

Enabling Developers, Protecting Users: Investigating Harassment and Safety in VR

Abhinaya S.B.
North Carolina State University
asrivid@ncsu.edu

Aafaq Sabir
North Carolina State University
asabir2@ncsu.edu

Anupam Das
North Carolina State University
anupam.das@ncsu.edu

Abstract

Virtual Reality (VR) has witnessed a rising issue of harassment, prompting the integration of safety controls like muting and blocking in VR applications. However, the lack of standardized safety measures across VR applications hinders their universal effectiveness, especially across contexts like socializing, gaming, and streaming. While prior research has studied safety controls in social VR applications, our user study ($n = 27$) takes a multi-perspective approach, examining both users’ perceptions of safety control usability and effectiveness as well as the challenges that developers face in designing and deploying VR safety controls. We identify challenges VR users face while employing safety controls, such as when finding users in crowded virtual spaces in order to block them. VR users also find controls ineffective in addressing harassment; for instance, they fail to eliminate the harassers’ presence from the environment. Further, VR users find the current methods of submitting evidence for reports time-consuming and cumbersome. Improvements desired by users include live moderation and behavior tracking across VR apps; however, developers cite technological, financial, and legal obstacles to implementing such solutions, often due to a lack of awareness and high development costs. We emphasize the importance of establishing technical and legal guidelines to enhance user safety in virtual environments.

1 Introduction

Content Warning: This paper studies harassment in VR. When necessary, this paper directly quotes participants, which may contain descriptions of offensive/hateful speech, profanity, and other potentially triggering content.

Virtual Reality (VR) is an emerging technology that enables users to partake in 360-degree virtual experiences using VR head-mounted displays [1–3]. VR offers full-body tracking and synchronous voice chat and has controllers that provide haptic feedback [4], allowing people to interact in newer, more immersive ways compared to traditional social media [5].

While VR presents these novel affordances, it also lowers the bar for unwanted behavior by malicious social actors. The anonymity it provides to users [6, 7], as well as the lack of their physical presence, not only increases the likelihood of harassment but also makes identification of harassers challenging [8]. While online harassment is not a new issue, the unique sense of embodiment and presence that VR enables [5, 9], even without haptic technology [10], may amplify certain forms of harassments (e.g., sexual harassment) in virtual spaces.

Harassment in VR is becoming prominent today [11–16], with an abusive incident estimated to occur every seven minutes [17]. VR-based harassment may include virtual violence, virtual groping [18], and haptic sex crimes [19]. To enable users to deal with harassment, VR applications have introduced safety¹ controls such as the personal bubble, power gesture [21], safe zone [22], etc. Targets of VR-based harassment have reported challenges in escaping and reporting problematic users [5], highlighting limitations of existing safety controls. The set of safety controls is not standardized across VR apps, with high variance in functionalities they provide [18], hinting at different usability challenges that may exist for even the same control across VR apps.

To identify gaps in our understanding of harassment controls and better address harassment in VR, it is essential to understand VR safety controls more deeply through the lens of targets subjected to VR-based harassment. First-hand accounts of these targets would elucidate the types of harassing activities that have a lasting impact and how the availability of safety controls, or lack thereof, contributed to these experiences. Further, it would enable the understanding of the usability and effectiveness of current safety controls and how they may be improved. When users’ perspectives are contrasted with those of VR developers, i.e., those involved in the implementation of these features, the combined knowledge

¹We consider “safety” in the context of interactions (with users, content, etc.) within the VR environment that threaten a user. We don’t consider physical safety implications arising intrinsically from VR hardware (e.g., headset, display), such as flashing lights causing epileptic seizures [19, 20], or injuries arising from careless use of VR (e.g., hitting a wall).

will highlight the gaps in practically addressing safety in VR.

Although prior works have identified types of VR-based harassment [5, 18, 23] and availability of safety controls [18] in VR, they have not studied the *effectiveness* of safety controls and reporting mechanisms, nor have they investigated users’ perceptions of their *usability* at the time of harassment. We address this research gap by conducting semi-structured interviews with targets (n = 18) of VR-based harassment. We also augment our findings by interviewing (n = 9) VR developers to understand their perceptions on designing and deploying proposed safety features for VR. Specifically, we seek to answer the following research questions:

RQ 1: *How do targets of VR-based harassment perceive the usability and effectiveness of existing safety controls and reporting mechanisms?* We delve into the participants’ thought processes behind (not) using safety controls and how such actions contributed to their experience of VR-based harassment.

RQ 2: *What are the expectations and recommendations by targets of VR-based harassment for making VR safer?* We understand what new safety measures our participants desire based on their experiences of VR-based harassment.

RQ 3: *What are VR developers’ perceptions of the design and deployment of safety controls?* We understand the challenges VR developers perceive in implementing VR safety controls. Further, we understand developers’ views on the feasibility of the safety features desired by targets of VR-based harassment.

We performed a *qualitative analysis* [24] of our participants’ responses using open coding [25]. In this paper, we make the following contributions:

- To the best of our knowledge, we are the first to conduct a multi-perspective study on VR safety through the lens of *targets of VR-based harassment* and *VR developers*.
- We identify contexts where existing VR safety controls and moderation practices are non-usable and ineffective. For instance, VR users face usability challenges in finding users in crowded virtual spaces to block them. Safety controls are also ineffective in providing feedback to the harassers.
- We highlight VR users’ expectations for making VR safer and contrast them with technical, legal, and financial challenges that VR developers perceive in implementing them. Users desire live moderation in social spaces and want users’ behavior to be tracked across VR apps; however, VR developers highlight difficulties in deploying live moderation at scale and the privacy risks in tracking users.
- We use our findings from this multi-perspective study to make recommendations to VR platform owners, app developers, and policy makers for improving safety in VR.

2 Background & Related Work

Online Abuse. Online harassment refers to behaviors that threaten or offend individuals through emails, instant messages, social media, etc. [26, 27]. The ability of users to retain their anonymity [28] exacerbates harassment, and the lack of

Table 1: Prominent safety controls available in VR apps.

Safety Control	Function
Mute	Disable voice chat of self, or other users in a VR space
Block	Hide or change the appearance of user(s) in a VR space
Proximity setting	Control the distance at which other users can interact with a user in a VR space
Quick travel	Travel to a different location within a VR app
Safe zone	A user’s private space accessible only to that user
Vote kick	Kick a user out of a VR space based on majority vote
Trust rank	Levels of trust assigned to a user

their physical presence on these platforms makes their identification challenging, causing emotional distress to victims [8].

Prior works have extensively studied online abuse in non-VR contexts. Thomas et al. [29] created a taxonomy of types of online hate and harassment, identifying seven classes of attacks including toxic content, content leakage and impersonation, based on attackers’ intents and capabilities. Of these, toxic content (e.g., bullying, trolling, hate speech) is viewed the highest priority threat among experts as it incurs significant emotional harm [30]. Additionally, what constitutes toxic content differs across demographics, beliefs, and personal experiences [31]. Researchers have studied technology-based abuse specific to at-risk populations such as youth [32] and sex workers [33–35], as well as how technology facilitates intimate partner abuse and surveillance via mobile [36] and IoT devices [37, 38]. Women, racial/cultural minorities, LGBTQ individuals, and persons with disabilities face more significant risks of online harassment [39, 40].

Safety in VR. VR can act as a medium for harassing activities such as virtual violence, virtual groping [18], and haptic sex crimes [19]. The sense of embodiment and presence facilitated by 3D avatars in VR environments [5, 9], and the ability of some users to feel as though they are their avatars even without haptic technology (termed as *phantom sense*) [10], makes harassment in VR realistic and thus traumatizing. While more female avatars report VR harassment [41], vulnerable populations such as minors face risks of harm through virtual grooming and erotic role-play abuse [42].

VR Apps & Safety Controls. VR apps can be downloaded into standalone or tethered VR headsets [43] through app stores such as Oculus Store [44], Steam VR [45], and Sidequest [46]. While every app has a primary purpose, such as watching movies (e.g., ‘BigScreen’ [47]) or competitive gaming (e.g., ‘Pavlov VR’ [48]), many apps have capabilities for multi-user interaction through social spaces and lobbies. Additionally, there are certain VR apps primarily meant for social interaction, called Social VR apps (e.g., ‘VRChat’ [49]).

VR apps have introduced safety controls such as muting, blocking, personal bubble and power gesture [21], personal boundary [50], and safe zone [22] to enable users to deal with harassment. Table 1 describes the most prominent safety con-

trols. Muting may be used either to disable one’s own voice such that it cannot be heard by other users, or disable other users’ voice in a VR environment. "Blocking" — also known as "ghosting" — may make another user’s avatar disappear or change its appearance, depending on the implementation. Proximity settings, also referred to as the "space bubble", "personal bubble" or the "personal boundary", enables a user to control the distance at which other users in a VR space can interact with them. "Safe zone" is a user’s private space that can be accessed only by the user, and may be invoked to "move away" from other users to a safe space. While "safe zone" teleports the user to a pre-determined location/VR space, "quick travel" can be used to move to other VR spaces within the same app (not just the safe zone). Some apps have custom implementations like in the case of the power gesture, which involves the action of "putting your hands together, pulling both triggers, and pulling them apart as if you are creating a force field" [21]. Some VR apps have additional safety features such as ‘Safety and Trust System’ (‘VRChat’) [51], ‘Comfort and Safety’ (‘RecRoom’) [52], etc.

Safety controls may be used proactively (in anticipation of harassment) or reactively (after the harassment incident). Some are inherently used reactively (e.g., blocking, muting) while others (e.g., space bubble) are used proactively [23]. Additionally, VR apps offer ways for reporting harassment, which can be done inside the app, through the headset, or via email to the app developers. While controls like muting and blocking involve only users, reporting involves multiple stakeholders such as users, moderators and automated systems for toxicity detection [53].

Prior work has explored how users engage in social VR [54] and uncovered tensions in specific interactions, such as between children and adults [55, 56]. Moderating sensitive content [57] and reporting users [58] are essential in dealing with inappropriate interactions in VR. Researchers have explored the ethics of acceptable behavior [9, 59], and the influence of body-gender transfer in VR [60].

Freeman et al. [23] interviewed social VR users to understand the new characteristics of harassment emerging in social VR and also investigated users’ strategies and recommendations to mitigate harassment. Blackwell et al. [5] investigated users’ expectations of social norms and moderation practices in VR communities. Zheng et al. [18] analyzed videos of VR activities posted by social VR users to identify types of safety risks in social VR, including virtual violence, abuse, and sexual harassment. Schulenberg et al. [61] explored the (re)purposing of existing social VR features (e.g., boundary settings) for preventing interpersonal harm in VR.

