Beyond Mute and Block: Adoption and Effectiveness of Safety Tools in Social VR, from Ubiquitous Harassment to Social Sculpting

Maheshya Weerasinghe* University of Glasgow, University of Primorska
Shaun Macdonald† University of Glasgow
Cristina Fiani‡ University of Glasgow
Joseph O'Hagan§ University of Glasgow
Mathieu Chollet¶ University of Glasgow
Mark McGill‖ University of Glasgow
Mohamed Khamis** University of Glasgow

Abstract— Harassment in Social Virtual Reality (SVR) is a growing concern. The current SVR landscape features inconsistent access to non-standardised safety features, with minimal empirical evidence on their real-world effectiveness, usage and impact. We examine the use and effectiveness of safety tools across 12 popular SVR platforms by surveying 100 users about their experiences of different types of harassment and their use of features like muting, blocking, personal spaces and safety gestures. While harassment remained common–including hate speech, virtual stalking, and physical harassment–many find safety features insufficient or inconsistently applied. Reactive tools like muting and blocking are widely used, largely driven by users' familiarity from other platforms. Safety tools are also used to proactively curate individual virtual experiences, protecting users from harassment, but inadvertently leading to fragmented social spaces. We advocate for standardising proactive, rather than reactive, anti-harassment tools across platforms, and present insights into future safety feature development.

Index Terms—Metaverse, Extended Reality, Online Safety

1 Introduction

Social Virtual Reality (SVR) platforms allow users to engage in immersive, embodied experiences that simulate real-world social interactions in virtual environments. These platforms create a heightened sense of presence, making virtual spaces feel increasingly perceptually realistic and socially engaging. The growth of SVR has raised concerns about user safety, especially in dealing with harassment and abuse [1, 2, 7] as its immersive nature amplifies the psychological and emotional impact of these negative experiences compared to traditional platforms [12, 33]. Harassment in SVR ranges from verbal abuse and pranks to virtual sexual assault [12, 35], with anonymity and embodiment intensifying the sense of presence [44], making these harassments feel more immediate and impactful than on 2D platforms [24]. The difficulty in moderating these spaces, coupled with the rapidly growing user base, further complicates efforts to ensure user safety [17, 33].

To mitigate these risks, SVR platforms have introduced safety features often inspired by traditional online platforms, such as muting, blocking, and reporting. These tools allow users to *react* quickly to harassment after it has occurred. SVR-specific safety features have also emerged, such as personal space boundaries and interaction shields, designed to *proactively* prevent harassment by allowing users to control their virtual environments and how others interact with them [3–5].

Despite these advancements, the implementation of safety tools across platforms remains inconsistent, with many relying on reactive rather than preventive measures [33, 39, 43, 45]. There is little empirical evidence assessing how existing safety tools are used and how well they are perceived to perform in practice. Existing studies have documented the prevalence and effects of harassment in SVR, but have not explored how users interact with these safety features or how effective they are [12, 19, 35]. Some studies have highlighted the potential for misuse or abuse of these tools [19, 45], and some suggest that existing safety features are inadequate to handle the more nuanced forms of harassment that SVR enables [12, 16, 19, 35]. However, existing safety features' use and impact in predominant SVR platforms have not been investigated.

In this paper, we fill this gap by assessing the scale of harassment in SVR as a prerequisite to understanding how users respond to these incidents, and their awareness, adoption and perceptions of the effectiveness of available safety features. We first conducted a review of safety tools across 12 popular SVR platforms, and then deployed an online survey (N=100) to gather both quantitative and qualitative evidence about users' experiences with harassment, whether as victims or bystanders, as well as their use and perceptions of safety features on these platforms. Our study is the first to evaluate how well these safety mechanisms function in practice, detail the nuances of their adoption and use, and identify gaps that need to be addressed in future work.

