used by perpetrators of harmful language. The challenges are then similar to those expressed in *(a.1) disruptive noise*, with visibility and moderation challenges being the major concerns. However, while third-party moderation could be difficult, [57] finds that users take advantage of blocking or personal space bubbles (i.e., a feature that removes a user from one's experience by making them invisible and not allowing them to be heard if they get too close) to prevent harassers from delivering discriminatory comments. The impact of such speech, however, has not been extensively studied. While [70] argue that these instances are more offensive when spoken in smaller groups, little is known about how users' virtual self-presentations in VR can influence the effects of this harm on them.

(b) Harms Enabled by Virtual Embodiment. Virtual bodies are new to the social media experience. In social VR, avatar bodies can emulate movement, touch, and interaction with items in the environment (e.g., picking up or throwing objects). These new capabilities create novel social interactions, but they also raise new possibilities for enabling harm between users. Below, we will discuss the eight types of potential harm we were able to identify that make use of this new form of interaction.

(b.1) Disruptive Movement. Disruptive movements are the most innocuous form of harm arising from embodiment in VR. Users were observed running between, around, or through other avatars having a conversation and disrupting their conversation *(O90, Rec Room)*. Running towards the stage area in performance-based events could also disrupt the flow of the events *(O[44, 11, 26], AltspaceVR)*. In *(O45, AltspaceVR)*, where an interview was taking place, a user attempted to get up on the stage multiple times, being re-spawned[1] when hitting the invisible stage blockers[2]. While the user was not able to gain access to the stage, the movements were disruptive and the intruding user was eventually removed. Some disruptive movements could be unintentional. For instance, taking one's headset off when in an environment could make your avatar's body go limp or float up *(O11, AltspaceVR)*. These movements could also be disruptive as they might block other users' views or distract the performers, however, they are not controllable by the user. In most cases, disruptive movements were moderated by the removal of users from the space. In *(O26, AltspaceVR)* however, when a user was constantly moving around in the circle of people who were sharing their experiences, the moderators engaged the user in a discussion, asking them to stop their continual movement. Freeman et. al. [57] report that users believe these types of disruptive movements to be specifically harmful as they disrupt the social atmosphere and ruin others' VR experience. Blackwell et. al. however find that users had different levels of tolerance towards such behavior with some finding it "harmless and somewhat entertaining" [31].

(b.2) Enactment of Physical Harm. We observed — and even experienced — physical harm multiple times, like being slapped *(O[89, 92], Rec Room)*, often for not responding to someone's comment or question. For example, we observed one avatar being slapped because they were having a conversation with another user which the offender was not a part of *(O90, Rec Room)*. The ability to interact with objects can also amplify this issue. For instance, we saw audience members in a performance throw bottles at the performers on stage because they believed their performance was not good *(O89, Rec Room)*. We also saw this kind of physical abuse targeted at children. Figure 1a shows one observation where the adult-looking and sounding individual was holding a stick saying they were going to "beat some kids up". A couple of moments later we also heard a child-sounding avatar saying "stop hitting her" *(O58, Horizon Worlds)*.

(b.3) Group Acts of Bullying Exacerbated by Physical Presence. We observed several instances of users using their physical presence to engage in group acts of bullying. Groups of users would circle around a target user and proceed to hit them *(O83, Rec Room)*, make fun of their avatar's appearance and clothing *(O93, Rec Room)*, or chant phrases such as "loser" *(O55, Horizon Worlds)*. In the latter case, we observed the target leaving the space crying. It is important to consider that because you view and experience the world from the eyes of your avatar (and not, for instance, from above), users circling around your avatar, completely blocks your view and renders you unable to see anything else other than their avatars. Consequently, this can feel not only like an invasion of space but also a loss of orientation to all space and inability to escape.

(b.4) Not Following Social Norms. All social media platforms have behavioral norms that are often not written down but still expected to be followed. These standards for behavior can be built around specific platform features (e.g., likes [147]) or a transformed version of real-world conventions. Not blocking the doorway/entrance of spaces, which is an expectation in most real-world settings, is seen in social VR spaces as well. In *(O4, AltspaceVR)*, we received a private message from a moderator asking us to move from the spawn point (the point at which users appear when joining a world). It is worth noting that standing in the spawn point does not prevent other users from joining the spaces, as walking through other avatars was implemented into these platforms in order to prevent users from getting stuck in corners [6, 7].

Gendered spaces are another real-world convention that was observed in VR. In *(O71, Rec Room)* we witnessed a heated argument between a group of users over a female-presenting avatar using the men's bathroom. Since avatars are not able to get fully nude or perform a simulated version of using a bathroom in the studied VR platforms, the reason behind the existence of such gendered spaces, as well as the strong emotions around the "misuse" are worth further analysis.

(b.5) Enactment of Sexual Assault. We observed hand and face movements being used to simulate sexual acts. Sexual acts can be performed to the offender's avatar (e.g., enactment of sexual self-gratification) or to the target's avatar. An example of a sexual act being performed on another user that our team observed, was an avatar kissing another avatar without consent *((O97, Rec Room) and (O64, Horizon Worlds)*. The user would bring their avatar's face and hands close to the other avatar's face, pretending to kiss them while making kissing noises. Personal space bubbles were added to VR platforms to prevent these types of incidences and ensure users have personal space [16]. However, this setting only removes a user from your view once they get too close and does not block them from getting closer[3]. As a result, while the user does not see any harmful behavior, the aggressor can still do things to the avatar's body. And the user would not know what is going on because the bubble prevents the user from seeing them. In a survey of 600+ regular VR users, Outlaw found that 49% of women had experienced at least one instance of sexual harassment [105]. Users have expressed finding such experiences in VR "jarring and uncomfortable" [57].

(b.6) Following and Stalking Avatars. The ability to move can also allow avatars to virtually follow another avatar throughout a space. Our team observed one scenario where a user was screaming at another user asking them to stop following them around different worlds *(O90, Rec Room)*. In this case, the offender was using their avatar's virtual body to follow the target in a home space where a party was being held, following the user to different floors in the house. While continuous unwanted interactions can occur in non-ephemeral social media platforms, blocking the users could solve the problem by fully removing them from the user's experience

[1] Placed back in the spot avatars are placed when they join a space.
[2] Stage blockers enable moderators to control access to specific sections of the environment, only allowing those with adequate access through.
[3] Meta explained their reasoning behind this design choice as follows: "avatars will still be able to move past each other, so users won't get trapped in a corner or doorway" [6, 7]

[40, 65]. In Rec Room, however, blocking would make the blocker and the blocked transparent and mute to the another. This new state prevents users from performing continuing obscene gestures or verbal harassment but could potentially still allow the user to follow as they are able to see a transparent outline of the avatar. Other VR platforms (e.g., AltspaceVR) hide the avatars from one another thus circumventing attempts to follow, but still allow both users to manipulate the environment, possibly redirecting the user towards other means of conducting harm.

(b.7) Enactment of Self-harm or Suicide. As currently designed, users can walk or jump off objects. This ability enables them to simulate self-harm or suicide scenarios. At one virtual event *(O94, Rec Room)* a user threatened to kill themselves during a conversation. They then proceeded to throw their avatar from the second floor of a virtual two-story building, pretending to be dead afterward. Self-harm and suicide-related (which we will refer to as self-harm content) content on post-based social media platforms have been studied in prior work [30, 42, 43, 111]. This content can be in the forms of suicidal ideation, suicide attempts, and non-suicidal self-harm behavior [122]. Expressions of self-harm content range from seeking or providing support, and memorials, to flippant references to suicide or other matters [35, 47]. Users posting non-suicidal self-harm images on Instagram expressed social needs ("connecting, disclosure, communicating") as their main motivation [34]. In the context of our observation *(O94, Rec Room)*, the user appeared to be viewing their act as humorous and as a method of engagement with others in the space as they performed the portrayal multiple times, returning to the conversation taking place around them after every jump. Thus further analysis is required to understand the real intentions and consequences of performing self-harm in VR spaces. Detection of self-harm content in ephemeral spaces would likely be quite challenging as the issue is still unsolved in post-based platforms with Morena et. al. finding that Instagram, for instance, is only able to provide content advisory warnings for one-third of such posts [101].