Distinction from Prior Work. While prior works have characterized harassment in VR and touched upon safety controls, they have not studied VR users’ thought processes in (not) using specific safety controls or how the (non) usage of controls influenced their experience. To the best of our knowledge, we are the first to investigate the usability and effectiveness of

safety controls through the lens of *targets of VR-based harassment*. We further augment our findings with *VR developers’* perspectives on designing and deploying VR safety measures.

3 Methods

We conducted a phased multi-perspective study with 18 targets of VR-based harassment (**Study-I**) and 9 VR developers (**Study-II**). The findings from **Study-I** were used to design parts of **Study-II**. Table 2 details our recruitment and interview procedure for both studies. We further highlight our study design, data analysis methods, and limitations.

3.1 Study Design & Ethical Considerations

For **Study-I**, we advertised the study in VR-specific online forums. Apart from the recruitment platforms specified in Table 2, we circulated our study in forums for women in VR (e.g., Ladies of Population One) to recruit participants representing marginalized groups in tech spaces [62–64]. A few participants we interviewed shared the study with other groups, facilitating snowball sampling [65] (four participants were from groups where we did not directly advertise). We selected a stratified sample of participants for interviews based on their responses to the screening survey (looking at VR usage, harassment types experienced), irrespective of where they obtained the study information.

We consulted a psychology expert during study design² [66] and determined to exclude those diagnosed with post traumatic stress disorder (PTSD) [67] or seeking help for emotional distress. The expert reviewed our interview script and advised us on phrasing sensitive questions without triggering individuals. In our screening survey, we stated that selected participants would be asked about their harassment experience in VR, but also specified that talking about traumatic events can be cathartic [68]. We curated a list of mental health resources in case of emotional distress during the interview, which included links to websites containing details about mental health to and hot line numbers to institutions such as Center for Mental Health Services.

During the interview, in order to study usability challenges in VR safety controls, we first collected contextual information: we asked participants to specify the VR apps in which they experienced harassment (we focused on four apps utmost), then probed about use of safety controls and reporting mechanisms (explaining the terms when necessary) during/after the incident, and finally asked about what they desired for improving VR safety. After the interview, we inquired about the participants’ emotional state, and none reported adverse effects. Regardless, we shared mental health resources with all the participants. Participants were given the opportunity to review the interview transcripts; 11 reviewed them, reporting accurate captioning.

²all study instruments are anonymously available at this OSF link: https://osf.io/c7fks/?view_only=6a3413e996c941db9882b21ca47bd328

Table 2: An overview of methods for **Study-I** (with targets of VR-based harassment) and **Study-II** (with VR developers).

	Study-I	Study-II
Recruitment	VR-specific Discord servers, subreddits, Facebook groups	VR-specific LinkedIn groups, subreddits
Inclusion Criteria	Ages 18-64, resident of the US, active user of VR, experienced harassment in VR, not diagnosed with PTSD, not seeking help for emotional distress at the time of the study	Ages 18 and above, resident of the US, VR developer
Screening Questions	Type of harassment experienced in VR (trolling, bullying, etc. presented as a list of categories), VR apps in which harassment was experienced, VR usage	Roles held as VR developer (UI/UX designer, XR gameplay and tools engineer, AR/VR maintenance and support, etc.), details of VR apps currently developing
Demographics Collection	Just before the interview, to minimize sensitive data collection at screening phase (as participants were targets of harassment)	Part of the screening questionnaire
Interviews	18 targets of VR-based harassment ³ , conducted via Zoom from November 2022 to February 2023, lasting 59 min on average	9 VR developers, conducted via Zoom from August 2023 to September 2023, lasting 49 min on average
Participant Distribution	6 female, 1 non-binary, 5 non-cis, and 7 black persons; 3 participants were prominent in certain VR communities (T12, T13, T15). Refer Table 5 Appendix A.1 for more details.	6 professional developers, 2 VR-based researchers, 1 hobby developer, including the founder of a VR game and a doctorate in XR. Refer Table 6 in Appendix A.2 for more details.
Interview Protocol	(1) Describe harassment incident experienced in VR, (2) Use of safety controls, (3) Use of reporting mechanisms, (4) Expectations for enhancing safety in VR	(1) VR development experience, (2) Solving usability challenges in VR safety controls, (3) Challenges in implementing VR safety features, (4) Feasibility of user recommended features

We developed **Study-II** to contrast the findings of **Study-I** with developers’ perspectives, by identifying developmental challenges, and practical solutions. Our screening survey for **Study-II** asked about participants’ role as VR developers, the apps they developed, demographic information, and their LinkedIn profile to ascertain their fit for the study. Further, the survey had questions about whether participants had developed apps containing features like socializing, multiplayer gaming, learning, or streaming. We tried to prioritize those with experience in social and multiplayer categories (since those categories of apps have more opportunity for harassment) but eventually included others too.

We started the interviews by asking participants about their VR development experience, followed by their perspective on handling harassment in VR. Then, we asked about the app development pipeline they were involved in, and the challenges they perceived in implementing safety features. We also gathered their perspectives on the feasibility of users’ desired features. Our protocol also included questions about reporting and moderation. However, none of our participants were content/user-moderators, thus we could not report related findings. Questions about the developers’ experiences of VR-based harassment were not part of the interview protocol, but some participants shared their experiences.

Both **Study-I** and **Study-II** were approved by our Institutional Review Board (IRB). All audio recordings were transcribed and de-identified immediately after the interviews. Participants also had the option to withdraw from the interview at any time.

³We initially conducted 20 interviews, and later excluded the imposter participant’s data, totaling the number of valid participants to 18.

3.2 Data Collection and Analysis

We hosted the surveys of our study (screening, demographics, compensation) using Qualtrics [69]. We refined our interview protocol by piloting with three participants for **Study-I** and one for **Study-II**. Participants of both studies were given a 20 USD Amazon gift card after they completed the interviews. All the interviews were audio-recorded and transcribed using Whisper [70]. One researcher checked each transcript for accuracy. Two researchers conducted thematic analysis [24, 25] on the interview transcripts by independently coding the transcripts for both studies. For **Study-I**, the transcripts were divided into four sections (harassment incident, use of safety controls, use of reporting, expectations for safer VR). Each section was coded for all transcripts (in batches of five) followed by discussions to generate/update the codebook, before moving on to the next. In the case of **Study-II**, the researchers coded two transcripts together to create the initial codebook, and the rest of the transcripts were coded independently (in batches of two) and discussed for updating the codebook. For both studies, multiple discussions were conducted to reach an agreement to generate a codebook. Since our coding process involved multiple iterations and discussions, intercoder reliability was not necessary to be checked [71].

3.3 Limitations

As is typical with interview studies, our recruited sample size was relatively small, due to the sensitive topic of study (**Study-I**) and challenges in recruiting VR developers (**Study-II**). While the exclusion criteria for **Study-I** prevented us from interviewing those impacted deeply by VR-based harassment,

Table 3: Harassment incidents reported by the participants of **Study-I**. Each harassment incident is categorized according to taxonomies from prior work on online abuse attacks by Thomas et al. [29] and VR safety risks by Zheng et al. [18].

Harassment Type	Participants	Excerpts from participants
Trolling / Virtual abuse	T3, T4, T7, T8, T10, T13, T14, T15, T16, T17, T18	"My accent is different than the average American, and most of the players in this game are either American or British, I've had players chase me around calling me f*gg*t or gay or little b*tch the entire match." (T13, Echo VR)
Profanity / Virtual abuse	T1, T2, T3, T4, T5, T9, T10, T14, T15 T16	"We're doing a team task or something like that. They just, started saying some inappropriate vulgar things and I didn't really feel comfortable. (T1, Asgard's Wrath)"
Hate speech / Virtual abuse	T5, T8	"He would just yell at me. I think he must have had anger management issues or something; he would say random over-the-top vitriolic things. He sounded furious." (T8, Pavlov VR)
Threats of violence	T16	"He started saying, 'I am going to gut your mother and skin your family,' kept repeating it over and over. That was a bit scary, like, really unsettling." (T16, Echo VR)
Bullying / Virtual violence	T3, T7, T8, T9, T11, T13, T14, T15, T17, T18	"In the game, you get punched in the head, and it stuns you. I've had players chase me around, trying to punch me in the head the entire match." (T13, Echo Arena)
(Virtual) Sexual harassment	T4, T7, T12, T13	"There would be a guy that starts to pretend to r*pe me and encourages others. Because there's presence in VR, they're doing this physically to my avatar, putting their avatar's cr*tch in my face, making slurping noises, sucking [my] b**bs." (T12, Echo VR)
Explicit content	T2, T6, T11	"Seeing content, I'm not okay with, people not fully dressed." (T6, YouTube VR)
Virtual crashing	T11	"I've played it before. Someone had toyed with that game." (T11, House of Terror)
Virtual trash actions	T14	"You can high-five each other. You'll see people slap your hand a bunch of times to get the high five. When you don't, they'll just wave in front of your hand." (T14, BigScreen)
Misuse of safety features	T8	"Because everybody looks the same, if he found out a specific person was me, he would team kill." (T8, Pavlov VR)

we minimized risks to participants. To preserve participants' anonymity while sharing sensitive experiences of harassment, we refrained from having identity verification. However, after the interviews, we identified one imposter participant [72] who participated twice, based on the high similarity of responses³. For **Study-II**, we required participants to provide their LinkedIn profile during screening, and enable their webcams prior to the interview.⁴ In **Study-II**, most developers we interviewed were from small teams, which may not exactly represent development experiences for platforms our participants interacted with. However, safety is relevant in all apps, even those developed by smaller teams. Both studies relied on self-reported information from participants, which may be subject to social desirability bias. To address these limitations in future research, participants may be recruited from a broader range of forums, languages, and cultural backgrounds.

⁴When one of the participants failed to enable their camera at the start of the interview in Study-II, the researcher did not proceed with the interview.

4 Users' Perceptions on VR Safety

In this section, we describe our findings from **Study-I** (§ 3), outlining results from every part of our interview protocol. In the first section of the interview, we asked participants about the harassment experiences they had in VR apps, to contextualize their use of safety controls (§ 4.1). In the second part of the interview, we focused on participants' awareness of safety controls available in VR apps, as well as their perceptions of the usability and effectiveness of safety controls (§ 4.2). In the third part, we studied the effectiveness of reporting mechanisms (§ 4.3). Finally, we asked about their expectations for enhancing safety in VR (§ 4.4). Our participants' diversity (Table 5 in Appendix A.1) enabled us to capture varying perspectives stemming from a wide range of VR experiences.