We found that many users consider harassment an inherent quality of SVR from which protection is not assured and the perceived impact of harassment did not significantly vary between victims and bystanders. Although users considered most SVR safety features similarly effective and preferable, they predominantly leverage well-established tools, such as reporting, blocking, or muting. Meanwhile, SVR-specific features such as personal space boundaries or safety gestures, are less consistently employed or available. Users also described leveraging SVR-specific features for pre-emptive protection and to sculpt their VR experience in previously undocumented ways, such as reducing visual clutter or public interaction. Based on our findings, we make recommendations for SVR safety features and further advocate for standardised SVR-specific features across platforms. Finally, we reflect on the implications of a social frontier which users can sculpt to their individual preferences. We contribute:

- A review of safety tools across 12 popular SVR platforms and highlight implementation gaps, including the need for proactive tools, cross-platform consistency, and community-based moderation to address SVR harassment.
- Insights into how users interact with SVR safety features in practice, their perceived effectiveness, and their impact on social interactions, from protection to social fragmentation.

*e-mail: Maheshya.Weerasinghe@famnit.upr.si (co-first author)
†e-mail: Shaun.Macdonald@glasgow.ac.uk (co-first author)
‡e-mail: C.Fiani.1@research.gla.ac.uk
§e-mail: Joseph.OHagan@glasgow.ac.uk)
¶e-mail: Mathieu.Chollet@glasgow.ac.uk
‖e-mail: Mark.Mcgill@glasgow.ac.uk
**e-mail: Mohamed.Khamis@glasgow.ac.uk)

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

- Findings highlighting the potential of bystanders in harassment situations and provide insights on how to leverage their involvement to enhance overall safety in SVR environments.
- Practical recommendations to improve SVR safety.

2 Related Work

2.1 Harassment in Social Virtual Reality

As opposed to 2D social media (e.g., Instagram) or 2D video games, SVR allows immersive, embodied social experiences that mimic real face-to-face interactions through advanced VR technologies like body tracking, eye tracking, haptics, and 3D audio [6]. Users can engage in activitiessuch as gaming and events while their senses (visual, auditory, and touch) are immersed, fostering a sense of presence [32]. Such multi-sensory immersion is key to generating a sense of presence, which in turn makes users more vulnerable to potential risks of online harm as they can feel even more directly impactful and immediately threatening. However, SVR platforms lack established social norms [24] and the level of immersion and synchronicity they afford while still enabling anonymity [12] has led to new forms of harm including physical harassment (e.g., an avatar getting close and entering the personal space of another user) and environmental harassment (e.g., displaying inappropriate content) [12]. For instance, a user reported feeling unsafe when a group of strangers surrounded them, shouting slurs and swears [24]. Another user observed that "more times than not, there would be one guy or a group of guys engaging in harassing behavior because that's how they enjoy their weekend." [29]. Additionally, others have highlighted recurring harassment by groups exploiting the affordances of VR for disruptive behaviors like flash mobs that "box in" users or overwhelm them with disruptive noises [26]. Prior work has also shown that children [19, 33, 35, 36] and marginalised users may experience higher risks of harassment in SVR [11, 41]. Events have been frequently documented in media articles, including rape in the metaverse [1], groping problems [2] and sexual assaults [7] towards women and children. Recent work has explored the experiences of individuals who have both been victims and accused of harassment [21], uncovering the need to understand the motivations behind people's behaviours and how harassment accusations can be used against marginalized SVR users. Therefore, the need for effective, usable mitigation tools in these platforms is crucial to protect users from harassment.

2.2 Existing Safety Tools in SVR

A recent study analysed YouTube videos of SVR users and identified key real-time safety features across four major SVR platforms, including VRChat and RecRoom [45]. They categorised SVR safety features into: *boundary settings* (e.g., proxemics, trust reputation, social spaces, shield levels, voice and avatar controls), *quick-reactions* (e.g., safety gestures, vote to kick, safety zone and safety reports) and *agreements* (e.g., code of conducts) [45]. The study found that relying solely on proxemics is insufficient, as safety risks often require distinct social cues (e.g., linguistic). Existing features lack natural reactive gestures as shortcuts to safety (such as "talk-to-the-hand gesture" to block), and incorporating such gestures could improve user response during safety risks [45]. Most safety features are reactive rather than preventive [45], placing responsibility on victims to use them effectively under distress. The study advocates for preventive measures, including using social cues or kinetics to detect risks before they happen [18, 45]. In addition to mitigation tools, human moderators address incidents in real time [12, 24, 40], with automated moderation proposed as a scalable alternative [17, 42].