(b.8) Enactment of Violence Due to Incitement in Realistic Settings. Incitement to violence is a recognized issue in social media platforms [85], where users write messages to incite others to perform acts of violence. VR has the potential to allow users to play-act these scenarios out. For instance, we observed an event (shown in Figure 1b) where female-presenting avatars were invited by the host to take their revenge on men *(O59, Horizon Worlds)*. The host announced, "We need to take our revenge. How you go about that is not for me to say, but I think you can come to that conclusion yourselves...". At the same time, the lid of a chest – containing a number of guns – opened with the message "KILL ALL MEN" written inside (see Figure 1b). The women then proceeded to shoot the men, as well as the other female avatars that were not participating in the shooting. Even though these scenarios are not real, they could potentially be more convincing than traditional social media posts in incitement to violence. On the other hand, they might help reduce the frustration of the user by taking it out on avatars that are not real. Video games, which are one of the most similar mediums to social VR, can include similar depictions of violence. While the link between violence in video games and aggression has been studied, the results are conflicting [50, 73, 130] and don't account for other aspects of video games (rather than the violent content) potentially influencing aggression [17]. Thus further research is needed to understand these types of activity and their impacts.

(c) Harms Enabled by Platform Affordances. How platforms are designed and the tools they provide for users can enable new communication methods, but can also enable new harms. We discuss three types of harm enabled by the capabilities of users within social VR platforms.

(c.1) Misuse of Moderation Power. Moderators removing or muting users even when they have not violated any rules is one of the harms we observed. For instance, in a talent show, moderators would remove users they did not enjoy the performance of, commenting on their bad performance after they were removed from the space *(O89, Rec Room)*. While abuse of moderation power is an issue in post-based social media platforms [157], the fact that content in social VR is ephemeral could make the control of this misuse of power harder as users are not able to provide any proof of wrongful moderation to the platform. Studies of this type of abuse on other platforms have found that it can result in a decline in participation by community members [157].

(c.2) Misuse of Platform Features. Horizon Worlds and Rec Room both offer a functionality where users are able to nominate a user to be removed from a space (removal polls). If this is done, all users in the space will see a message box pop up in front of them asking for their vote. An example of this pop-up is shown in Figure 1d. This functionality can be positive when it gives users the ability to remove disruptive or disrespectful users from a space, in the absence of moderators.

However, one major issue is that if a group of users were to team up, they could remove people from spaces for no valid reason as decisions are not reviewed by any outside parties before taking effect. For instance, during the unwanted kissing incident discussed in *(b.5) Enactment of Sexual Assault (O64, Horizon Worlds)*, another individual in the space was trying to help the target by pointing out the settings that would enable the use of personal space bubbles and reporting features. Since this user was helping, someone who supported the offender created a poll to remove the user[4]. We also observed another case where a group of users were successfully able to remove three individuals who were part of the organization team of an event *(O98, Rec Room)*. Thus, if there is a large enough group of individuals, unwarranted removal of people would be possible.

Another issue is that because worlds are often very large, not all participants are visible to an avatar (e.g., blocked by walls or objects). Additionally, spatial audio results in one's inability to hear avatars unless they are in one's immediate vicinity. As a result, some of the people who are asked to vote might not be aware of the harm that took place. While users are able to provide reasons as to why the poll was created (e.g. *"disrupting the experience" (O64, Horizon Worlds)*, *"inappropriate behavior" (O61, Horizon Worlds)*), we don't believe these short texts provide enough context for someone who was not present to make a decision. Moreover, as previously mentioned, they could be untrue.

[4]We don't know which user created the poll since removal polls don't display the name of the creator.

(c.3) Pornographic Content. This type of content is consumed across post-based platforms as well. However, while non-ephemeral social media enable moderators to see anything that is posted to a user's page and be able to take action if they violate the guidelines, this is no longer true in VR platforms. This is because users can employ different strategies to keep what they are doing hidden: (1) since VR worlds are often very large, users could move their avatar away from the main gathering of people to another spot in the world where they would not be observed. (2) Users could also move behind objects in a space to remain hidden. Or (3) they could hold the prohibited content in the form of small, disposable objects that would not be easily seen by others unless explicitly shown to them and could dispose of them if they sensed a moderator had joined the space. In *(O88, Rec Room)* we saw a group of avatars that sounded like children exchanging Polaroid images among themselves (see Figure 1c). They shared the Polaroids with us, which revealed they contained pornographic images. However, if they had decided not to show the images to us, it would have been much harder for us to find out what they were. Another example we observed was an avatar wearing a T-shirt with a pornographic image on it *(O94, Rec Room)*. Similar to the previous example, since users can change their avatar's clothing on the fly, they could change the T-shirt at any time and not be caught sharing such content.

3.3.2 *Observations of Emergent Moderation Strategies.*

During our virtual field research, we took note of moderation practices as well as observed potential harms. Building on existing moderation frameworks discussed in Section 2, we categorize these practices into (1) preemptive, (2) facilitatory, and (3) punitive. As moderation of spaces is often handled by volunteers, 41% of spaces are currently unmoderated. Moreover, in line with our categorization of moderation practices, we observe more actions conducted by moderators than purely punitive measures in response to harmful behavior. More concretely, moderation responses were observed in 39% of events, while moderation actions that were in response to harms (i.e., punitive) were only observed in 14% of events. Table 1 displays the proportion of events with occurrences of harmful behaviors in which moderators and moderation responses were observed.

(1) Preemptive. These are moderation actions that attempt to prevent harmful behavior. This prevention could be in the form of informing users what the rules are or by reducing users' capabilities in a space. In VR, moderators could inform users what the rules are by greeting users at the spawn point and telling them *(O12, AltspaceVR)*, writing the rules in the description of the event *(O11, AltspaceVR)*, showing the rules to users in the form of a pop-up in front of their avatar's face when they join an event, or writing them down on objects in the space *(O27, AltspaceVR)*. However, a large portion of observed events did not specify their rules anywhere.

Preemptive moderation could also be in the form of limiting what users are capable of doing to ensure the event would not be disrupted. For instance, moderators might mute all users upon entry *(O[16, 19, 32], AltspaceVR)* and allow participation only through hand-raising and the approval of hosts. In these events, users would not be able to unmute themselves. As observed in *(O82, Rec Room)* when the moderator decided to lift this mute all, everyone started screaming. Using "stage blockers" could also be a form of preemptive moderation that is seen in VR *(O[11, 44], AltspaceVR)*. This design tool allows users with the correct permissions to gain access to blocked areas (e.g., the stage)while everyone else would be thrown to the spawn point if they attempted to gain access to these restricted areas.

(2) Facilitatory. Moderators might also perform acts to facilitate the use of spaces and platform features by the attendees. Greeting people when they join, introducing what the event is, and what to expect in the event was one type of facilitation we observed *(O[18, 22], AltspaceVR)*. Moderators were also observed conversing with the attendees, and answering questions about how different controllers work and how certain actions can be performed in the platform *(O63, Horizon Worlds)*.

(3) Punitive. Punitive responses can include warning *(O10, AltspaceVR)*, muting *(O7, AltspaceVR)*, or removal *(O11, AltspaceVR)* of individuals in response to harmful behavior. Similar methods are used across all platforms. Warnings could be given through private messages *(O10, AltspaceVR)* or verbally *(O26, AltspaceVR)*. Punitive actions could be in response to a moderator observing something firsthand or having someone report an occurrence to them. For instance, in *(O91, Rec Room)* we observed a user being removed because another user reported they were being transphobic. While the moderator did remove the reported user, they did not ask for any evidence or what the transphobic behavior or language was before removing the user reported for the harm. This lack of proof is due to the lack of persistent records in ephemeral spaces.

We further elaborate on strategies utilized by moderators in order to address real-time interactions, the ephemerality of content, and specific features of the platforms based on interviews with moderators in Sections 4.2.1 and 4.2.2.

4 STUDY 2: MODERATOR INTERVIEWS

Despite the ephemerality of interactions and the limited abilities of moderators (e.g., narrow field of view and limited range for hearing), moderators have developed a set of practices for addressing potentially harmful behavior. Since these practices and tools might not have been captured through our observations, we conduct interviews with content moderators to hear the practices and challenges of moderation from their perspective.

4.1 Method

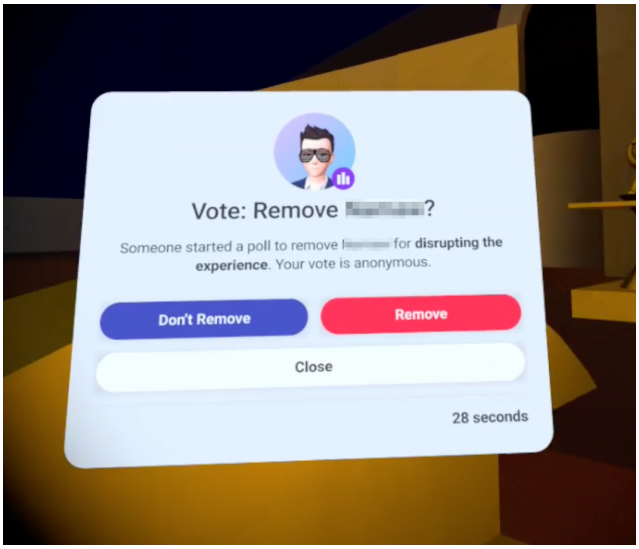
To learn about the experiences of content moderators in ephemeral platforms, we conducted semi-structured interviews with 11 moderators. Interviews were conducted in July–August 2022. The study was approved by our institution's IRB.