4.1 Harassment Experienced by Participants

We asked participants to specify the VR apps in which they experienced harassment (Table 5 in Appendix A.1 lists the apps). Overall, we consider 35 participant-app pairs from 18 participants, across 4 social VR, 11 gaming VR, and 2 streaming VR apps. We categorized the harassment incidents

Table 4: Sources of harassment identified by participants.

Source	Participants	Excerpts from participants
Gender	T3, T4, T7 T12, T17	<i>"Oh, come over here, you b*tch, let me do blah, blah, blah. Women shouldn't even be in this game anyway. Why don't you get off and go make me a sandwich". (T12, Echo VR)</i> <i>"It was really frustrating, women are made to be some kind of joke in VR." (T17, Echo VR)</i>
Race	T4, T13, T15, T16	<i>"I wanted to play a game with another user and the person wasn't interested in playing with me because he was white and I was black." (T4, Beat Saber)</i>
Avatar attributes	T15, T16, T9, T11	<i>"One time I was wearing a turban and jeez, a bunch of kids walked up and wouldn't leave me alone. They kept pestering me and asking if I was Arabic, or Muslim. I didn't really know what to do." (T16, RecRoom)</i> <i>"The special season avatar, it's like a dog head. There is a cat avatar, like a Panther. If I have those on, I'll be called a furry, a lot." (T15, Echo VR)</i>
Physical attributes	T8, T13, T17	<i>"I had a deviated septum, where one part of the nasal passage is smaller than the other. So my voice sounded really nasally, like, the stereotypical nerd voice. People would make fun of me because of that a lot." (T8, VRChat)</i> <i>"Someone figured out I was mute and started going off on me making fun of my disability." (T17, VRChat)</i>

reported by participants according to existing taxonomies on online abuse [29] (trolling, profanity, hate speech, etc.) and VR safety [18] (virtual violence, virtual crashing, virtual sexual harassment, etc.). Table 3 lists example quotes for each type of harassment recounted by our participants. Here, we do not introduce new findings about harassment in VR, rather, we illustrate the different harassment types to set the context for discussing safety features in the following sections.

We identified that most of our participants experienced different forms of toxic content, apart from VR-specific scenarios such as virtual crashing (using tactics or bugs to ruin others' experience) and trash actions (activities typically intended to spoil the experiences of others) [18]. Participants described their experiences as *'annoying'*, *'uncomfortable'*, *'violating'*, and *'jarring'*. Participants reported experiencing harassment based on their *gender* (as in [23, 41]), with five out of six of our female participants reporting gender-based harassment by a male-like avatar. They also reported that *race*, unique *avatar attributes*, and *physical attributes* that helped identify a user's gender [18, 23], contributed to harassment, as illustrated in Table 4.

In summary, although many accounts of VR-based harassment can be categorized as "toxic content" [29], VR amplifies many forms of abuse (e.g., virtual sexual harassment, bullying) due to its immersion and presence [18] when compared to traditional social spaces online. These incidents leave a significant psychological impact, as reported by our participants.

4.2 Perceptions on Safety Controls

After asking about the harassment incident experienced by the participant, we asked them if they were aware of safety controls at the time of harassment, and if they used them. Based on their usage, we asked them about the usability and effectiveness of the safety controls they used. The list of safety controls was not picked beforehand, rather, participants came

up with controls themselves during the interview.

Awareness of Safety Controls. Of the 35 participant-app pairs considered, in 23 of the cases, participants were aware of at least one safety control at the time of harassment, which included muting (20), blocking (12) and ghosting (4), using safety bubble (3), vote kicking (1), a form of teleportation such as quick travel (2) or changing lobbies (1). In addition, one participant reported using parental guidance in 'YouTube VR' as a safety control. Participants learned about safety controls in a variety of ways. In seven of the cases, they were informed because of a tutorial available on the VR app. In two cases, the nudges in the app ('Echo VR') prompted them to learn about the controls. In three cases, their knowledge was due to contextual information from playing other video games. Participants also discovered the controls by chance, through exploring the app or searching the Internet.

However, the participants who were unaware of safety controls before encountering harassment did not necessarily attempt to learn about them after the incident:

"No, absolutely not. I just left it. [Using the app] is like a fun thing for me to do. I can just find another platform to just move on to." (T1, Asgard's Wrath)

T14 added that it was not worth his time to learn the safety controls and hinted at an unintuitive design: *"VRChat is a complicated app. I don't know how it works. It's not, to me, user intuitive. The offenses are so prevalent that it is not worth my time putting in the time to learn"*.

Some participants (T4, T11) reported having checked for controls, yet did not make changes to their settings. While T4 did not understand the controls, T11 felt that the trauma of his experience was too much for him to continue using VR.

Activation of Safety Controls. Participants used safety controls both proactively and reactively [23], depending on previous experience of harassment and when they learned about

them. Muting was the most predominantly used control (12), followed by blocking (7) and ghosting (3). The use of quick travel, vote kicking, and parental guidance was mentioned by one participant each. Some controls had several flavors based on the context, for instance, "mute" could be "mute self," "mute others," or "mute all."

T5 talked about the dependence of certain VR games (e.g., 'Beat Saber') on external platforms (Discord) and how controls from those platforms need to be used to mute people while playing the VR game: *"Beat Saber is multiplayer, but there's no voice chat in the game. You have to rely on Discord servers to get into a multiplayer lobby with players, [but there] it's mostly 10-year-olds saying [the] N-word."*

Some participants had clear preferences for using one safety control over another. T5 preferred blocking to muting: *"I'm not going to disable the voice chat in-game [Beat Saber] because it's not everyone [who harasses]. It's only a few people, and blocking them usually helps".* T8 described his preference for vote kicking in 'Pavlov VR': *"You can mute and also vote kick people. You can call a vote, and if it gets enough votes, the person is kicked from the lobby. It is a necessity to hear everything people are saying, so I don't mute."*

Of the 23 cases where participants were aware of the safety controls in the apps, in 19 of them, they were enabled, while in four, they were not. T12 did not use safety controls as she did not think they stopped the harassment:

*"If you mute or block them, it's not going to stop the harassment. It's just going to stop me from being aware of it. If they stick their cr*tch in my face, others can see that even if I can't."* (T12, Echo VR)

T17 added that safety controls were implemented only in the lobbies and did not stop harassment while playing 'Echo VR'. T8 found that it was faster to leave 'VRChat' than to individually block every offender.

Usability of Safety Controls. Among 23 instances where participants were aware of safety controls and one instance where they learned about them after a harassment incident, in 14 cases, they found the process of enabling the controls to be easy. In four cases, they found it to be cumbersome and challenging. Moreover, individuals' perceptions of ease of use varied considerably for the same set of features in a VR app. For example, T15 felt: *"The in-game [Echo VR] features to mute or ghost are very simple. You bring up your menu and select the individual or choose the easier option to mute or ghost all".* In contrast, T12, also a long-term user of 'Echo VR,' said: *"There is a block, but it's hard to use because you have to be able to point at the avatar, which is difficult sometimes".* T16 recalled how it was hard to find the mute button on 'RecRoom' as a new user:

"You had to go into some sub-menu, [with] all the people in the lobby listed, and find their name and click on it. The names are not always super easy to see if they're moving around." (T16, RecRoom)

Participants also noted that specific controls were more usable than others. T8 found it easier to mute himself than to mute others on 'VRChat': *"It's easy to mute yourself. So that's what I ended up doing most of the time. Blocking people isn't necessarily difficult, but not as fast"*.

Effectiveness of Safety Controls. Participants found safety controls to be useful in certain situations, and we find that the effectiveness of safety controls is highly context-dependent. In social VR, users perceive muting to effectively filter out inappropriate comments and overcome verbal disruption. In gaming VR, which shares the culture of trash-talking with other forms of online gaming [73], users state that muting is effective. T16, who muted his offensive team player on 'Echo VR' said: *"Even if the person said horrible things, I assume they still want to win the game. So they're not going to come over and intentionally play badly"*.

Participants used blocking to effectively limit further interaction with the harassers, especially when there were only a few of them to deal with. They also found quick travel to be effective in escaping from the harasser:

"They don't know where you're going, so they can't chase you around." (T7, Zenith)

Ineffectiveness of Safety Controls. Although participants found the use of safety controls effective in a few scenarios, in a vast majority of harassment incidents, they found the controls lacking in several ways. Social VR users felt that safety controls affected social interactions with non-harassers. T16 noted that enabling safety controls *did not provide feedback to the harassers*, and thus did not stop their behavior:

"I don't think that the game tells them that I have muted them. So they would have no feedback [that] this person can't hear me." (T16, RecRoom)

T7 complained that blocking did not remove the harasser from the game and only changed their appearance on 'Orbus'. She added that quick travel disrupted her experience on 'Zenith': *"You're doing something, and you have to stop what you're doing. It's like you're being punished because you're the one being harassed"*. T5, who used 'Beat Saber' and connected to Discord for voice chat, explained how blocking users on the game did not mute them on Discord, and had to use the safety controls on multiple platforms: *"Because blocking the person doesn't mute the person on Discord, I'll just mute the person"*. T13 expressed how certain safety controls would affect communication, which was essential in a strategic game like 'Echo VR': *"Muting players during the game makes it harder because it's a team game, and you can't communicate if you're muting people"*. T9 felt that none of the safety controls would be effective as the harasser could simply create a new account on the app: *"If the guy creates another profile, the blocking and muting would be in vain. You'd have to block this new avatar now. Back to square one"*. T13 and T18 also felt that safety controls would not stop the culture of abuse on 'Echo VR':

"They have zero effect on the larger culture of abuse. They're a testament to the failure of the larger structure to have any form of protection or any meaningful way to stop or intervene in whatever negative dynamics are going on." (T18, Echo VR)

T8 explained how safety controls could be misused, as described in Table 3. T14 felt that the use of safety controls could not prevent disruption when a new user joined the room on 'BigScreen': *"If someone new comes in and doesn't have that person blocked, then you'll have someone being like, oh, who is the screecher? And everyone has to be like, it's this person, block them"*. T14 also felt that the safety controls were ineffective in the unique context of 'BigScreen':

"You can pull up the usernames, but it'll be every username sitting in a theater. If you see someone screeching, you can see a microphone going off. But they'll be quiet when they're doing hand slapping. So you can't see who's talking. Sometimes they will get underneath the seats and slap your hand where you can't see their username. So everyone at that point has to stop the movie and find the offensive person." (T14, BigScreen)

Takeaways: Muting is the most used safety control, in line with trolling and profanity, which are reported to be the most dominant types of harassment. Blocking and proximity settings that help mitigate bullying and sexual harassment are the second most frequently used controls. The key usability challenges with safety controls arise when selecting them from a dense hierarchy of menus, while identifying the offending user's name from a long list of lengthy usernames (often with many special characters) or pointing at an offending user's avatar in order to take action on them, while they are still moving. Although safety controls provide a temporary escape to the target, they affect communication with non-harassers, fail to provide feedback to harassers upon muting or blocking, do not remove the harasser from the game but merely change their appearance, and could be misused to cause further harassment. Additionally, they fail to prevent others from witnessing the incident and stop further instances of harassment.