2.3 Limitations of Existing Safety Tools

2.3.1 Traditional Safety Tools in SVR

Some safety tools have been translated from social media into SVR, such as reporting and blocking. But their effectiveness has not been studied. Prior work argued they are inadequate and highlighted the importance of addressing the nuanced social cues—such as proxemics, linguistics, paralinguistics, and kinetics—associated with both attackers and victims [45]. Furthermore, there is a notable inconsistency across different VR platforms like VRChat (e.g., Trust Rank) and RecRoom (e.g., hand gestures to block), leading to varying levels of safety and user protection. This inconsistency makes it difficult for users to have a uniform experience and can leave some users more vulnerable than others [45]. Privacy concerns also arise with the collection of personal information for advanced tools like body and eye tracking [9, 10, 36, 45].

2.3.2 AI-Driven Moderation and Tools

Human moderators can be effective in managing disruptions, but they are limited in scalability and availability. A recent study found that only 24% of incidents in SVR were addressed by human moderators [40], illustrating their inability to cover all issues in a rapidly growing VR ecosystem and emphasising the need for more automated moderation tools. While there have been attempts to incorporate AI-driven moderation [16, 17], the technology is not yet mature for effective deployment. Similarly, other AI-driven features, such as voice analysis to detect hate speech and profanity, are not yet fully reliable and often lack the necessary accuracy. This inaccuracy can lead to both false positives and false negatives, undermining trust in these systems [16, 42]. While initial efforts toward AI-driven moderation have been made, there is still a gap in understanding the practical use and trustworthiness of these tools from the perspective of everyday users, which we address.

2.3.3 Misuse and Bias in Safety Tools

Previous research also highlighted the risk of safety tools being misused or abused [45]. For example, Trust rank systems, which prioritise users based on their reputation and history, can be biased and unfairly favor long-term or more popular users, potentially marginalising newcomers or less active participants who may not have had the opportunity to build a high rank [15]. Although prior work highlights the misuse and potential biases in these systems, little is known about how these shortcomings affect perceived safety and social dynamics in practice, which we address in our work.

2.3.4 Effectiveness of SVR Safety Tools

To the best of our knowledge, only one study by Zheng et al. [45] investigated safety tool effectiveness in SVR, but focused primarily on direct observations of safety features in action without exploring how users interact with these tools in the context of their personal experiences. Our work, on the other hand, fills this gap by presenting the first study of the perceptions, perceived effectiveness, adoption and impact of SVR safety on social interactions and the broader implications on SVR environments, from the perspective of users and bystanders. Addressing this gap is key to refining the development of more effective and usable SVR safety tools.

3 Methodology

To assess the scale of harassment in SVR and how users engage with available safety tools, we first identified the existing safety features in the most popular SVR platforms, and the most common uncomfortable and unsafe situations associated with SVR based on a literature review. This was then followed by a survey of 100 SVR users to explore their harassment experiences and investigate how effectively they used self-protection tools, and examine the bystanders' role in these interactions.

3.1 Identifying the Social VR Platforms to Consider

We selected the 12 most popular SVR applications based on numbers of downloads and monthly active users across popular VR app stores in January 2024 including Steam VR, Meta Horizon Store, and Sidequest. We define an SVR platform as a VR application in which users are embodied by avatars and can remotely interact with other users in fully immersive 3D virtual environments whilst wearing a headset. The use of these embodied avatars allows users to express both verbal, like speech and voice chat, and non-verbal behaviours, like gestures, body language, facial expressions, and other forms of non-verbal cues detected using full or partial body tracking [22, 23]. This means that our selected platforms included not only purpose-built SVR platforms, but also games, such as Roblox, where social interaction is central.

Following the criteria above, the selected platforms are: *VRChat*, *RecRoom*, *Horizon Worlds*, *Altspace* VR[1], *BigScreen*, *NeosVR*, *Spatial*, *Mozilla Hubs*, *Cluster*, *Roblox*, *Sandbox* and *Sansar*.