4.1.1 Recruitment. Messaging moderators directly in VR can be challenging. First, for many events moderators were not present. Second, moderators could not always be recognized. Third, messaging in VR imposes a very low character limit and includes restrictions based on who you're friends with. As a result, in addition to VR messages, we joined Discord servers dedicated to VR platforms. On these servers, we would post in channels dedicated to requesting help for research, as well as privately message people

(a) An adult avatar beating a child avatar with a stick (an example of *enactment of physical harm*).

(b) Cabinet with "KILL ALL MEN" message and loaded guns underneath it (an example of *enactment of violence due to incitement*).

(c) Polaroids containing pornographic images being shared with other avatars. The pornographic content has been blurred (an example of *pornographic content*).



(d) The removal poll interface is shown to users when asking to remove someone in Horizon Worlds. The creator of the poll selected "disrupting the experience" as the reason for requesting the removal of a user attempting to help a target of virtual harassment (an example of *misuse of platform features*).

Figure 1: Images of some of the harmful behaviors in VR. Usernames on top of the avatars are redacted to prevent user identification.

who had talked about having been VR moderators in messages posted on these servers. We contacted moderators on the following Discord servers: Virtual Reality, VR Club, Oculus community, Rec Room, Official AltspaceVR, Horizon Worlds, Educators in VR, and Students in VR. Unfortunately, we were only able to recruit moderators who were active in AltspaceVR. This could have a number of reasons, the most fundamental of which is that moderators in VR are scarce. These reasons may also include: 1) the reticence of official moderators (those affiliated with the platforms) to participate in interviews, 2) the newness of the platforms (e.g., Horizon Worlds was released in December 2021), and 3) moderators who are children. The last is particularly surprising, but we observed that in Rec Room most moderators sounded like children. These children would be ineligible to participate in our study (participants must be 18 or older).

	Age	Gender	Moderation Experience
P1	22	Female	2 months
P2	53	Female	8 months
P3	22	Genderfluid	4+ years
P4	40	Non-binary	4+ years
P5	60+	Female	4+ years
P6	52	Male	4+ years
P7	51	Female	1 year
P8	46	Male	2 years
P9	75	Male	10 months
P10	29	Male	1 year
P11	19	Genderfluid	4+ years

Table 3: Demographics of the moderators interviewed.

4.1.2 Participants. We conducted 11 interviews with current AltspaceVR moderators. Some of the moderators had experience moderating other social VR platforms, but all were most active in AltspaceVR. All were volunteer, rather than official, platform-employed, moderators. Table 3 shows the demographics of our interviewees as well as their experience levels.

4.1.3 Interviews. Participants were asked open-ended questions about their role, the potentially harmful behaviors they have witnessed, their moderation practices, and their thoughts on current and future moderation tools. The full protocol is included in the supplementary materials. Interviews were conducted through Zoom, recorded, and later transcribed for analysis. Participants were compensated for their time in accordance with the minimum wage of the researchers' location ($15/hour). Interviews took 45–120 minutes, with most interviews lasting around 75 minutes. The variance in length was largely accounted for by the depth of experience the user had in the platform (as a user or moderator) as well as the amount of exposure they had to harmful behavior.

4.1.4 Analysis. Using the interview transcripts, we review responses that address the research question we aim to answer, these correspond to any snippets discussing (1) description or perceptions of moderation roles, (2) methods or examples of performing moderation actions, (3) comments on world design or tools, and (4) discussions of evolving norms. As discussions were wide-ranging, these responses could be extracted from any portion of the interview and not necessarily map to specific questions asking about these issues.

After all relevant sections of interviews were identified, the lead researcher grouped related portions together. The research team then used affinity diagrams [87] and performed multiple rounds of clustering, identifying underlying characteristics and groupings within these responses. Among these clusters, we present those that speak to the differences in moderation practices, challenges, and design that are due to the presence of ephemeral content or the use of voice or VR technologies.

4.2 Findings

In this section, through analysis of interviews with moderators, we present strategies implemented by moderators to address the ephemerality of content and peculiarities in platform design choices. Next, the addition of content creation to the role of moderators is presented and compared with prior work on views on moderation roles [127]. We then elaborate on evolving perceptions of harm. Finally, we close this section by examining issues with current tools and the design of VR platforms, presenting the solutions offered by the participants.

4.2.1 Moderators Had to Be More Proactive for Real-time Interaction.

Having real-time interactions introduces disagreements and harms that would need to be moderated in real-time as well. As de-escalation in real-time can be very challenging, moderators made use of strategies that allow them to proactively facilitate conversations, rather than needing to reactively moderate them. VR moderators' recognition that embodiedness could make interactions very real could be another reason why they made an effort to prevent harm rather than respond to it.

> *In VR, it's literally designed to feel in-person. Sure, everyone's cartoon characters and either of you[5] could be anyone anywhere, but there's an inherent presence to it. So then it gets into how you feel about it IRL, more or less.* ***(P3)***

This understanding also resulted in some of the moderators having a "victim first" **(P3)** mentality[6] and having more severe punitive responses when harms did occur as they believed "victim first is what I find often works best for a welcoming environment" **(P3)**. We begin by describing the preemptive and facilitatory work undertaken by moderators that help reduce punitive responses. These strategies include: (a) informing users of the rules, (b) strategically positioning the moderator avatar to influence users' awareness of their presence, (c) identifying and helping those who might be struggling, and (d) keeping a closer eye on specific avatars.

(1) Informing Users of the Rules: Making sure users know what the rules are could help prevent infractions due to lack of knowledge. Rules could be written in the description of the event, as pop-ups, or on objects within the space. Participants were apprehensive about doing so, however. For instance, while they did inform whether events are for mature audiences in descriptions, they expressed "when we say 18+, you know, you would hope that would deter people, but sometimes it's more of an invitation" **(P10)**. Moreover, they often did not write rules within spaces because they believed "the users in VR, just don't typically take the time to stop and look at a plaque that you've made, that has your code of conduct on it" **(P10)**. Instead, if they want to cause harm they will "just blast past that and jump off the wall and re-spawn and do stupid stuff" **(P10)**.

(2) Strategically Positioning the Moderator Avatar to Influence Users' Awareness of Their Presence: The choice of where moderators would stand, and whether users are aware of their presence during events was also used as a preemptive measure by moderators. For instance, one participant expressed that if there is "something where there are audiences moving around a lot, and there's a lot of things going on" **(P5)**, they would want the moderators to be facing the audience

[5] Referring to the instigator and target of harm.
[6] Prioritising the victim through immediate response to the harm.

from the stage where attendees would be able to see them when they look onto the stage, similar to a "police presence" **(P5)**. In other cases, they might put moderators "in different places along the back, or on the sides" **(P5)**, not drawing much attention to their presence.

(3) Identifying and Helping Those Struggling: Ensuring users know how to use the VR technology and controllers, what to expect, and what is expected of them in events could also help prevent harms such as disruptive behavior. To do so, moderators would try to identify themselves to attendees (e.g., through banners with the image of the moderator's avatar, greeting users at the spawn point, using specific badges) and would then "constantly tell everyone else to be asking for help, if there is anything they need" **(P1)**. They would also approach users who seemed to be struggling or "standing alone and not talking to anyone" **(P1)** and try to help them.

(4) Keeping a Close Eye on Specific Avatars: Finally, moderators expressed that strangely dressed avatars or usernames would draw their attention as possible bad actors. While they expressed that "you can't judge them based on their avatar" **(P8)** and tried their best not to make their moderation decision based on these factors, they also mentioned removing users with strange usernames such as "give me food or die" **(P5)** or "f*ckface" **(P5)**).

4.2.2 *Moderators Adapted to Platform Design Choice.*

Ephemerality and limited sight and hearing are limitations that have been built into users' experiences in VR platforms. While these features can help mimic people's experiences of real-world interactions, they severely limit how moderators perceive and respond to events. Additionally, due to the lack of recorded logs within ephemeral spaces, finding evidence of harmful behavior after the fact is very difficult. The combination of these factors encourages moderators to try to increase their chances of observing anything that takes place in the spaces they moderate. However, limitations in what you can see or hear mean observing everything is not straightforward. To reach optimal visibility moderators made use of the following strategies: (a) open spaces without objects obscuring the view, (b) moderator positioning, (c) using objects to affect participant positioning, (d) teamwork, and (e) roaming around in the space.