4.3 Perceptions on Reporting

After asking participants' experiences with safety controls, we also asked them about their experiences with using the reporting mechanisms in VR. Reports can be made in-app, via the headset, or through the website depending on the VR app and platform in question. We were particularly interested in the reporting process as it is a multi-stakeholder process involving not only the users (unlike the rest of the safety controls) but also involving developers/moderators and in cases, automated toxicity detection systems. Since reporting often involves feedback to the reporting/reported users, we investigated participants' satisfaction with the reporting systems in the VR apps they used.

We find that participants only submitted reports in half of the cases. Their failure to report was either due to a lack of reporting mechanisms, knowledge of reporting, or evidence. Some users did not want to spend effort on reporting. In cases where users submitted reports, their ease of reporting was dependent on the type of evidence they needed to provide.

Why Users Do Not Report. Of the 35 cases of harassment, only 16 cases were reported. Several participants did not know how to report. T16 did not know the controls to report and preferred to leave the app ('RecRoom'), while T7 perceived reporting to be difficult and did not want to "mess" with the process of reporting:

"I'm not sure how you report someone in Zenith and also I just didn't feel like messing with it." (T7, Zenith)

T8 and T16 both mentioned that 'Pavlov VR' did not have the feature to report and expressed their lack of faith in the developers to take necessary action. Some users who were new to the app assumed that the process would be cumbersome based on their experience with other apps. For example, T7 recalled her experience using the in-app keyboard on 'Orbus' and assumed that the process would be similar on 'Zenith':

"I'm sure it would be like Orbus, where you have to fill something out in the game on a keyboard. It's really cumbersome to do that. Plus, they were being very annoying, and I didn't want to stay long enough to do it." (T7, Zenith)

In a few cases, participants reported that it did not occur to them to report. Some participants also failed to report as they believed that the existing policies would not consider a certain type of harassment as a violation. For instance, T5 ('Beat Saber') felt: *"You can't do anything against him because he's technically not breaking any rules"*. Several other participants echoed the notion that reporting did not have any effect. T18, who was banned for 24 hours on 'Echo VR' due to a retaliatory report by his harasser, expressed outrage:

"That incensed me even more. I got banned, and it's only for 24 hours. You expect me to experience the negativity, to videotape the negativity, to go out of my way and submit the report to you. And then you're going to kick them out for 24 hours or 48 hours? How much of a punishment is that? That's a total waste of my time." (T18, Echo VR)

In some cases, even when the user knew how to report, they could not because they did not have evidence. T17 said: *"You have to provide footage of the incident, and I didn't"*.

Ease of Reporting. Our participants submitted reports through the in-app button, headset, or an external website affiliated with the VR app. Six participants reported that they found the reporting process to be simple and easy, and we note that in all of these cases, the only information they needed to provide was a description of the incident or choose the type of incident from a drop-down menu.

T7 found using the in-game keyboard on 'Orbus' for reporting cumbersome and frustrating. T17 felt that reporting was

complex as she could not capture everything her harassers said before she started recording. T13 noted that reporting could be difficult for beginners, but it might get easier with practice. T14 added: *"The only reason I know it so well is because I do it so much. I put time into the community, and I'm on all the Facebook and Discord groups. If a new user were to do it, it's not so clear how to report somebody [on Echo VR]. It's not part of the tutorial at all."*

Of the five cases in which the user had to provide a video of the incident, four participants found the process difficult.

"Each report takes 15 minutes if you're doing it properly. You have to get off the game, go through your footage, do a small edit of it, and write out the email. Unfortunately that's why many people don't do it." (T15, Echo VR)

T12 said that reporting could take five minutes to a week, and highlighted some challenges with headset-based reporting:

"On Quest 2, you click a report button, and you can record a snippet or upload a video. It'll search for the person, and you just send in the report, which is relatively easy. But you can't shorten it. You can't do that in the headset, you need to go to your computer." (T12, Echo VR)

How Reporting was Handled. For 11 out of 16 reports, participants believed a moderator had processed their reports. In five of these cases, they were automated responses acknowledging the report. In some cases, the participants assumed that an action had been taken when they did not see the harasser's account anymore (T10, 'Second Life') or did not encounter the type of content they reported anymore (T6 and T11, 'YouTube VR'). Four participants received a response specifying whether action was taken.

In nine cases, participants were satisfied with the responses, while in the other two, they were not. T13 felt that the systems were underdeveloped and human resources were lacking:

"I reported to Meta, and I got an email saying that it does not breach their terms of service before I got an email saying, 'we will review your complaint'. So, it's not exactly encouraging." (T13, Echo VR)

T14 added that while he was satisfied with the platform's actions, he expected more from the app developers: *"With Meta, [I'm satisfied]. I think that more things should get action taken than does. But I wish Echo VR, the company, was taking action and didn't just pass it off to Meta"*.

T16, who did not hear back from the moderators of 'Echo VR,' found reporting to be disincentivizing as he had *"no idea whether it's actually doing anything"*. He believed it was essential for the feedback to include what action was taken and what abilities had been restricted or revoked for the harasser. He also suggested not including the harasser's username in the feedback in order to protect their privacy. T2 added that reports should be taken seriously, considering the target's mental health, while T3 wanted moderators to enquire about the well-being of the user who reported.

Takeaways: If users do not report, it may be due to a lack of reporting mechanisms, knowledge of reporting, or evidence. Although users might find reporting easy once they get acquainted with the system, some aspects of existing reporting mechanisms, such as entering text through a keyboard while in VR or capturing video evidence, are cumbersome. Moreover, users expect timely responses to their reports, with feedback on the action taken.

4.4 Users' Expectations for Safer VR

At the end of the interview, we asked participants about their expectations for safer VR experiences. Participants had insightful suggestions for improving safety controls and new ways of tackling harassment in VR.

Live Moderators in VR. Participants indicated a need for real-time assistance in various situations. T11 said: *"There should be a guide, or an assistant to contact in case things go sideways"*. T17 wanted live moderators who were regular users capable of flagging those who violated the terms of service. T15 described moderators as *"in-game security guards,"* while T12 compared them to the police:

"99% of the time, most of us don't see a police person. But if we call one because we really need one, they come. If you could just push a button and have a person called to you, [they] can come and assess the situation." (T12)

T18 expressed outrage at the notion of social spaces without police officers and emphasized the need for social accountability. T15, a leader in a community for VR gamers (*Virtual Reality Party League*), specified actionable ways to promote live moderation with the help of VR community leaders:

"Have community leaders be involved in the main social aspects of games. Whether that is community leaders becoming mods or creating community members who want to step up and become mods. On the back end, have it built properly so they are properly trusted and educated on what they're supposed to do [with] their tools." (T15)

He also suggested offering perks to the moderators or having it as a paid position: *"They might get a hierarchy rank in that development team. They're allowed to go to certain events. And they're really there for the idea of that community growth. Or it could be paid, you have to be on this headset hour by hour to cover this block and we'll pay you the sum amount"*.

Tracking Users' Behaviour. Participants suggested that VR apps and platforms track VR users' behavior. T1 felt that having access to every user's history was paramount, primarily to determine what action should be taken against them: *"When they ascertain that this person has a history of harassment, [they] can give the person a warning first, and if it continues, just block the person completely, and make sure that the person will not have access to the platform again"*. T12's wanted users' behaviors to be tracked such that they had a certain "trust" level in social spaces inside an app:

"It's good to have certain levels where you trust people or give them abilities or access to places, depending on whether they've earned that. If somebody has been reported for harassment, I'd be booting them right back to level one. They should have to earn back the privilege to be 'normal' again." (T12)

T13 expressed that if a user was flagged for being toxic in one VR space, their "toxicity" should "follow from one app to another". He felt that the transfer of repercussions across apps was essential for a fundamental shift in the behavior of people online. He added that it had to be implemented at the platform ("Meta") level, with a coalition among popular VR apps: *"The 10 most popular VR MMO games go, we're going to combine efforts and if you get three strikes combined in any of our games, you can't play any of our games anymore".*

Detect Distress in Users. Some participants wanted their headsets to detect when they were feeling distressed. T11 felt that VR headsets should be sensitive enough to determine when a person needed help: *"It should tell when a person needs help when he or she no longer feels comfortable in the game".* T3 added that body movements such as rapid blinking of the eyes or sensors in the headset could be used as indicators of distress: *"When I'm distressed, it gets this information and shuts down. If I am talking to you and I start blinking rapidly, it should know, this person isn't okay, and slow down things".*

In-app Interventions. Several participants indicated their preference for having in-app interventions such as disclaimers, warnings, or prompts to inform them before engaging in potentially toxic environments. T2, who encountered offending scenes on 'YouTube VR', wanted a disclaimer about the contents of a video. T11 suggested the inclusion of a prompt that would ask the user if they wanted to exit an app when harassment was detected: *"If your headset can detect [harassment], there should be a prompt [asking] if you would want to leave the game you are in".* T8 and T9 wanted features that would warn users before they entered social spaces where they might encounter harassers:

"If somebody is playing in a certain lobby and they'll know you, recognize you, and get angry at you again, I don't think the game should allow you to join it, or tell you who is in the lobby before you join it. It should have a disclaimer that says you've tagged this person." (T8)

T9 added that users joining areas with problematic users (e.g., trolls) should get a notification warning them.