3.2 Identifying Unsafe SVR Situations

We conducted a review of existing literature [12, 19, 24, 26, 29, 33–36, 40, 41, 45], focusing on verbal, visual, physical discomfort and safety concerns within these environments, and classified them into three categories (see Table 1 for more details):

- **Verbal Uncomfortable/Unsafe Situations**: those are a result of inappropriate language, or unwanted conversations. Examples include hate speech, sexualised language, and voice-trolling.
- **Visual Uncomfortable/Unsafe Situations**: these stem from exposure to offensive or disturbing content. Examples include virtual scaring, or displaying inappropriate (sexual or violent) content.
- **Physical Uncomfortable/Unsafe Situations**: these come from the immersive embodied nature of SVR platforms, where virtual promixity can feel like real-world invasions of personal space. Examples include physical assault and unwelcome touching.

3.3 Identifying SVR Safety Features

We conducted a review of existing safety features provided by the above 12 SVR platforms. We did this by a) trying the application ourselves, and b) reviewing the platform's community safety websites. This resulted in a list of safety features and coping strategies (see Table 2), which we categorised to a) boundary setting features that are mostly preventative and set *before* harassment incidents, and b) quick reaction features which are to be used *during* or *after* the harassment incident. Note that because the studied platforms use different terminologies for the same functionality, we sometimes group safety features under names that are different from those used on some platforms e.g., some platforms refer to "Personal Space" as "Safety Bubble".

3.4 Survey: Participants' Experiences of Harm in SVR and Perspectives on Effectiveness of Safety Tools

To assess the adequacy and effectiveness of the identified safety features, we designed an online questionnaire on `https://www.qualtrics.com/`Qualtrics. The survey was divided into five main sections covering the following:

1. **Demographics:** Information on participants' backgrounds, such as their age, gender, country of residence, and nationality.
2. **SVR Usage:** Participants' experiences with VR and their familiarity with SVR platforms, with particular focus on the 12 popular platforms mentioned in Section 3.3 and analysed in Table 2.
3. **Experiences in SVR:** Reporting any unsafe or uncomfortable situations participants had faced or witnessed in SVR environments (Table 1), including the frequency of these situations and their how they felt during those moments.
4. **Use of Safety Features:** Which specific safety features the participants used to address these experiences (Table 2), such as boundary settings or safety reactions provided by SVR platforms, and how these tools influenced their overall experience and sense of safety in SVR.
5. **Perceptions and Preferences:** Opinions about existing safety tools, including their ease of use, perceived effectiveness, and preferences for future safety features and improvements.

The core sections (3), (4) and (5) were organised with a mixture of rating scales and open-ended questions. This allowed us to both ascertain quantitative elements of SVR harms, such as the frequency of exposure to harm and the distribution of participant sentiment on the severity of harassment types and the effectiveness of safety tools, as well as collect rich qualitative feedback on experiences. The survey questions were reviewed by senior VR and HCI researchers to ensure they aligned with our research objectives. We also conducted multiple pilot tests within our departments to refine the questions and ensure they addressed the research topic. These details have been added to the revised paper for clarity on the survey's validity. The full survey is available in the supplemental material.

[1]We included AltspaceVR for its popularity before discontinuation in 2023.

3.5 Participant Recruitment and Ethical Considerations

Our study received approval from both University of Glasgow's ethics board (Approval number 300230070) and REPHRAIN's Ethics Board. SVR users aged 16 and above were invited to take part in the study by advertising on online discussion boards, including on relevant Reddit sub-communities (r/virtualreality, r/OculusQuest, r/VRresearch, r/VisionPro, r/SocialVR), on SteamVR's Discussions page and on LinkedIn, in February 2024. At the beginning and end of the survey, we included statements acknowledging the potential sensitivity of the topics discussed, along with links to mental health resources for participants in case recalling sensitive incidents may cause distress (e.g., `https://www.mind.org.uk/information-support/`). We also asked participants to refrain from disclosing personally identifiable data. Data was checked to identify any accidental disclosure of personally identifiable data, compromising anonymity or bypassing consent criteria, and corresponding data was deleted.