(1) Open Spaces without Objects Obscuring View: Good world design that would support moderation was something all moderators strongly believed in, stating that "if you create a space, where people can, you know, affect other people without being watched, then you're part of the problem" **(P10)**. This often translates into the use of open space (i.e., a space with no objects such as walls or buildings blocking the view). One participant **(P3)** expressed that these open spaces tend to attract more loud ill-intentioned individuals as "they want to be observed doing their thing, whatever reaction they're looking for out of people" **(P3)**. This potential, however, "does not discourage from" **(P3)** using open spaces and it is still the go-to design because of the visibility it offers.

(2) Moderator Position: Whatever the overall design of the space, moderators tried to work with designers to build "some vantage point somewhere for moderators to be" **(P3)** from which point "it's easy to see what's going on, you know, maybe have to turn your head" **(P3)**. This type of visibility is often achieved through designing spots that "stand up above everybody else so that you're not, you know, blocked by other avatars" **(P10)**.

Since most VR spaces have spatial audio, this type of positioning for moderators could mean that while they might be able to see everyone, they would not be able to hear what users are saying (as users might be far away). As a result, moderators detect if someone is speaking or not by "seeing their talk bubble go up" **(P2)**. If they are speaking, moderators try to study the body language of people close to them to detect if a harmful act is taking place. In scenarios where they suspect that might be the case, they would either ask "someone that I know in the audience that's closer to the event" **(P2)**, "reach out to people that are around it" **(P2)**, or investigate it themself by "teleporting over to try and hear what they're saying" **(P2)**. These methods, however, could have a number of issues. Firstly, if multiple groups of people are talking at any given time, further investigations of the types mentioned above would not be possible as they could be very time-consuming. Secondly, while the body-tracking technology in VR is good, it might not reflect similar signals as one would have in real-life interactions, thus body language of an avatar might not always be an accurate measure of whether the person is feeling comfortable or not. Context is also important. For example, "you have to see when a person is actually friends with someone or not" **(P2)** as some language might be acceptable with friends but not with strangers. Because of these issues, participants suggested the addition of a tool that would allow them to "tune into a person that's far away" **(P2)** (e.g., "click on a person and actually hear what they're saying through their mic" **(P2)**) to help monitor speech more easily.

(3) Objects and Their Effects on Participant Positioning: Even if no objects blocked a moderator's field of vision, if users spread through the space unevenly, or huddle up in one location, this could still limit what can be seen as some users could be hidden by other avatars. Thus moderators needed to ensure people would spread in the space in a way that would give them the most access. Moderators expressed that if they leave the space empty, with no obvious objects users are expected to sit or stand on, users "started becoming wall huggers. They'd come into an event [...], and they'd line up against the walls" **(P5)**. Because of effects like this, they believed designing designated seating areas for users could help them spread in a way in the room that would make moderation easier.

> *the audience actually sticks to sitting in those chairs even though we're all just standing or sitting in our own chairs [...] So it's actually much easier to moderate the room with chairs because everyone's more evenly spread, you know, instead of just hugging any other surface around them* ***(P11)***

(4) Teamwork: Another strategy used to observe all concurrent activity in a space was the use of teamwork. The most basic use case of having a team of moderators is increased visibility such that "if I don't see something, chances are one of the other team members will" **(10)**. It could also mean that one moderator could

move and follow specific individuals in the space knowing others will be able to monitor other users. Consulting others when making decisions about appropriate moderation responses is another benefit of teamwork, allowing members to question "is that a little fishy or not? And we kind of make a collective decision" **(P8)**. The communication between team members could be through messages in VR, the party option on Oculus, or through a voice channel on a secondary application such as Discord **(P1, P5, P7, P8, P9)**. The choice of which depended on the team and what they found most comfortable. While some found being on a Discord call at the same time as being present in VR easy, others found it challenging. One participant **(P3)** also mentioned not using any specific method for communication. If something was so confusing that they would need to talk about it, they would do so in the VR space by moving to a side and briefly discussing the matter.

When a group of moderators is in charge of moderating an event, there could be scenarios with conflicting opinions on what the right course of action is. In these cases, one VR moderator believed the host had complete authority to make the decision, and "if you don't like it, don't do it. You know, don't moderate don't participate" **(P5)**. Another moderator (**P1**) also expressed that they were told they needed to ask permission from the lead moderator if they wanted to take specific moderation actions such as removing users from the space. Another participant, however, thought the best approach "is conversation and voting, unless it's like something that needs to be immediately addressed" **(P3)**. If something needed to be addressed immediately it would have to be a "very present and real and big threat" **(P3)**, in which case they would "remove this threat first" **(P3)** and then discuss how to move forward once things had settled down.

(5) Roaming Around in the Space: When world design fails and there is no team to rely on, moderators try to move in the space in a way that would allow them to get an idea of what is going on at any spot at any given time.

> *you can expect moderators to be roaming around and to just, you know, generally go where there are clumps of people and listen for things.* ***(P3)***

The moderator acknowledged that "Of course, some things may slip through the cracks" **(P3)**, as users might wait until the moderator leaves their vicinity before they perform the harmful action, "especially if people know where the mods are" **(P3)**. However, they believed at some point there is nothing more you can do and "if you miss something you can't really like, you know, fault yourself" **(P3)**. To help with visibility, one participant suggested the addition of a tool that would "enable a user to be seen through walls so that you know where they are" **(P3)**. This ability would allow moderators to see all users and not have to follow each avatar that leaves their sight. This would mean that moderators would only move their own avatar if they sensed there was a need for them to be there, and thus not lose their control over the main event space just because they followed one person. It could also potentially help with evidence gathering, for cases where harmful behavior is occurring behind an object.

4.2.3 Moderators Also Played Roles in Content Creation.

Event and world creation are a large component of social VR platforms. Since the monetization of VR content is still in its early stages, all interviewed moderators were volunteers and were often not paid for their work. As a result, at times they were not only moderators in the traditionally understood role of being in charge of the control or removal of harmful content, but they were also content creators, hosts, and entertainers. While prior work has analyzed different metaphors and perceptions of the social roles of moderators [127], "entertainer" or "content creator" is not among the descriptions of the role extracted from the participants' responses. VR moderators, however, discussed being in charge of the creation of physical or event content at times. For instance, if moderators are given world-building permissions they could act as "terraformers". In this role, they would be able to change the design of the space while the event is in progress (e.g., by adding or removing objects as needed): "If they need something in the room, I can put something in the room. If they need like an extra chair or something, I'd be able to do that" **(P2)**. As the design of physical spaces and the objects within them can affect the harms and moderation practices in VR spaces, having this role can help mitigate specific moderation challenges or influence behavior.

This mix of responsibilities brings up questions regarding ways to incentivize content generation while ensuring moderation is occurring correctly. It is important to ensure this incentivization does not come at the price of breaking platform rules. **P6** described an example of how this shared responsibility resulted in them allowing, and even planning to include, acts into their show that are against platform rules.

> [talking about an official platform moderator joining the event] *if they don't know how you moderate the event, they may end up kicking somebody you don't want kicked [...] I had something worked out with somebody when they were going to do something that was pre-arranged. And an admin came in and saw it and thought they were trolling and they kicked the person out.* **(P6)**

The potential for these types of conflict between the content generation and moderation roles is an important characteristic to keep in mind when thinking about how moderation happens in social VR spaces.

4.2.4 Moderators Explained Evolving Norms and Perceptions of Harm.

Moderators mentioned changes in perceived harm of certain behaviors over time. In most cases, participants shared examples of behaviors that had previously been viewed as harmless but grew to be seen as harmful over time. These were often focused on invasions of previously-unconsidered personal space. Two participants mentioned extreme movements in front of other avatars, where users would pass through other avatars or move back and forth repeatedly (**P1, P11**). As one participant stated, initially this jumping around "was thought to be a form of greeting, but it has become a way of, you know, harassing someone because you're constantly in someone's face" **(P1)**. Similarly, moving objects near another avatar – even though early perceptions, as with Burning Man wands that glowed and sparkled, were just "this person's having fun with the

item" (**P11**)– became increasingly frowned upon. As norms around personal space in VR evolved, those previously acceptable actions were considered inappropriate.