Segregation of Users. Participants had concerns about the safety of kids and wanted age-based restrictions for VR usage:

"VR shouldn't be for kids, honestly. Oculus is already requiring apps to remove 13-year-olds on it. I don't think they're going to stop using Oculus, they're just going to create a normal account and start communicating." (T5)

T18 added, *"No children should be allowed to play with adults. Period. There needs to be age segregation, [because] there are adult predators transgressing boundaries of morality".* T9 echoed the notion of segregation and suggested that the entire app be divided into age-brackets:

"There would be a bracket [with] young kids, another intermediate bracket, [with] teenagers. Adults, somebody will have to agree to some terms and conditions, if you're joining this section, know you may experience this and this. In other sections, all of that is banned." (T9)

T14 felt that paid users could be segregated from unpaid users to filter out children: *"It would be cool if I could pay money to only play with people who paid money because it would get rid of a lot of screechers".* T5, who had negative experiences with child users saying racial slurs and destroying virtual artifacts inside games, wanted adult-only lobbies. He also believed that checking IDs could be a way to enforce this:

"I'm not saying that kids shouldn't be able to enter, just that there should be a way that adults can just stick with each other, maybe like ID check." (T5)

Takeaways: VR users believe that live moderators and age-based segregation would significantly reduce harassment in VR. They also recommend automatic detection of harassment situations and tracking users' toxicity histories across VR apps. Further, they want in-app interventions to inform them of harassers in the vicinity or in VR spaces they enter.

5 Developers' Perceptions on VR Safety

Based on our findings from **Study-I**, we conducted **Study-II** (§ 3) where we interviewed (n = 9) VR developers. We asked them about the feasibility of the safety features desired by our users (§ 5.1), challenges in designing and deploying safety controls (§ 5.2), and ways to improve existing controls (§ 5.3). With professional developers, VR-based researchers, and hobby developers represented by our participants (Table 6 in Appendix A.2), we believe our findings provide an extensive view of the challenges and limitations in this domain.

5.1 Feasibility of User's Expectations

One portion of our interview protocol was about eliciting developers' perceptions of the features desired by the participants of **Study-I**. We started by asking the developers' perspective of what was needed to make VR safer, and followed it by presenting the main user-desired features — live moderators in VR, tracking users' behaviour, detecting distress in users, in-app interventions, and segregation of users — and asking about their feasibility. Since segregation of users largely stemmed from users' desire to have identity or age-based segregation, we asked developers about the feasibility of identity verification.

Live Moderators in VR – Feasibility. Developers largely agreed that having live moderators in every social space was

infeasible due to the human resources required, the financial challenges, and the difficulty in scaling. D8 said: *"That wouldn't be very scalable, especially for teams like mine where we only have three people developing"*. D7 argued that the economic model would not work, with D9 adding: *"a small development studio, [with] six people and 100 servers couldn't fund that many people to do that. Even a big company who could do that, [would] have to fund thousands [of] people to listen in on every conversation that's said"*.

To tackle these issues, D3 suggested leveraging the VR community to ease the burden on human moderation, echoing T15: *"some platforms have community outreach, where respected members of the community can act as deputies of sorts"*. Drawing from his experiences in community management for his VR game, D7 added that live moderation was taxing and could be effective long-term only if moderators were part of a community: *"They have to get something out of the experience in the first place, or it's just something that burns people out"*. D3 advocated using a small, well-trained team to ensure consistent moderation practices. Multiple developers echoed that moderation at scale could be done using AI-based abuse detection but also recognized the inherent technical and privacy challenges:

"Is it practical? How will people react to it? These are the questions that we need to be asking because we're already doing voice, but what about content and avatars? When are we going to have models that look at a 3D mesh and identify if there's any perfect content on it?" (D3)

Tracking Users' Behaviour – Feasibility. Developers argued that tracking users' behavior was feasible from a technical standpoint; however, they were not in favor of its deployment. D1 pointed out that if a user violated the terms and conditions of a VR community once, it may not be fair to exclude them from other apps, especially if they corrected their behavior. D9 added: *"If you connect every app ever and you're toxic on one app and it now spreads to everything, obviously people wouldn't want that. You could also have silent toxic people that just lower your rating"*. D3 further highlighted the importance of ensuring that users did not get a negative impression of a user before interacting with them: *"I would not put [the rating] in something which is always stuck to your head. But if I made the conscious decision to require more information about [the user], it's present"*.

Developers highlighted potential challenges in implementing this technique. D8 believed that the tracking could happen only at the platform level (such as Meta, Steam, etc.), due to the existence of several VR stores, each using different accounts. D6 added that there must be an incentive for users to use the same profile on all apps/platforms. However, they also found this solution to be privacy-infringing:

"As a user, I don't like to be pervasively tracked by corporations. Whether or not they're providing me tangibly usable tools, they're not trustworthy entities." (D7)

D2 argued that, while the data could be anonymized if it was used for training an AI model to detect harassment, real-time tracking involved legal challenges. D3 wanted VR companies to be transparent about data collection practices to users: *"Transparency on the company's part is necessary to reaffirm faith in users"*. Despite their thoughts on how this solution compromised users' privacy, developers also perceived the value of tracking users:

"It's good to have that unified information. As a developer, I would love to have a jump list of problematic users and an ability to just, at the very least, pre-sort them into [a similar] category of servers." (D7)

Detect Distress in Users – Feasibility. Our participants agreed that the automatic detection of distress in users to turn off the headset or quit a VR app seemed a good solution, but raised concerns about the availability of sufficient input for accurate detection:

"We'll have to be able to break down the feasibility in terms of being able to understand when any stress is happening, how to interpret it, in order to implement it." (D2)

D4 believed there could be many false positives, where emotions such as excitement may be wrongly detected as distress. D6 added that false positives would result in a *"terrible user experience"*. D9 suggested that obtaining feedback from the user upon detecting distress might achieve a good trade-off between model improvement and user experience.

In-app Interventions – Feasibility. The developer participants agreed that having in-app interventions was *feasible* solution. D1 suggested the use of voice-activated commands for interacting with pop-up notifications and encouraged designs with minimal visual interaction:

"The less you have to disturb the user in terms of having to do manual input, that's good. You can [use] pop-ups and [they] just [have] to click, that's useful. It can even be voice-activated commands [without the] user having to interact too much with hand controls." (D2)

Identity Verification – Feasibility. Since our user participants wanted to segregate users based on age or identity verification, we asked the developers about their perspectives on implementing identity verification. They felt that users would not be willing to share official identification as it was invasive, compromised anonymity, and posed risks in cases of data breach, while failing to address the problem:

"Someone who's someone is immaterial to whether or not they are a problem. Having their government ID is just putting more power in the hands of tech platforms." (D7)

D1 and D2 suggested the integration of social media profiles to VR accounts; however, such initiatives have received backlash in the past, as described by D9: *"[Meta] were trying to connect everything from Facebook, Instagram, Oculus, and they got in trouble for doing that"*. D2 and D5 recommended

the use of hardware identifiers or IP addresses for detecting a user. D2 suggested third-party verification, but also highlighted challenges in complying with different technology laws across countries. D3 shared that some VR clubs required real-life identification to partake in the community:

"Some communities have third-party services to check if the person's real. And that can vary. You take a selfie with your ID, let your personal information out. Or it can be a Discord bot that takes a picture of your driver's license, talks to a background checking service." (D3)

D5 suggested computer vision techniques for face detection to predict a user's age, while D6 recommended using physiological attributes such as movement and posture for inference.

Takeaways: Sustainable live moderation in VR communities may rely on moderators embedded in the community and augmented with AI-based abuse detection. While tracking users would enable VR developers to create a block list of problematic users, it may be limited to a single platform (e.g., Meta, Steam, etc.) and pose legal risks to VR companies. Detecting distress in users requires a thorough evaluation of whether VR systems can access sufficient input for accurate detection. Developers advocate third-party verification for non-invasive user identification and propose in-app interventions.

5.2 Challenges in Designing Safety Controls

Several questions in our interview were about extracting the challenges in the development of safety controls. We asked developers whether and when safety control design appeared in the pipeline of the apps they developed, and what according to them was hindering the development of effective safety controls in VR apps. We also asked them if they believed an industry standard of VR safety controls could be created, and what technical, social, and economic barriers they perceived.

Safety Not A Priority. A majority of our developer participants believed that VR platforms and app developers do not currently prioritize safety. D3 explained that moderation in VR tended to be reactive rather than proactive, as the development efforts would be focused on appealing to the investors and clients. D4 and D8 felt that VR developers usually had small teams and did not have the bandwidth to develop features for safety. D9 added that developers would not implement safety features until their users reported issues:

"Companies are very money-first, fix later. [Maybe] this is why it's getting pushed off and not many people are talking about it or fixing it." (D9)

Several participants agreed that there was a lack of financial motive to prioritize safety. D7 highlighted the need for tech companies to justify developmental efforts, especially in light of recent economic conditions with mass layoffs: *"the primary challenge is justifying all that development effort, when that could be put towards new user acquisition or directly monetizing engagement"*. D7 added that big companies did

not stand to lose out monetarily due to the users that stopped using apps due to harassment:

"You can always try to acquire new users faster than you are losing older users. If [companies] can compensate for X women who are leaving, and it's just awful being a woman online period, but have a referral program that gets more teenage boys to recommend the platform and they get a \$3 kickback and that makes their numbers go [up], then [they] don't have a problem." (D7)

Participants also believed that well-implemented safety controls may limit interactions among users, impact overall app engagement, and disincentivize companies to focus on them:

"Platforms wouldn't wanna do that because engagement is easiest to generate via conflict. If you actually allow people to peaceably exist separate from each other, there are repeat logins that people don't end up doing." (D7)

Lack of Guidelines/Standards. Developers argued that they did not have any guidelines for ensuring safety in VR apps and stressed the importance of creating awareness among developers to design for safety. For instance, D5 shared:

"We don't have a list to go through [that] says you should have [these] models in your app for safety. I'm not sure if it's already there. But clearly, it's not in my mind. It would be nice to have clear regulations to follow." (D5)

When asked about the feasibility of creating an industry-wide standard for VR safety controls, most of our participants foresaw challenges in bringing together the major stakeholders in the ecosystem. D2 felt that effective enforcement of a standard required coordination among all the big companies, while D7 illustrated how companies tended not to cooperate:

"Apple goes out of their way to not use any of the same words to describe things. They just announced a VR headset without using the words virtual reality once. It's anti-theoretical for them that there's some collective best way to handle user moderation. It had to be forced by the EU to put a charger on their phone. That tells how much these corporations want to cooperate on anything." (D7)

D2 illustrated the technical challenges in implementing uniform data collection practices across platforms: *"To implement such a standard, the way data is gathered in application A should be similar to application B. You have to develop [a] middle framework capable of [collecting] this data and parsing it"*. D6 felt that enforcing entities to adhere to standards was hard. D5 believed non-profit organizations could facilitate companies working together.