Participants were incentivised to participate through a lottery giving a chance to win online shopping vouchers worth £20. In total, 100 international participants enrolled in the study from 22 countries, with a majority of participants from the Philippines (n=26), the USA (n=21) and the UK (n=13), with an average age of 27.3 years ($\sigma = 9.6$), and 65 participants identifying as Male, 33 as Female, and 2 as Non-Binary.

3.6 Data Analysis

3.6.1 Quantitative Analysis

The distributions of participant responses to rating scales were analysed using descriptive statistics. We identified the proportion of participants reporting experiences of the 15 harassment types outlined in Table 1, on participants' subjective experiences of harassment in SVR in terms of its perceptual qualities, e.g. immersiveness and realism; their emotional responses to harassment; their feeling of being (not) in control when witnessing or experiencing SVR harm; and their responses to these situations. Participants who did not experience a specific type of harassment were not posed follow-up questions regarding it. We then analysed distributions of participants' familiarity with and general sentiment regarding currently implemented SVR safety features.

3.6.2 Qualitative Analysis

To analyse free-text responses we applied thematic analysis [13, 14][2]. Two researchers each conducted an initial inductive coding pass on an identical 25% subset of the data then met to normalise codes. With this initial codebook, one researcher then re-coded the entire data set, making alterations as required. After consulting with the second researcher on the resultant coding scheme, both researchers conducted Axial coding to form the categories and themes. Throughout the process the codes, categories and themes were reviewed and revised in discussion between both researchers, to reduce individual subjectivity in the outcome. For the final codebooks, see supplementary material.

3.7 User Scenario

To help readers understand how a user might experience harassment and use safety tools in SVR, we present a sample user scenario. After logging into a public lobby with full immersion, Alex encounters Chris, who begins harassing Alex by hurling verbal insults and invading their personal space by following very closely. Chris then displays offensive visuals, such as inappropriate or graphic images. Feeling uncomfortable, Alex uses platform tools to Mute or Block Chris, removing their presence. Seeking further comfort, Alex moves to a Private Social Space with trusted friends and activates personal space boundaries. Finally, Alex submits a report to document the incident.

4 Results

We present two key sets of findings. First, we confirm existing research on SVR harassment patterns and provide new insights into bystander involvement. Second, we introduce novel insights highlighting previously unexplored aspects of safety tool adoption and perception.

[2]Following Braun and Clarke, we used a qualitative approach to validity, emphasising researcher influence over quantitative Inter-Coder Reliability [14].

Table 1: Commonly encountered verbal, visual, and physical discomforts or unsafe situations experienced by users in SVR.

Verbal Uncomfortable/Unsafe Situations	**Visual Uncomfortable/Unsafe Situations**	**Physical Uncomfortable/Unsafe Situations**
Hate speech or other forms of discrimination *(e.g. Racist, homophobic, violent comments or threats)*	Virtual scaring *(e.g. Employ scary-looking avatars to scare other users, and either rush towards them or appear in front of them out of nowhere)*	Sexual assault or abuse *(e.g. Groping or any sexual suggestive touching, such as grabbing someone's avatar's private parts)*
Personal insult *(e.g. Inappropriate jokes, offensive name-calling, or teasing)*	Displaying sexualised content *(e.g. Showing unsolicited sexually related images/videos, showing sexualised avatars or gestures)*	Physical assault or virtual violence *(e.g. Punching, kicking, slapping, or throwing objects at someone's avatar)*
Sexualised language *(e.g. Sexual jokes or ask sexually related questions)*	Displaying abusive or inappropriate messages *(e.g. Showing hate or threaten comments, sending sexually explicit or obscene messages)*	Interrupting or preventing movements *(e.g. Blocking the path of the someone's avatar using objects or another avatar when moving around)*
Make inappropriate sounds *(e.g. Kissing sounds, whistling, or smacking lips)*	Fraud-impersonation *(e.g., Purposefully identifying oneself as another individual or group, such as a social VR employee or an existing social VR user)*	Inappropriately hugging or unwelcome touching *(e.g. Hugging someone's avatar in a way that is intimidate or threaten, touching on the avatar's body, hair, or clothing)*
Voice-trolling *(e.g., Gender-mismatch voice or contrasting voice like a lovely avatar with a frightening voice to scare other users)*	Virtual crashing *(e.g. Use tactics or bugs to ruin others' experience, such as adding particle effects like fire, electric bundle animation in avatars to cause damages)*	
	Displaying violent contents *(e.g. Showing sensitive--i.e. killing, abusing, scaring, etc.-- images or videos)*	

Table 2: Analysis of major safety features across popular SVR platforms.