Finally, one participant mentioned that they were actively at work to change norms. Compared to offline interactions, **P4** found that ghosting — leaving a space without saying anything, even "leaving mid-sentence" — had become pervasive: it's "very, very common to have people just vanish out of relationships in a way you can't do in physical reality, you have to make up an excuse, you have to like, have an appointment, you have to have reasons to move away from each other in physical reality and that's not true in VR."**(P4)** This participant described a number of interventions they used in their moderation to deter ghosting and establish better community norms. These included a variety of more or less direct conversations with the ghosting user, where the moderator emphasized that these are "real normal human behaviors we do to check on each other and say you matter" **(P4)**, making the online world feel less anonymous and impersonal. But in addition to these efforts directed at the ghosting user, the moderator described a new collective behavior that enforces the desired leave-taking norms:

> *What will happen in the groups that I'm around is if somebody vanishes, everybody stops and waves at this person, even though they're gone. And that behavior actually creates the awareness that if somebody else in that group decided to leave, they can either be present for everybody waving at them, or they can leave and have everyone just feel that sadness and loss. (**P4**)*

This participant made it clear that they did not view ghosting as simply a rude but inconsequential behavior. Instead, they viewed ghosting as a signal of "neglect" and an "objectification of each other, we can use each other and vanish and not have to pay homage to [the fact that] we were building relationships there" **(P4)**. While the VR environment enabled this behavior (or occasionally even caused it, if a headset died), the moderator felt it was toxic and important to prevent this norm from taking root. Instead, through a variety of interventions, they sought to establish better norms.

4.2.5 Moderators Offered Suggestions for Moderation Abilities and Tools.

Finally, moderators were asked about their opinion of existing moderation tools, and suggestions for additions to the moderation toolkit that they believed would improve their experience. We present these suggestions in this section, dividing suggested tools based on the behavior they aim to address.

(1) Tried and True Design Issues: A number of moderators commented on required changes to platform designs and tools that have already been identified in the literature. These issues mostly revolve around the lack of access control. Moderators expressed issues with revoking moderation access when a moderator is misusing their power. The current design of AltspaceVR requires that this change be made from outside the event after which the world has to be restarted for the changes to take effect. This difficulty had caused issues for some of the participants (**P5, P10**). This is however not a new problem as the need to build access control into systems has long been identified [131].

(2) Create Nuanced Processes for Removing Participants: Removal of users from spaces is a central tool moderators make use of. In its current design, when a moderator tries to remove a user from a space, they are prompted to select the reason why. Multiple moderators expressed that this can be a problem because while they are filling in the information, trying to discern which options fit the behavior best, "the person is still doing the sh*t standing right there in front of you" **(P5)**. There were two suggestions for solving this issue. One was the addition of a "loading space" or limbo state the user would be taken to while the moderator filled in the information. They believed this type of space would allow the moderator to cancel the removal request if "you click on them by accident" **(P11)** and would mean that moderators "wouldn't have to worry about [the offender] continuing to do bad actions while you're trying to write up about the bad actions" **(P11)**. Another suggestion was the addition of specific objects for immediate removal of a user when the "person is just being a very immediate and very present threat" **(P3)**. These tools would no longer require the moderator to open their host panel to take action, "just point, laser, bam and then they're gone" **(P3)** thus saving time that could have allowed the user to continue the harmful behavior.

The reporting system described above for the removal of users sends the report to AltspaceVR. The information is however not visible in the user's profile for other moderators on the platform. One moderator believed having general statistics on previous harmful behaviors of the user could be helpful because "if they join any other events [...] other moderators also get to know that this person was creating chaos in some other events, so they can be aware of it" **(P1)**. Even though this could help ensure moderators keep a close eye on potential bad actors, it could easily bias moderators in new events against a user, resulting in them taking more severe moderation action on the user for a small infraction of the rules. It also wouldn't allow for growth or improvement in behavior on the user's part as they will always be branded with their past misbehavior.

Moderators also brought up a concern with another aspect of the user removal procedure. While AltspaceVR does have a host panel that allows moderators to remove users by clicking on their name on the panel, the moderators need to know what the user's name is to be able to use this functionality. However, "you can only get their display name from clicking on them" **(P11)**. This could be challenging if "someone is rapidly moving around because you can't figure out what their username is" **(P11)**.

(3) Provide More Tools to Control Speech and Movement: Muting of individual users for an extended period of time was one mentioned need. The current moderation tools offered by AltspaceVR allow for the muting of a single individual or the muting of all users in a space. When a single person is muted, they can immediately unmute themselves while muting all people would not allow them to unmute unless given specific permissions or the mute all is lifted. A number of moderators believed that it would be "a lot more helpful to be able to mute specific people permanently" **(P11)**, not necessarily for harmful speech, but because they don't understand why they were muted and "think it's like a system that's muting them not a person" **(P11)** so they unmute immediately. As a result, being able

to permanently mute them would give moderators enough time to "message them about why they needed to be muted" **(P11)**.

Moreover, participants commented on the lack of tools for the control of movements. As mentioned in Section 3.3.1 movements could be disruptive or harmful. However, the only way moderators are currently able to deal with these movements is to warn the users and then remove them from the space if they don't listen. However, one moderator believed that while some types of movement could be harmful, not all of them are bad enough to warrant completely removing the user from the space (**P10**). **P10** suggested the addition of a "rowdy crowd blocker" that would "restrain a person to a certain area for a certain amount of time" **(P10)** allowing the user to stay in the space, while also ensuring they are not bothering anyone else in the event.

> *they can be pushed back to a safe distance, so where they can, if they choose to, still be there and engaged in the show, but not necessarily affect the other people that are in there. **(P10)***

Providing new and more granular tools would support the more nuanced moderation our participants called for, and would expand a middle ground for users to learn evolving norms without being banished completely.

(4) Emphasize Education and Tolerance: One moderator touched on the need for a tool that would "educate and create reconciliation cycles" **(P4)** for moderators, specifically those who "are experiencing the belief that they had to cut somebody out of an environment for understanding the consequences of what that's like" **(P4)** since they believed actions such as kicking individuals from spaces "is violent and it trains people to be much more violent with each other" **(P4)**. They expressed there should be tools or platform design strategies that "reward less extreme moderation behavior and educate on the consequences of more extreme moderation behavior" **(P4)**. The simplest form of such a design would be "giving people a tally of like the average number of people kicked in an event" **(P4)**, allowing the user to compare themself with the average values and reflect on their own behavior. Another way would be to introduce a "gradient of interventions" **(P4)** and not just muting and kicking "but also some kind of education on how to intervene differently and how to debrief and how to educate" **(P4)**.

Another moderator suggested an alternative to completely removing the user. In this option, you would move the user to an empty space where no other user was present and no one would be able to see the user, but they would still be able to see the people who are "broadcasted" (e.g. hosts, and other performers in the space), "like a timeout room" **(P10)**. This way, the offender would not be able to do any harm to the users in the space, but can still enjoy the event if they wish to do so.

5 DISCUSSION

Our study provides an important first step in identifying harms and moderation challenges in social VR platforms. These challenges stem from real-time activity, wide territory, and the array of concurrent activities occurring in these spaces. Additionally, the lack of persistent records and the need for real-time observation and response to harm are major challenges for existing workflows of content moderation. Our observations indicate that these ephemeral social media platforms have potentially largely relied on volunteer moderators and community governance to address this need. For continuously accessible spaces (i.e., worlds) however, platforms have opted for a warning regarding entrance into unmoderated worlds and not forcing the presence of any moderators.

Even if a moderator is present, no avatar can observe everything that is happening in the space. This is due to the spatial and temporal constraints of being present at one place at any time. If there are several groups or individuals engaged in different activities, that would mean more work for moderators present in the space. When multiple moderators are present, spatial coordination will be required before or during the event to ensure physical coverage of a space's wide territory. The coordination and the response to harmful events will also need to be conducted in a seamless way that minimizes the disruption to other participants in the space. Concurrent activity is another issue in VR platforms. The spatial audio property in VR on top of the design of worlds as large spaces that can be explored introduce distinct interactions that would need to be moderated.

For the most part, VR platforms offer no commercial content moderation (CCM) in user-created spaces. This could be due to the lack of availability of a workforce that would be able to support their needs. The need for moderators to be present in events while they are live, coupled with the time difference with locations where CCMs might be located, could result in conflicts with the working hours of these moderators. The lack of CCM presence in these spaces could cause problems such as biases in moderation or inaction towards specific types of harm.

Moreover, human moderation is regarded as civic labor and can have emotional effects on moderators [96, 116]. In non-ephemeral contexts, moderators can have time to reflect and consult with other moderators whether the governance structure is flat or hierarchical. This is especially useful when moderators are new or the content being assessed is subtle or implicit. However, in ephemeral spaces, moderators have to react and respond to harmful behavior within seconds. This could result in extra burdens or taxation on moderators. This would be exacerbated by the diverse roles that they engage in.

5.1 Design Implications for Moderating Social VR

5.1.1 Mindset Shift Towards More Preemptive and Facilitatory Approaches to Moderation. The difficulties of real-time moderation and demobilization of harm have caused moderators to make an effort to prevent harms from taking place and control the flow of events. This desire has resulted in an evolution in the perception of moderation as only punishment for harmful behaviors to include actions that aim to assist or ward off harm.