Practical Challenges. Developers identified several technical and logistical challenges in the design and deployment of effective safety mechanisms for VR. D2 felt, *"it's complicated to prove that someone is harassing you because it's through internet, you have the VPN, can hide your IP. It's tricky to prove that someone is harassing you"*. D7 stressed the importance of recruiting good moderators for his VR game:

"You need effective moderators. People who are aware of the rules, who see eye to eye to you about what they mean, why they exist, what your goals for the community are, who you wish to include and exclude in that community, and willing to put in the hours being present. They shouldn't just feel like silent overlords because that can build resentment patterns between user bases." (D7)

An inherent challenge about safety controls that developers called out was that they offloaded a lot of responsibility to the user, as users needed to remember the right controls and use them at the time of need. This is further exacerbated by complications in usability testing, specifically in replicating stressful situations to test how usable the tools are:

"The biggest barrier is actually simulating those experiences to see the tools that are actually working." (D8)

D7 also emphasized the complicated interface development work involved in synthetically recreating reality for VR apps.

Compromise Privacy. Most of our participants expressed concerns about using methods that infringed privacy for the sake of safety and questioned the necessity of invasive practices for moderation. D3 said, *"most people are turned off by the idea of [something] constantly monitoring them, possibly storing their conversations on a database somewhere for an extended period of time"*. While discussing ways to collect evidence for reporting, many developers suggested the using a video buffer that stored the recording from a specific time period (e.g., the last 10 seconds); however, all of them agreed that it was privacy-violating:

"Say the platform has the recording of everything in the last 24 hours. The user can always go back to see what happened around [them]. So everything will be recorded, but I don't know how that will conflict with privacy." (D5)

Takeaways: VR companies tend not to prioritize safety due to a lack of financial incentive and high development costs. Further, VR developers lack awareness about safety risks in VR and may not have legal or technical guidelines for safety design. Developers also highlight challenges in simulating VR safety risks for user testing of the safety controls and raise questions about balancing privacy with safety.

5.3 Improving Safety in VR

We asked developers about the similarities and differences in dealing with harassment in VR when compared to traditional forms of social media. We then asked them what they perceived to be lacking in safety controls, and how they may be improved. We also presented the usability challenges from **Study-I** and asked developers how they would solve them.

Abuse Detection. Most of our developer participants agreed on the challenges involved in moderation at scale and recommended using several abuse detection methods as the first step in the moderation life cycle. D3 believed, *"there's certainly the chore that is the actual implementation, once those*

systems are in place, they're largely autonomous. I don't need to touch it.", and explained how word analyzers may be used to identify verbal abuse and notify users:

"Using semantic analyzer, we have been able to identify the intent of what someone is saying. If we identify that as harmful or having foul language, we can give the user a reminder; like, hey, you should be watching your language. If it persists, we can take moderation action." (D3)

D2 added that a common list of bad words may be detected and used to notify relevant users: *"We're thinking of being able to remove some words from the usage of the users"*. For detecting other forms of abuse in a non-invasive manner, D6, a VR-based researcher proposed the use of logged events:

"I would keep track of user events that can occur, positions of their avatars, the proximity of different avatars. If I have that record, even without video evidence, I would know if they were in proximity, if some user events occurred. If I have the username reported, I can see if they were in those proximity and had the opportunity to interact." (D6)

Effective Grouping of Users. Drawing on his experience running a VR game, D7 shared that effective community management with like-minded users grouped together prevented conflicts, reinforcing users' preferences from **Study-I**:

"One of the bedrocks of sustainable community management is about establishing a set of norms and expectations and creating an environment that funnels those correctly so that you are effectively grouping users together who share expectations. And if they're sharing expectations, you don't get people in conflict [like when] one person thought they were here for X and another thought they were there for Y and they're now fighting over that." (D7)

Developers further elaborated on interaction filtering and social graph pruning to group users. D7 said: *"I'm in party mode. I hit a button and it makes my social protocols in terms of who can talk to me and this becomes broad momentarily because I trust the people running this party. Then I switch to a different virtual space and I hit a button, that changes who can DM me for a private chat, who appears on my screen"*. He added that all connections of specific users should be prevented from ever interacting with another user to effectively cut off problematic social graphs.

Solving Usability Challenges. To solve some of the usability challenges that our participants faced in **Study-I** (pointing at a user in virtual space to block them, choosing a user from a long list of usernames, typing on a keyboard in VR), developers proposed several potential solutions, and generally advocated for designing intuitive controls that felt natural to use. D3 believed that blocking tools needed more granularity, providing the capability to block different aspects of a user, such as *"their voice, avatar, ability to scale, and other atomic elements that users can have fine-tune control over"*. D7 added that users should be able to block other users for

different periods of time: "you should be able to shut someone up for 10 minutes if they're just being annoying".

To effectively identify users for taking action, D4 recommended designing ways to pause the scene and recalled how Meta Horizon Worlds implemented this: "you just hold up your arm and press a button. It immediately freezes everybody, but you still see other people in the room. If someone is harassing you, you can pause everything, report [or] block them". D7 added that the list of users can be accompanied by a snapshot of their avatars:

"If that list is difficult to deal with, why is that list not sorted by distance to you physically? Or grouped into sets of bands? You're trying to identify someone like oh it's a guy with a hat and a coat. If you think about the way you would search a physical space in that context and mark people off, effectively modeling that cognitive process as a UI tool is a way to do that." (D7)

D8 also recommended implementing voice-based memos for actions such as reporting in order to avoid typing in VR.

Creating Awareness among Developers. Three out of nine developers who took part in **Study-II** argued that more awareness was needed among VR developers about harassment issues in VR. D9 expressed that independent developers needed to be made aware of safety issues and provided with the relevant resources for them to adopt a safety-oriented design:

"When I was developing, the last thing I was going to add was safety features. So bringing more awareness to the little guys, the small developers, giving them resources that allow them to integrate these things quickly..." (D9)

D8, who was part of a three-member development team, argued, "if it was a big enough issue, even though we're a small team, if it was brought to our attention, probably have a sprint to focus on that, at least one of us".

Open-sourced Safety Libraries. Several participants shared that open-sourced libraries available through popular game engines that allowed developers to integrate standard safety controls into their apps would be valuable. D9 said: "if you had a library you could pull from, easy implementation. Especially if it just works out of the box". He further added that if it was open-source when somebody found a way to break it, the community would be able to fix it. D2 added:

"It should be a plug and play tool that could be brought to the game engines, and you just drag and drop." (D2)

D8 pointed out the importance of having well-established requirements in app stores for targeting efforts towards safety: "If some stores have requirements where your app needs [certain] features to be admitted, that would force all the devs to have at least something to protect its users". D1 and D7 stressed that seminars and workshops may be organized for developers to brainstorm on creating such tools. D5 also suggested creating standardized tutorials for VR users, informing them how to protect themselves.

Takeaways: Using word filters and semantic analyzers may aid in abuse detection. Developers recommend implementing extensive controls to limit or allow user interaction in various settings. Granular controls allowing users to restrict interaction temporally and spatially address certain limitations of existing safety measures. Seminars and workshops can be conducted to inform developers about VR safety and foster innovation in safety design. Additionally, integrating open-sourced safety-focused libraries with popular game engines can streamline the development of safety controls.

6 Discussion

In this paper, we outlined our findings about VR safety features based on a multi-perspective study with 18 targets of VR-based harassment and 9 VR developers. In this section, we discuss the broader findings from our study and offer recommendations to help VR platform owners, app developers, and policymakers enhance VR user safety.

What's New With VR-based Harassment? Much like various other manifestations of online harassment, VR-based harassment exposes users to harmful content, including bullying, threats of violence, and sexual harassment [29]. While certain forms of harassment may be universally applicable across different mediums, the distinctive features of VR applications give rise to unique manifestations of harassment. For instance, incidents such as virtual sexual harassment, where the harasser positions their avatar's "crotch" on the target's face, and virtual violence, where a harasser can manipulate the target's avatar to cause disorientation, illustrate the specific challenges posed by VR, as outlined in Table 3.

VR has parallels with other forms of online interaction such as gaming, so it follows that many solutions that apply to the latter (profanity filters, reporting systems, behavioral analytics, etc.) apply to the former. However, implementing the same solutions effectively in VR requires significant redesign to accommodate the three-dimensional and immersive nature of interaction that lead to more traumatizing forms of harassment (e.g., virtual sexual harassment). Further, safety controls are not equally effective across VR contexts, particularly while removing harassers from gaming and streaming settings, highlighting the uniqueness of safety design for VR. Proposed solutions have included robust reporting mechanisms [23], non-player characters as safety companions [18], and consent-based designs [61, 74] for boundary settings. We add to this by identifying ways to improve safety in VR — keeping the bad actors in check while minimizing load on users, and enabling developers to implement these solutions.

Minimizing Load On Users. With the burden for staying safe online currently falling on users [30], the fact that many participants from **Study-I** were unaware of VR safety controls calls for improving user awareness about safety controls. Apart from making the information available in-app as tutorials, nudges to users, and displaying cues in VR in the style

of "Madison Avenue" advertising [75], we also recommend that VR app/platform owners inform users through official websites, app stores, and social media campaigns. However, as our findings highlight, understanding the usage of safety controls is a time investment that VR users must be willing to make. A good design of safety controls should make this learning process seamless and consider particular contexts in which users would employ them, such as gaming or streaming settings. Standardizing the baseline safety controls across the VR ecosystem may help, as participants report that safety controls get easier to use with familiarity.

Stopping Bad Actors. Stopping bad actors requires automated or human-driven identification of bad behavior and enforcement of sufficient, non-biased punitive action. The action may range from warnings and penalties to bans that prevent harassers from returning to the platform [5]. While several ML-based techniques have been developed to detect obscene imagery [76] and abusive language [77] there is a need to direct efforts towards detecting a variety of toxic content from a 3D environment.

Since user reports play a major role in VR moderation, designing usable reporting mechanisms with multi-modal availability is critical. Again, standardizing the reporting features across VR would reduce the burden on users. Effective moderation requires VR apps/platforms to have diverse, well-trained moderators in the community. AI-based moderation techniques [78] may be considered for prompt processing and providing timely feedback to the reporter. Models of norm enforcement such as *responsive regulation* have been identified in the literature [5], where the penalties are proportional to the severity of the offense and the intent of the perpetrators.

Identifying cases of false reporting and providing constructive feedback to the ones reported for corrective behavior should be an integral component of moderation. Prior work has proposed the inclusion of a "limbo" space where a reported user could be taken to, while the moderator provides them feedback about why they were reported [79]. Non-invasive forms of identity verification need to be implemented to prevent toxic users from returning to the platform in case of a permanent ban.