Safety Feature	**VRChat**	**RecRoom**	**Horizon Worlds**	**Altspace VR**	**BigScreen**	**NeosVR**	**Spatial**	**Mozilla Hubs**	**Cluster**	**Roblox**	**Sandbox**	**Sansar**
Boundary Settings												
Intimacy Proxemics: A feature that allows users to establish and maintain a distance with other users in the virtual environment.	✓	✓	✓	✓		✓	✓					
Social Spaces: A feature that allows users to create and designate specific virtual areas as safe spaces.	✓	✓	✓	✓		✓	✓	✓	✓			
Intimacy Rank: A "trust rank" feature with several levels which users can customise their trust for interactions with other avatars.	✓											
Content Gating: A feature that allows to control who can see and interact with certain types of content of the users.	✓	✓	✓	✓		✓	✓					
Interaction Shields: A set of adjustable settings that users can customise to control how their avatars interact with others.	✓	✓		✓		✓						✓
Parental Control: A set of tools and settings that allow parents or guardians to manage their children's VR experiences.		✓	✓							✓		
Quick Reactions												
Safety Gestures: A feature that allows users to activate safety features using hand movements quickly and easily.		✓										
Safe Zone Teleport: A feature that allows users to instantly teleport themselves to a designated Safe Zone.	✓	✓		✓		✓						
Freeze Controls: A feature that allows users to temporarily freeze the controls of their avatars for preventing other users engaging.	✓											
Vote Kick: A feature that allows users to initiate a vote to remove another user from a specific virtual space.	✓	✓										
Mute and Block: The Mute feature enables users to silence specific individuals and the Block feature prevent specific individuals temporarily or permanently interacting with them.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Report: A feature that allows users to flag inappropriate or harmful behaviour to moderators or platform administrators.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other												
AI Moderation: A feature that utilises artificial intelligence (AI) to monitor user interactions, content, and behaviour to identify potential violations of community guidelines and safety concerns.	✓	✓		✓						✓		

4.1 Harassment Experiences

4.1.1 Prevalence of Harassment Types in Social VR

Participants reported their experience as "victim" or "witness" of the fifteen types of SVR harassment in Table 2, divided into three high-level categories: verbal, visual and physical harassment. For most harassment types a similar but slightly larger proportion of participants reported having been a victim of harassment versus having been a witness. The number of participants that observed each harassment type is reported in Table 3. A roughly equal proportion of harassment was coded as Severe (including content such as Hate Speech and Sexual Assault) and less severe Trolling or Flaming (such as Virtual Scaring, Virtual Crashing or Jokes). In total, across all types of harassment, 95% of the participants reported experiencing or witnessing some form of harassment in Social VR.

Verbal harassment was most prevalent in our dataset, experienced in some capacity by 80% of participants: 70% as the victim and 49% as a witness. Hate Speech was most common, with 44% of participants experiencing it as victims (34% as witnesses), followed by Personal Insults (36% as victims, 22% as witnesses), Sexualised Language (25% as victims, 22% as witnesses), others making Inappropriate Sounds (21% - 13%) and Voice Trolling (17% - 11%).

Visual harassment types were experienced by 72% of participants: 56% as victim and 46% as witness. Of these, 20% of participants were subjected to other users displaying images of Inappropriate Messages, followed by Virtual Scaring (19%), Virtual Crashing (18%) and displaying Sexualised (18%) or Violent (8%) content. Only 4% of respondents had been subjected to users changing their avatar visuals to attempt Impersonation of Fraud.