It is important to note that while moderation responses have been extended to include those that are preemptive or facilitatory, the list of the types of punitive responses that are available in these ephemeral spaces has reduced significantly. Punitive responses in ephemeral spaces are currently only comprised of warning, muting, or removal which are limited forms of norm-setting and excluding (i.e., "depriving the community of the contributions that those who are excluded could have made" [63]).

This limit in methods for handling harmful behaviors could result in moderators having severe responses because they don't have the possibility to take proportionate action. For example, in some cases of disruptive movements that are more distracting than harmful, moderators don't have any response that would be somewhere between warning and removal in intensity. Some participant suggestions for such tools were discussed in this study but future design workshops might consider creating a gradient of moderation responses.

5.1.2 More Powers for Moderators to See and Hear. To combat the lack of recorded evidence of harmful behaviors, all VR moderators make every effort to increase the visibility of individuals within ephemeral spaces, attempting to emulate the same type of perceptibility that might be strived for in real-world events or gatherings. However, this belief in imitating face-to-face communications in online spaces has been shown to be an inefficient strategy [66] and could limit making use of the power of existing technologies to their fullest extent.

As a result, while designing tools to increase the possibility of observing all individuals in attendance at an event could be a fruitful research direction, future studies into the design of automated moderation tools might benefit from moving away from this endeavor and instead focusing on designing tools that would allow monitoring and getting cues for different behaviors without personally viewing them. For example, a moderation dashboard that would display summary statistics for all user activity, such as the performance of specific movements or the utterance of particular phrases or noises, all in one place could be of great value and even help with evidence gathering in cases of rule infractions. This dashboard's backend could employ event sensing mechanisms for objects that could be used for violent actions (such as knives, guns, and sticks), chasing of avatars, and the invasion of personal space, as well as harmful behavior detection for text representations in these spaces (e.g., "kill all men" banner). Additionally, speech-to-text systems could be incorporated to leverage prior natural language processing work on harmful language detection including violent and hateful speech [46, 114, 126, 150] and cyberbullying [67, 120, 144] for near real-time detection of such problematic language.

5.1.3 Teams of Moderators with Different Roles to Reduce Burden and Potential for Bias. We identify three major challenges for moderation: the possibility of multiple concurrent harms, and the vastness of spaces that make it difficult/impossible to hear/see all harms, and the limitations of current tools that make near-real-time moderation necessary. These three challenges make the job of moderation very taxing or even impossible for moderators working alone. Additionally, since there is no time to consult with others within the community (due to the need for real-time responses), working alone can increase the probability of biased responses.

The use of teams with clearly defined roles that reduce some of these burdens for individual moderators could be a helpful strategy. Successful groups assigned different tasks to different moderators. But many teams did not differentiate roles. For example, this included failing to split up spaces. Moreover, none of the systems support communication between moderators. As a result, a number of VR moderators join Discord channels while present in the VR world – forcing them to process multiple conversations simultaneously. Systems should focus on designing easier interaction and collaboration between moderation team members.

5.2 Limitations and Future Work

In this study, we investigated harms and moderation practices through observations and interviews. Observations were limited to public gatherings and events on the platforms due to a lack of access to private events. We also excluded events with low participant counts. Future work could investigate differences in behavior in smaller groups of individuals. Interviews were also limited to AltspaceVR moderators, as we faced challenges (e.g., the taciturnity of official platform moderators, the newness of platforms, and the age of moderators) in the recruitment of participants from other social VR spaces. Future work could explore methods to address the concerns of moderators and incentivize them to participate in research studies. Moreover, while some interviewees had moderation experience in other platforms, examination of differences between moderation practices on various VR platforms, including moderation practices of children (e.g., moderators on Rec Room as our observations indicated that most moderators on this platform were not adults) could be of value. Future research could also extend to other types of ephemeral social spaces (e.g., augmented reality and emerging voice-based platforms). While much of the identified harms and moderation challenges could be present in these spaces as well, further examination of these platforms could reveal new harmful behaviors or potential modifications to moderation practices.

6 CONCLUSION

This study addressed the gap in understanding moderation practices and harms in ephemeral social spaces. Through conducting 100 virtual observations and 11 interviews for three different platforms, we find that the real-time nature of interactions, ephemerality of content, and limitations in visibility which are all characteristics of virtual reality spaces, can introduce significant challenges for moderators. These challenges have made moderators move away from viewing moderation responses as purely punitive, and incorporate preemptive and facilitatory responses to prevent occurrences of harmful behavior as much as possible. Moreover, these features of ephemeral platforms and the addition of capabilities such as having virtual bodies, have resulted in the inception of new forms of harm. The limited moderation tools provided by the platforms, however, make it hard to have a gradient of moderation responses when these harms occur. Finally, we discussed how real-time activity, wide territory, and concurrent activity can introduce challenges for current moderation frameworks, and discuss how current practices have evolved to address some of these issues.

ACKNOWLEDGMENTS

Dow was supported in part by the National Science Foundation under Grant #2009003. We would also like to thank the Virtual Reality Lab for lending an Oculus Quest VR headset to us for the duration of this study.

REFERENCES

[1] Microsoft. (2023). AltspaceVR: Host tools overview. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/getting-started/roles. Accessed: 2023-02-13.

[2] Jacinda Santora. (2022). Clubhouse Statistics: Revenue, Users and More (2022). https://influencermarketinghub.com/clubhouse-stats/#toc-0. Accessed: 2022-08-16.

[3] ThinkImpact. Clubhouse User Statistics. https://www.thinkimpact.com/clubhouse-statistics/. Accessed: 2022-08-16.

[4] Fandom. Discord Wiki. https://discord.fandom.com/wiki/Discord. Accessed: 2022-09-02.

[5] David Pierce. (2020). How Discord (somewhat accidentally) invented the future of the internet. https://www.protocol.com/discord. Accessed: 2022-09-02.

[6] Kris Holt. (2022). Meta adds 'personal boundaries' to Horizon Worlds and Venues to fight harassment. https://techcrunch.com/2022/02/04/meta-adds-personal-boundaries-to-horizon-worlds-and-venues-to-fight-harassment/. Accessed: 2022-08-16.

[7] Adi Robertson. (2022). Meta is adding a 'personal boundary' to VR avatars to stop harassment. https://www.theverge.com/2022/2/4/22917722/meta-horizon-worlds-venues-metaverse-harassment-groping-personal-boundary-feature. Accessed: 2022-08-16.

[8] Alex Heath. (2022). Meta's social VR platform Horizon hits 300,000 users. https://www.theverge.com/2022/2/17/22939297/meta-social-vr-platform-horizon-300000-users. Accessed: 2022-08-16.

[9] Rec Room. A Parent's Guide to Rec Room. https://recroom.com/parents-guide. Accessed: 2022-08-16.

[10] Rec Room. REC ROOM: Appealing a Ban. https://rec.net/ban-appeal. Accessed: 2022-12-12.

[11] Meta. (2022). Supplemental Horizon Worlds Terms of Service. https://www.meta.com/legal/quest/terms-of-service/. Accessed: 2022-08-16.

[12] Meta Horizon. (2022). Tweet by Horizon Worlds' official Twitter account about the number of worlds. https://twitter.com/MetaHorizon/status/1494007916990373895. Accessed: 2022-08-16.

[13] Mark Pappas and Julie Hurvitz Aliaga. (2021). Voice-Based Social Networks: Understanding and Leveraging the Clubhouse Trend. https://cmimediagroup.com/resources/voice-based-social-networks-understanding-and-leveraging-the-clubhouse-trend/. Accessed: 2022-09-02.

[14] Rec Room. What is a Junior Account? https://recroom.zendesk.com/hc/en-us/articles/4426900227735-What-is-a-Junior-Account-. Accessed: 2022-08-16.

[15] Ashwin Ram. (2021). Why audio-based social media is the future. https://www.zoho.com/social/journal/audio-based-socialmedia.html. Accessed: 2022-09-02.

[16] Caty McCarthy. The "space bubble" ensures you always have personal space in VR. https://killscreen.com/versions/users-can-no-longer-encroach-personal-space-thanks-altspaces-space-bubble/. Accessed: 2022-08-16.

[17] Paul JC Adachi and Teena Willoughby. 2011. The effect of violent video games on aggression: Is it more than just the violence? *Aggression and Violent behavior* 16, 1 (2011), 55–62.

[18] Haider M al Khateeb and Gregory Epiphaniou. 2016. How technology can mitigate and counteract cyber-stalking and online grooming. *Computer Fraud & Security* 2016, 1 (2016), 14–18.