Enabling Developers. A key part of solving the safety problem is enabling VR developers. The design of safety controls requires innovation in usability as they need to be usable in three-dimensional space, in contrast to their counterparts in other online social platforms. As more and more inexperienced developers enter the burgeoning VR industry, awareness is critical to ensure a safety-oriented design of VR apps. Although the VR ecosystem consists of a diverse population of independent developers, gaming companies, and corporations, likely with conflicting priorities [5], technical and legal guidelines mandating standards for safety and data collection practices would ensure cross-platform compliance and reduce the burden on developers. VR-specific regulations for appro-

priate codes of conduct, tutorials, moderation systems, and the required set of safety features in VR platforms/apps would simplify the design process for developers. Open-source development of VR safety-oriented libraries would add immense value, particularly if integrated into game engines.

Considering Multiple Perspectives. Although VR users and developers had shared opinions in our study, such as using community members for effective moderation, developers expressed more concern about users' privacy compared to users, while thinking about safety design for VR. Although users recommended behavioral tracking across VR apps to ensure safety, by implementing these pervasive tracking mechanisms, VR app companies may be subject to legal scrutiny, and lose their users' goodwill in cases of false reporting. From **Study-II**, developers had conflicting views on when safety controls needed to be included in the development pipeline for a VR app: some wanted to design them at the start to be in tune with the affordances of the app, while others wanted to focus on the core functionality first. Although developers recommended safety-focused research, they also highlighted the challenges of simulating harassment scenarios for user testing. It is imperative to reflect on these contrasting perspectives to implement solutions that achieve a good balance of safety and privacy with ease of development.

7 Conclusion

Since VR serves as an emulation of the real world, it presents many of the challenges in society, perhaps creating a notion that harassment in VR is a societal problem. However, VR-based harassment is yet another instance where technology facilitates malicious social actors to thrive, similar to social engineering attacks such as phishing or robocalls. Thus, similar to developing solutions for phishing or robocalls, we must engineer robust solutions to address VR-based safety risks. Techniques such as AI-based abuse detection, semi-automated moderation, and identity verification raise questions about users' privacy; therefore, it is vital to identify practical, privacy-preserving solutions that serve VR platforms, developers, and users. To allow for anonymous VR interactions while creating safe spaces, VR apps may create tiered experiences, with some spaces having identity verification and others not. To prevent vulnerable populations such as children from partaking in potentially dangerous VR spaces, users' gait, height, and facial features may be used to run a model on-device to infer their age. Problematic and disruptive users may be removed from other users' views by pruning their social graphs. The onus is on major players in the VR ecosystem to come together to solve the problem of harassment in virtual environments.

Acknowledgment

We are deeply grateful to the VR users and developers who participated in our studies, as well as Dr. Vanessa Volpe, De-

partment of Psychology, North Carolina State University, who helped us refine Study-I. We extend our thanks to Elizabeth Lin, Greg Tystahl, Sathvik Prasad and Zahra Shiraz for their feedback on the interview protocols. We also thank our anonymous reviewers and shepherd for their constructive and insightful feedback, which helped improve this paper significantly. This material is based upon work supported in parts by a Meta Research gift award. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Meta Inc.

References

- [1] A. Kolesnichenko, J. McVeigh-Schultz, and K. Isbister, "Understanding emerging design practices for avatar systems in the commercial social VR ecology," in *Proceedings of the 14th Designing Interactive Systems Conference (DIS)*, 2019, pp. 241–252.
- [2] J. McVeigh-Schultz, A. Kolesnichenko, and K. Isbister, "Shaping pro-social interaction in VR: an emerging design framework," in *Proceedings of the 39th Conference on Human Factors in Computing Systems (CHI)*, 2019, pp. 1–12.
- [3] J. McVeigh-Schultz, E. Márquez Segura, N. Merrill, and K. Isbister, "What's it mean to 'be social' in VR? mapping the social VR design ecology," in *Proceedings of the 13th Designing Interactive Systems Conference (DIS) Companion*, 2018, pp. 289–294.
- [4] G. M. Garrido, V. Nair, and D. Song, "Sok: Data privacy in virtual reality," *arXiv preprint arXiv:2301.05940*, 2023.
- [5] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz, "Harassment in social virtual reality: Challenges for platform governance," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 3, pp. 1–25, 2019.
- [6] M. Kim, M. Ellithorpe, and S. Burt, "Anonymity and its role in digital aggression: A systematic review," *Aggression and Violent Behavior*, 2023.
- [7] M. Duggan, "The broader context of online harassment," Pew Research Center, 2017. [Online]. Available: <https://www.pewresearch.org/internet/2017/07/11/the-broader-context-of-online-harassment/>
- [8] S. Burke Winkelman, J. Oomen-Early, A. D. Walker, L. Chu, and A. Yick-Flanagan, "Exploring cyber harassment among women who use social media," *Universal Journal of Public Health*, vol. 3, no. 5, p. 194, 2015.
- [9] D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles, "Ethics emerging: the story of privacy and security perceptions in virtual reality," in *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS)*, 2018, pp. 427–442.
- [10] Z. Qingxiao, D. T. Ngoc, W. Lingqing, and H. Yun, "Facing the illusion and reality of safety in social VR," *arXiv preprint arXiv:2204.07121*, 2022.
- [11] W. Duffield, "A grope in Meta's space," Tech Dirt, 2021. [Online]. Available: <https://www.cato.org/commentary/grope-metas-space>
- [12] J. Belamire, "My first virtual reality groping," Medium, 2016. [Online]. Available: <https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee>
- [13] S. Frenkel and K. Browning, "The Metaverse's dark side: Here come harassment and assaults," *New York Times*, 2021. [Online]. Available: <https://www.nytimes.com/2021/12/30/technology/metaverse-harassment-assaults.html>
- [14] "Hate in social VR," Anti-Defamation League, 2018. [Online]. Available: <https://www.adl.org/resources/report/hate-social-vr>
- [15] "Metaverse: another cesspool of toxic content," SomeOfUs, 2022. [Online]. Available: https://www.eko.org/images/Metaverse_report_May_2022.pdf
- [16] B. Duranske, "Reader Roundtable: 'Virtual Rape' Claim Brings Belgian Police to Second Life," Visually Blind, 2007. [Online]. Available: <http://visuallyblind.com/2007/04/24/open-roundtable-allegations-of-virtual-rape-bring-belgian-police-to-second-life/>
- [17] "One incident of abuse and harassment every 7 minutes," Center for Countering Digital Hate Inc, 2021. [Online]. Available: <https://counterhate.com/research/facebooks-metaverse>
- [18] Q. Zheng, S. Xu, L. Wang, Y. Tang, R. C. Salvi, G. Freeman, and Y. Huang, "Understanding Safety Risks and Safety Design in Social VR Environments," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 7, pp. 1–37, 2023.
- [19] M. A. Lemley and E. Volokh, "Law, virtual reality, and augmented reality," *University of Pennsylvania Law Review*, vol. 166, p. 1051, 2017.
- [20] L. Tychsen and L. L. Thio, "Concern of photosensitive seizures evoked by 3D video displays or virtual reality headsets in children: current perspective," *Eye and brain*, pp. 45–48, 2020.
- [21] A. Stanton, "Dealing with harassment in VR," UploadVR (UVR Media, LLC), 2016. [Online]. Available: <https://uploadvr.com/dealing-with-harassment-in-vr/>
- [22] "Use the safe zone in meta horizon worlds," Meta, 2022. [Online]. Available: <https://www.meta.com/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safe-zone-in-horizon/>
- [23] G. Freeman, S. Zamanifard, D. Maloney, and D. Acena, "Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 6, pp. 1–30, 2022.
- [24] R. E. Boyatzis, *Transforming qualitative information: Thematic analysis and code development*. Sage, 1998.
- [25] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.
- [26] S. D. Hazelwood and S. Koon-Magnin, "Cyber stalking and cyber harassment legislation in the united states: A qualitative analysis," *International Journal of Cyber Criminology*, vol. 7, no. 2, p. 155, 2013.
- [27] C. Southworth, J. Finn, S. Dawson, C. Fraser, and S. Tucker, "Intimate partner violence, technology, and stalking," *Violence against women*, vol. 13, no. 8, pp. 842–856, 2007.
- [28] N. Lapidot-Lefler and A. Barak, "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition," *Computers in Human Behavior*, vol. 28, pp. 434–443, 2012.
- [29] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *Proceedings of the 42nd Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 247–267.
- [30] M. Wei, S. Consolvo, P. G. Kelley, T. Kohno, F. Roesner, and K. Thomas, "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment," in *Proceedings of the 43rd Conference on Human Factors in Computing Systems (CHI)*, 2023, pp. 1–17.
- [31] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Proceedings of the 17th Symposium on Usable Privacy and Security (SOUPS)*, 2021, pp. 299–318.
- [32] D. Freed, N. N. Bazarova, S. Consolvo, E. J. Han, P. G. Kelley, K. Thomas, and D. Cosley, "Understanding Digital-Safety Experiences of Youth in the US," in *Proceedings of the 43rd Conference on Human Factors in Computing Systems (CHI)*, 2023, pp. 1–15.
- [33] R. Bhalerao, N. McDonald, H. Barakat, V. Hamilton, D. McCoy, and E. Redmiles, "Ethics and Efficacy of Unsolicited Anti-Trafficking SMS Outreach," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 6, pp. 1–39, 2022.
- [34] A. McDonald, C. Barwulor, M. L. Mazurek, F. Schaub, and E. M. Redmiles, "It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online," in *Proceedings of the 30th USENIX Security Symposium (USENIX Security)*, 2021, pp. 375–392.
- [35] A. Strohmayer, J. Clamen, and M. Laing, "Technologies for social justice: Lessons from sex workers on the front lines," in *Proceedings of the 39th Conference on Human Factors in Computing Systems (CHI)*, 2019, pp. 1–14.
- [36] M. Almansoori, A. Gallardo, J. Poveda, A. Ahmed, and R. Chatterjee, "A Global Survey of Android Dual-Use Applications Used in Intimate Partner Surveillance," *Proceedings on Privacy Enhancing Technologies (PoPETs)*, vol. 4, pp. 120–139, 2022.