[19] Marina Amâncio. 2017. "Put it in your Story": Digital Storytelling in Instagram and Snapchat Stories.

[20] Joseph Aneke, Carmelo Ardito, and Giuseppe Desolda. 2021. Help the User Recognize a Phishing Scam: Design of Explanation Messages in Warning Interfaces for Phishing Attacks. In *International Conference on Human-Computer Interaction*. Springer, 403–416.

[21] Florian Arendt. 2019. Suicide on Instagram—Content analysis of a German suicide-related hashtag. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 40, 1 (2019), 36.

[22] Florian Arendt, Sebastian Scherr, and Daniel Romer. 2019. Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society* 21, 11-12 (2019), 2422–2442.

[23] Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law and Technology* (2018).

[24] Emmanuel W Ayaburi and Daniel N Treku. 2020. Effect of penitence on social media trust and privacy concerns: The case of Facebook. *International Journal of Information Management* 50 (2020), 171–181.

[25] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction* 5, CSCW1 (2021), 1–34.

[26] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[27] Joseph B Bayer, Nicole B Ellison, Sarita Y Schoenebeck, and Emily B Falk. 2016. Sharing the small moments: ephemeral social interaction on Snapchat. *Information, Communication & Society* 19, 7 (2016), 956–977.

[28] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[29] Joanna Bergström, Tor-Salve Dalsgaard, Jason Alexander, and Kasper Hornbæk. 2021. How to evaluate object selection and manipulation in VR? Guidelines from 20 years of studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[30] Candice Biernesser, Craig JR Sewall, David Brent, Todd Bear, Christina Mair, and Jeanette Trauth. 2020. Social media use and deliberate self-harm among youth: A systematized narrative review. *Children and youth services review* 116 (2020), 105054.

[31] 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.

[32] Danielle Blunt, Ariel Wolf, Emily Coombes, and Shanelle Mullin. 2020. Posting into the void: Studying the impact of shadowbanning on sex workers and activists. *Retrieved September* 6 (2020), 2021.

[33] Tom Boellstorff, Bonnie Nardi, Celia Pearce, and Tina L Taylor. 2012. *Ethnography and virtual worlds*. Princeton University Press.

[34] Rebecca C Brown, Tin Fischer, David A Goldwich, and Paul L Plener. 2020. "I just finally wanted to belong somewhere"—Qualitative Analysis of Experiences With Posting Pictures of Self-Injury on Instagram. *Frontiers in psychiatry* 11 (2020), 274.

[35] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*. 75–84.

[36] 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.

[37] 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.

[38] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring# GamerGate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international conference on world wide web companion*. 1285–1290.

[39] Mauro Coletto, Luca Maria Aiello, Claudio Lucchese, and Fabrizio Silvestri. 2016. Pornography consumption in social media. *arXiv preprint arXiv:1612.08157* (2016).

[40] Elisabetta Costa. 2018. Affordances-in-practice: An ethnographic critique of social media logic and context collapse. *New Media & Society* 20, 10 (2018), 3641–3656.

[41] Lorrie Faith Cranor and Simson Garfinkel. 2005. *Security and usability: designing secure systems that people can use.* " O'Reilly Media, Inc.".

[42] Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Eleventh International AAAI Conference on Web and Social Media*.

[43] Michele P Dyson, Lisa Hartling, Jocelyn Shulhan, Annabritt Chisholm, Andrea Milne, Purnima Sundar, Shannon D Scott, and Amanda S Newton. 2016. A systematic review of social media use to discuss and view deliberate self-harm acts. *PloS one* 11, 5 (2016), e0155813.

[44] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition* 38, 8 (2010), 1087–1100.

[45] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[46] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 345–363.

[47] Charlotte Emma Hilton. 2017. Unveiling self-harm behaviour: what can social media site Twitter tell us about self-harm? A qualitative exploration. *Journal of clinical nursing* 26, 11-12 (2017), 1690–1704.

[48] Daniel A Epstein, Siyun Ji, Danny Beltran, Griffin D'Haenens, Zhaomin Li, and Tan Zhou. 2020. Exploring design principles for sharing of personal informatics data on ephemeral social media. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[49] Rubia Fatima, Affan Yasin, Lin Liu, and Jianmin Wang. 2019. How persuasive is a phishing email? A phishing game for phishing awareness. *Journal of Computer Security* 27, 6 (2019), 581–612.

[50] Christopher John Ferguson. 2007. The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric quarterly* 78, 4 (2007), 309–316.

[51] Emilio Ferrara, Stefano Cresci, and Luca Luceri. 2020. Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science* 3, 2 (2020), 271–277.

[52] 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.

[53] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[54] Guo Freeman and Dane Acena. 2021. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *ACM International Conference on Interactive Media Experiences*. 84–95.

[55] Guo Freeman and Dane Acena. 2022. " Acting Out" Queer Identity: The Embodied Visibility in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[56] Guo Freeman, Divine Maloney, Dane Acena, and Catherine Barwulor. 2022. (Re) discovering the Physical Body Online: Strategies and Challenges to Approach Non-Cisgender Identity in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[57] 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.

[58] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. 2020. My body, my avatar: How people perceive their avatars in social virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[59] Bharath Ganesh and Jonathan Bright. 2020. Countering extremists on social media: Challenges for strategic communication and content moderation. , 6–19 pages.

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

[61] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

[62] 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.

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

[64] 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.

[65] Benjamin Hanckel, Son Vivienne, Paul Byron, Brady Robards, and Brendan Churchill. 2019. 'That's not necessarily for them': LGBTIQ+ young people, social media platform affordances and identity curation. *Media, Culture & Society* 41, 8 (2019), 1261–1278.

[66] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 119–125.

[67] Qianjia Huang, Diana Inkpen, Jianhong Zhang, and David Van Bruwaene. 2018. Cyberbullying Intervention Based on Convolutional Neural Networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 42–51. https://aclanthology.org/W18-4405/

[68] Wen-Chun Hung. 2022. Exploring the factors that influence users' intention to continue using audio-based social media: The Clubhouse case. (2022).

[69] Adrienne Holz Ivory, James D Ivory, Winston Wu, Anthony M Limperos, Nathaniel Andrew, and Brandon S Sesler. 2017. Harsh words and deeds: Systematic content analyses of offensive* user behavior in the virtual environments of online first-person shooter games. *Journal For Virtual Worlds Research* 10, 2 (2017).

[70] 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.

[71] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill. 2021. Effects of emotion and agency on presence in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[72] Kyuha Jung, Yoobin Park, Hanwool Kim, and Joonhwan Lee. 2022. Let's Talk@ Clubhouse: Exploring Voice-Centered Social Media Platform and its Opportunities, Challenges, and Design Guidelines. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[73] Steven J Kirsh. 2003. The effects of violent video games on adolescents: The overlooked influence of development. *Aggression and violent behavior* 8, 4 (2003), 377–389.

[74] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.

[75] 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.

[76] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*. 85–94.

[77] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 905–914.

[78] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[79] Daniel S Lane, Vishnupriya Das, and Dan Hiaeshutter-Rice. 2019. Civic laboratories: youth political expression in anonymous, ephemeral, geo-bounded social media. *Information, Communication & Society* 22, 14 (2019), 2171–2186.

[80] Song Mi Lee, Cliff Lampe, JJ Prescott, and Sarita Schoenebeck. 2022. Characteristics of People Who Engage in Online Harassing Behavior. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[81] Keren Lehavot, Dror Ben-Zeev, and Robin E Neville. 2012. Ethical considerations and social media: a case of suicidal postings on Facebook. *Journal of Dual Diagnosis* 8, 4 (2012), 341–346.

[82] Zachary Leibowitz. 2017. Terror on your timeline: Criminalizing terrorist incitement on social media through doctrinal shift. *Fordham L. Rev.* 86 (2017), 795.

[83] Lawrence Lessig. 2006. Code 2.0: Code and other laws of cyberspace.

[84] Lingyuan Li, Guo Freeman, and Donghee Yvette Wohn. 2020. Power in Skin: The Interplay of Self-Presentation, Tactical Play, and Spending in Fortnite. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 71–80.

[85] Lyrissa Barnett Lidsky. 2011. Incendiary speech and social media. *Tex. Tech L. Rev.* 44 (2011), 147.

[86] Daniel Link, Bernd Hellingrath, and Jie Ling. 2016. A Human-is-the-Loop Approach for Semi-Automated Content Moderation.. In *ISCRAM*.

[87] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *IFIP conference on human-computer interaction*. Springer, 231–248.

[88] Cayley MacArthur, Arielle Grinberg, Daniel Harley, and Mark Hancock. 2021. You're making me sick: A systematic review of how virtual reality research considers gender & cybersickness. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[89] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019), e0221152.