- [37] S. Stephenson, M. Almansoori, P. Emami-Naeini, and R. Chatterjee, "It's the Equivalent of Feeling Like You're in Jail": Lessons from Firsthand and Second-hand Accounts of IoT-Enabled Intimate Partner Abuse," in *Proceedings of the 32nd USENIX Security Symposium (USENIX Security)*, 2023, pp. 105–122.
- [38] S. Stephenson, M. Almansoori, P. Emami-Naeini, D. Y. Huang, and R. Chatterjee, "Abuse vectors: A framework for conceptualizing IoT-enabled interpersonal abuse," in *Proceedings of the 32nd USENIX Security Symposium (USENIX Security)*, 2023, pp. 69–86.
- [39] J. Fox and W. Y. Tang, "Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies," *New media & society*, vol. 19, no. 8, pp. 1290–1307, 2017.
- [40] L. McLean and M. D. Griffiths, "Female gamers' experience of online harassment and social support in online gaming: A qualitative study," *International Journal of Mental Health and Addiction*, vol. 17, pp. 970–994, 2019.
- [41] K. Shriram and R. Schwartz, "All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality," in *Proceedings of the 25th Virtual Reality Conference (VR)*. IEEE, 2017, pp. 225–226.
- [42] E. Deldari, D. Freed, J. Poveda, and Y. Yao, "An Investigation of Teenager Experiences in Social Virtual Reality from Teenagers', Parents', and Bystanders' Perspectives," in *Proceedings of the 19th Symposium on Usable Privacy and Security (SOUPS)*, 2023, pp. 1–17.
- [43] V. Angelov, E. Petkov, G. Shipkovenski, and T. Kalushkov, "Modern virtual reality headsets," in *Proceedings of the 1st International congress on human-computer interaction, optimization and robotic applications (HORA)*, 2020, pp. 1–5.
- [44] "Oculus store." [Online]. Available: <https://www.oculus.com/experiences/quest>
- [45] "Steam." [Online]. Available: <https://store.steampowered.com/vr>
- [46] "Sidequest." [Online]. Available: <https://sidequestvr.com/all-apps>
- [47] "Bigscreen," 2023. [Online]. Available: <https://www.bigscreenvr.com/software>
- [48] "Pavlov VR," 2023. [Online]. Available: https://store.steampowered.com/app/555160/Pavlov_VR/
- [49] "Vrchat," 2023. [Online]. Available: <https://hello.vrchat.com>
- [50] V. Sharma, "Introducing a personal boundary for horizon worlds and venues," *Meta*, 2022. [Online]. Available: <https://about.fb.com/news/2022/02/personal-boundary-horizon/>
- [51] "Safety and trust system." [Online]. Available: <https://docs.vrchat.com/docs/vrchat-safety-and-trust-system>
- [52] "Comfort and safety." [Online]. Available: <https://recroom.com/safety>
- [53] "Modulate," Modulate, 2023. [Online]. Available: <https://www.modulate.ai/tox-mod>
- [54] D. Maloney and G. Freeman, "Falling asleep together: What makes activities in social virtual reality meaningful to users," in *Proceedings of the 7th Symposium on Computer-Human Interaction in Play (CHI PLAY)*, 2020, pp. 510–521.
- [55] D. Maloney, G. Freeman, and A. Robb, "It is complicated: Interacting with children in social virtual reality," in *Proceedings of the 1st Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 343–347.
- [56] D. Maloney, G. Freeman, and A. Robb, "A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality," in *Proceedings of the 7th Symposium on Computer-Human Interaction in Play (CHI PLAY)*, 2020, pp. 472–483.
- [57] M. Khelif, "Virtual reality consistency with common concerns of humanity: an overview," in *Proceedings of the 7th Conference on Information Technology Trends (ITT)*. IEEE, 2020, pp. 218–223.
- [58] X. Deng and J. Ruan, "Users' privacy in the Second Life Library," in *Proceedings of the 2nd Symposium on IT in Medicine & Education*, vol. 1, 2009, pp. 337–340.
- [59] L. A. Sparrow, M. Antonellos, M. Gibbs, and M. Arnold, "From 'Silly' to 'Scumbag': Reddit discussion of a case of groping in a virtual reality game," in *Proceedings of the 12th DiGRA International Conference, The Digital Games Research Association (DiGRA)*, 2020.
- [60] L. Wu, K. B. Chen, and E. P. Fitts, "Effect of body-gender transfer in virtual reality on the perception of sexual harassment," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, 2021, pp. 1089–1093.
- [61] K. Schulenberg, L. Li, C. Lancaster, D. Zytco, and G. Freeman, "'We Don't Want a Bird Cage, We Want Guardrails': Understanding & Designing for Preventing Interpersonal Harm in Social VR through the Lens of Consent," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 7, pp. 1–30, 2023.
- [62] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe, "Classification and its consequences for online harassment: Design insights from heartmob," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 1, pp. 1–19, 2017.
- [63] K. Collins, "Tech is overwhelmingly white and male, and white men are just fine with that," Quartz, 2017. [Online]. Available: <https://qz.com/940660/tech-is-overwhelmingly-male-and-men-are-just-fine-with-that>
- [64] "Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity," US Department of Labor, Bureau of Labor Statistics, 2018. [Online]. Available: <https://www.bls.gov/cps/cpsaat11.htm>
- [65] L. A. Goodman, "Snowball sampling," *The annals of mathematical statistics*, pp. 148–170, 1961.
- [66] R. Bellini, E. Tseng, N. Warford, A. Daffalla, T. Matthews, S. Consolvo, J. P. Woelfer, P. G. Kelley, M. L. Mazurek, D. Cuomo *et al.*, "Sok: Safer digital-safety research involving at-risk users," in *Proceedings of the 45th Symposium on Security and Privacy (SP)*, 2024, pp. 71–71.
- [67] J. Bisson and M. Andrew, "Psychological treatment of post-traumatic stress disorder (ptsd)," *Cochrane database of systematic reviews*, no. 3, 2007.
- [68] "To get over something, write about it," 2023. [Online]. Available: <https://hbr.org/2014/11/to-get-over-something-write-about-it>
- [69] Qualtrics Survey. [Online]. Available: <https://www.qualtrics.com/>
- [70] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *OpenAI Blog*, 2022.
- [71] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice," *Proceedings of the ACM on human-computer interaction (CSCW)*, vol. 3, pp. 1–23, 2019.
- [72] J. M. Roehl and D. J. Harland, "Imposter participants: overcoming methodological challenges related to balancing participant privacy with data quality when using online recruitment and data collection," *The Qualitative Report*, vol. 27, no. 11, pp. 2469–2485, 2022.
- [73] A. C. Cote, "'I can defend myself' women's strategies for coping with harassment while gaming online," *Games and culture*, vol. 12, no. 2, pp. 136–155, 2017.
- [74] D. Zytco and J. Chan, "The Dating Metaverse: Why We Need to Design for Consent in Social VR," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2489–2498, 2023.
- [75] M. Arzaghi and J. V. Henderson, "Networking off madison avenue," *The Review of Economic Studies*, vol. 75, no. 4, pp. 1011–1038, 2008.
- [76] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, "Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms," in *Proceedings of the 11th Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019, pp. 464–469.
- [77] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The bag of communities: Identifying abusive behavior online with preexisting internet data," in *Proceedings of the 37th conference on human factors in computing systems (CHI)*, 2017, pp. 3175–3187.
- [78] K. Schulenberg, L. Li, G. Freeman, S. Zamanifard, and N. J. McNeese, "Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality," in *Proceedings of the 43rd Conference on Human Factors in Computing Systems (CHI)*, 2023, pp. 1–17.
- [79] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, and M. Elsherief, "Challenges of Moderating Social Virtual Reality," in *Proceedings of the 43rd Conference on Human Factors in Computing Systems (CHI)*, 2023, pp. 1–20.

Appendix

A Participant demographics

A.1 Targets of VR-based harassment

Table 5: Demographic information of targets of VR-based harassment (self-reported).

ID	Age	Gender	Sexual Orientation	Race	Usage (years)	Usage (hrs/week)	VR apps with harassment experience		
							Social VR	Gaming VR	Streaming VR
T1	18-24	Non-binary	Other	Black	6	4-20		Asgard's Wrath	
T2	25-34	Male	Heterosexual	White	5	4-20			YouTube VR
T3	25-34	Female	Heterosexual	Black	4	20-40	Second Life		
T4	25-34	Female	Heterosexual	Black	3			Beat Saber	BigScreen
T5	18-24	Male	Heterosexual	Black	3	4-20	VR Chat	Beat Saber	
T6	25-34	Male	Heterosexual	White	<2	4-20		Star Wars: Squadron	YouTube VR
T7	45-54	Female	Other	White	3	4-20		Orbus, Zenith	
T8	18-24	Male	Bisexual	White	4	4-20	VR Chat	Pavlov VR	
T9	18-24	Male	Heterosexual	Black	4	4-20	Sinespace	Archangel	
T10	25-34	Female	Heterosexual	Black	<1	1-4	Second Life		
T11	25-34	Male	Heterosexual	Black	<1	1-4	VR Chat	House of Terror	YouTube VR
T12	45-54	Female	Heterosexual	White	6	20-40		Echo VR	
T13	35-44	Male	Heterosexual	White	2	4-20	VR Chat	Echo VR	
T14	35-44	Male	Other	American Indian	7	4-20	VR Chat	Echo VR	Big Screen
T15	25-34	Male	Heterosexual	White	4	20-40	VR Chat	Echo VR, Gorilla Tag	
T16	18-24	Male	Heterosexual	White	1	1-4	RecRoom	Echo VR, Pavlov VR, Walkabout Mini-Golf	
T17	18-24	Female	Heterosexual	American Indian	<1	4-20	VR Chat	Echo VR	
T18	35-44	Male	Homosexual	Other	4	20-40		Echo VR	

A.2 VR developers

Table 6: Demographic information of VR developers (self-reported).

ID	Age	Gender	VR dev experience	Education (degree)	Tools used	Roles performed					
						UI/UX Designer	XR Gameplay & Tools Engineer	Software Developer	Researcher	AR/VR Maintenance & Support	Other Roles
D1	25 – 34	Male	>= 4 years	Bachelor's	Unity	✓		✓		✓	
D2	25 – 34	Male	2 – 3 years	Doctoral	Unity		✓	✓	✓		Graphics Engineer
D3	18 – 24	Male	1 – 2 years	Bachelor's	Unity	✓	✓	✓	✓	✓	
D4	35 – 44	Male	>= 4 years	Bachelor's	Unreal Engine		✓			✓	
D5	25 – 34	Female	3 – 4 years	Master's	Unity	✓			✓		
D6	25 – 34	Male	>= 4 years	Master's	Unreal Engine				✓		
D7	35 – 44	Male	>= 4 years	Master's	Unity, Maya Creative cloud, Substance suite	✓	✓	✓			Product, Manager, Marketing
D8	18 – 24	Male	1 – 2 years	Vocational training	Unity, Blender	✓	✓	✓			
D9	18 – 24	Male	1 – 2 years	Bachelor's	Unity	✓	✓	✓			