[90] Divine Maloney and Guo Freeman. 2020. Falling asleep together: What makes activities in social virtual reality meaningful to users. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 510–521.

[91] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. " Talking without a Voice" Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[92] Carmia Margaret and David Alinurdin. 2021. A Christian Response to the Use of the Clubhouse Apps During the Covid-19 Era. *Societas Dei* 8, 2 (2021), 229–245.

[93] 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.

[94] Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.

[95] Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media* 22, 2 (2021), 205–224.

[96] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[97] Philipp Mayring et al. 2004. Qualitative content analysis. , 159–176 pages.

[98] Sarah McRoberts, Haiwei Ma, Andrew Hall, and Svetlana Yarosh. 2017. Share first, save later: Performance of self through Snapchat stories. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6902–6911.

[99] 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.

[100] George B Mo, John J Dudley, and Per Ola Kristensson. 2021. Gesture Knitter: A Hand Gesture Design Tool for Head-Mounted Mixed Reality Applications. In

Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–13.

[101] Megan A Moreno, Adrienne Ton, Ellen Selkie, and Yolanda Evans. 2016. Secret society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health* 58, 1 (2016), 78–84.

[102] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

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

[104] Helen Nissenbaum. 2009. Privacy in context. In *Privacy in Context*. Stanford University Press.

[105] Jessica Outlaw and Beth Duckles. 2018. Virtual harassment: The social experience of 600+ regular virtual reality (VR) users. *The Extended Mind Blog* 4 (2018).

[106] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review* (2020).

[107] 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.

[108] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open* 10, 4 (2020), 2158244020973022.

[109] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[110] Gustav Bøg Petersen, Aske Mottelson, and Guido Makransky. 2021. Pedagogical agents in educational vr: An in the wild study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[111] Jacobo Picardo, Sarah K McKenzie, Sunny Collings, and Gabrielle Jenkin. 2020. Suicide and self-harm content on Instagram: A systematic scoping review. *PloS one* 15, 9 (2020), e0238603.

[112] Roosa Piitulainen, Perttu Hämäläinen, and Elisa D Mekler. 2022. Vibing Together: Dance Experiences in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[113] Michael L Pittaro. 2007. Cyber stalking: An analysis of online harassment and intimidation. *International journal of cyber criminology* 1, 2 (2007), 180–197.

[114] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 118–123. https://aclanthology.org/N18-2019/

[115] Donna J Reid and Fraser JM Reid. 2007. Text or talk? Social anxiety, loneliness, and divergent preferences for cell phone use. *CyberPsychology & Behavior* 10, 3 (2007), 424–435.

[116] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

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

[118] Sarah T Roberts. 2017. *Content moderation*. University of California, Los Angeles.

[119] Sarah T Roberts. 2018. Digital detritus:'Error'and the logic of opacity in social media content moderation. *First Monday* (2018).

[120] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[121] Michelle Roter. 2016. With Great Power Comes Great Responsibility: Imposing a Duty to Take Down Terrorist Incitement on Social Media. *Hofstra L. Rev.* 45 (2016), 1379.

[122] Sebastian Scherr. 2022. Social media, self-harm, and suicide. *Current opinion in psychology* (2022), 101311.

[123] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[124] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2018), 1–27.

[125] Jonas Schjerlund, Kasper Hornbæk, and Joanna Bergström. 2021. Ninja hands: Using many hands to improve target selection in vr. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[126] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, 1–10. https://aclanthology.org/W17-1101/

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

[128] 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.

[129] Jana Shakarian, Andrew T Gunn, and Paulo Shakarian. 2016. Exploring malicious hacker forums. In *Cyber deception*. Springer, 259–282.

[130] John L Sherry. 2001. The effects of violent video games on aggression: A meta-analysis. *Human communication research* 27, 3 (2001), 409–431.

[131] William Shotts. 2017. *The Linux command line*. LinuxCommand. org.

[132] Ellen Silver. 2018. Hard questions: who reviews objectionable content on Facebook—and is the company doing enough to support them. *Facebook Newsroom* 26 (2018).

[133] Olivia Solon. 2017. Underpaid and overburdened: the life of a Facebook moderator. *The Guardian* 25, 05 (2017), 2017.

[134] Daniel J Solove. 2008. Understanding privacy. (2008).

[135] Francesca Stevens, Jason RC Nurse, and Budi Arief. 2021. Cyber stalking, cyber harassment, and adult mental health: A systematic review. *Cyberpsychology, Behavior, and Social Networking* 24, 6 (2021), 367–376.

[136] Jenna Strawhun, Natasha Adams, and Matthew T Huss. 2013. The assessment of cyberstalking: An expanded examination including social networking, attachment, jealousy, and anger in relation to violence and abuse. *Violence and victims* 28, 4 (2013), 715–730.

[137] JoAnne Sweeny. 2019. Incitement in the Era of Trump and Charlottesville. *Cap. UL Rev.* 47 (2019), 585.

[138] Philipp Sykownik, Divine Maloney, Guo Freeman, and Maic Masuch. 2022. Something Personal from the Metaverse: Goals, Topics, and Contextual Factors of Self-Disclosure in Commercial Social VR. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[139] Sabine Trepte and Leonard Reinecke. 2011. *Privacy online: Perspectives on privacy and self-disclosure in the social web*. Springer.

[140] Penny Trieu and Nancy K Baym. 2020. Private responses for public sharing: understanding self-presentation and relational maintenance via stories in social media. In *proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[141] Hsin-Ruey Tsai, Yuan-Chia Chang, Tzu-Yun Wei, Chih-An Tsao, Xander Chin-yuan Koo, Hao-Chuan Wang, and Bing-Yu Chen. 2021. GuideBand: Intuitive 3D Multilevel Force Guidance on a Wristband in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[142] 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.

[143] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[144] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and Fine-Grained Classification of Cyberbullying Events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 672–680. https://aclanthology.org/R15-1086/

[145] Jorge Vázquez-Herrero, Sabela Direito-Rebollal, and Xosé López-García. 2019. Ephemeral journalism: News distribution through Instagram stories. *Social media+ society* 5, 4 (2019), 2056305119888657.

[146] Elena Villaespesa and Sara Wowkowych. 2020. Ephemeral storytelling with social media: Snapchat and Instagram stories at the Brooklyn Museum. *Social Media+ Society* 6, 1 (2020), 2056305119898776.

[147] Anna JM Wagner. 2018. Do not click "like" when somebody has died: The role of norms for mourning practices in social media. *Social Media+ Society* 4, 1 (2018), 2056305117744392.

[148] Endang Wahyuningsih and Baidi Baidi. 2021. Scrutinizing the potential use of Discord application as a digital platform amidst emergency remote learning. *Journal of Educational Management and Instruction* 1, 1 (2021), 9–18.

[149] Dennis Wang, Marawin Chheang, Siyun Ji, Ryan Mohta, and Daniel A Epstein. 2022. SnapPI: Understanding Everyday Use of Personal Informatics Data Stickers on Ephemeral Social Media. *Proceedings of the ACM on human-computer interaction* 7 (2022).

[150] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://aclanthology.org/N16-2013/

[151] Julie Williamson, Jie Li, Vinoba Vinayagamoorthy, David A Shamma, and Pablo Cesar. 2021. Proxemics and social interactions in an instrumented virtual reality workshop. In *Proceedings of the 2021 CHI Conference on Human Factors*

in Computing Systems. 1–13.

[152] Julie R Williamson, Joseph O'Hagan, John Alexis Guerra-Gomez, John H Williamson, Pablo Cesar, and David A Shamma. 2022. Digital Proxemics: Designing Social and Collaborative Interaction in Virtual Environments. In *CHI Conference on Human Factors in Computing Systems*. 1–12.

[153] Janis Wolak, Kimberly J Mitchell, and David Finkelhor. 2007. Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *Journal of adolescent health* 41, 6 (2007), S51–S58.

[154] Paulina Wu. 2015. Impossible to regulate: Social media, terrorists, and the role for the UN.. In *Chi. J. Int'l L.*, Vol. 16. 281.

[155] Arum Nisma Wulanjani. 2018. Discord application: Turning a voice chat application for gamers into a virtual listening class. In *English Language and Literature International Conference (ELLiC) Proceedings*, Vol. 2. 115–119.

[156] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 656–666.

[157] Yukun Yang. 2019. When power goes wild online: How did a voluntary moderator's abuse of power affect an online community? *Proceedings of the Association for Information Science and Technology* 56, 1 (2019), 504–508.

[158] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[159] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. Gaze-supported 3d object manipulation in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[160] Binhe Zhu. 2021. Clubhouse: A popular audio social application. In *2021 International Conference on Public Relations and Social Sciences (ICPRSS 2021)*. Atlantis Press, 575–579.