Physical harassment was experienced by 60%, 49% as victim and 41% as witness, involving other users Interrupting the Movement of the participant (27%), followed by Virtual Violence (14%), Sexual Assault (11%) and Unwelcome Touching (7%).

Participants also commented on the prevalence of harassment in SVR. Many felt *harassment was inherent* and *expected* in SVR, with some describing such incidents as "typical" (P39), "part of being on the internet" (P96) or " very familiar and mundane" (P70). Those with prior experience in online spaces or gaming identified SVR as just the latest venue for a familiar issue, with P85 writing: "online games are rife with people being horrifically discriminatory in every possible way and until the industry does something about it online spaces will never feel like a safe space." This familiarity led some to feel desensitised to harassment, such as P87: "personally I am an internet veteran, so

Verbal		Hate Speech	Personal Insult	Sexualised Language	Inapprop. Sound	Voice Trolling	
	Victim	44%	36%	25%	21%	17%	
	Witness	34%	22%	22%	14%	11%	
Visual Display		Virtual Scaring	Sexualised Content	Violent Content	Fraud	Virtual Crashing	Inappropriate Message
	Victim	19%	18%	8%	4%	18%	20%
	Witness	18%	14%	4%	2%	14%	18%
Physical		Sexual Assault	Unwelcome Touching	Virtual Violence	Interrupt Movement		
	Victim	11%	7%	14%	27%		
	Witness	13%	6%	12%	23%		

Table 3: Proportion of participants (n=100) who experienced each of the 15 harassment types surveyed as a victim of harassment or a witness.

maybe I am a bit desensitised to this type of thing".

Others, however, were unprepared for the harassment they experienced. P45 wrote "I was in unexpected situations and it left me on shock and I felt uncomfortable and angry. No one tried to help or [get] involve[d]", while P78 was surprised by the contrast to their offline experiences: "I just felt kinda scared/bad because I'm not used to witnessing racist or homophobic comments in real life". Respondents like P2 felt that anonymity was a key enabler for increased harassment: "people tend to make racist and homophobic remarks a lot because they're hidden by their anonymity when using the VR platforms. I'm pretty sure majority of those that make these remarks wouldn't be making them if they were playing in the VR room under their real identity". This in turn led to calls for "more complete virtual identity authentication methods" (P97) [8, 25, 38].

4.1.2 Participant Experiences of Harassment in SVR

After asking how often participants had experienced each harassment type, they were then asked to report their experiences with each type of harassment they had experienced by indicating agreement through 5-point scale questions (see Sec.3.4). These questions asked participants to indicate if 1. the harassment experience was consistent with real-life experiences; 2. if the harassment felt real; 3. if they felt pleasant during the experience; 4. if they felt in control during the experience and 5. if they felt emotionally intense during the experience. Proportional responses per question and harassment type for victims and witnesses are shown on Figures 1/2.

Realism and Immersiveness of Harassment: For each harassment type, respondents were asked to report how consistent their experience is with its offline counterpart and how real the experience felt. An average of 33% of respondents felt verbal harassment was consistent with their offline experiences. Hate speech felt consistent (52%) and real to most participants (45%), with P33 noting that "the anonymity in the online platform makes you want to fight it, but hate only increases more. This is something that mimics real-life situations". Visual harassment was least immersive, with 22% of respondents agreeing their experience was consistent with offline ones across all types. Physical harassment types like Movement Interruption and Physical Assault were also considered inconsistent, with 52% and 71% of respondents disagreeing that their experiences felt real or were consistent with offline experiences, respectively. This may be due to how differently these harassment manifest in VR without physical contact.

Sexual assault, however, was felt to be consistent with offline experiences by 64% of respondents and far more reported that it felt real (45%) than not (18%), with P65 writing "It was an horrible experience. I felt abused, I didn't know what to do, it seemed so real and I was disgusted." Others also noted that when harassment felt unrealistic was less impactful: "I know it's not a real-life situation but just an imitation of it so I was not necessarily moved" (P33, regarding Virtual Scaring). P85 highlighted how the consequences for harassment are inconsistent with their real-world experiences, leading to higher prevalence:"since the real-world consequences of hate speech online haven't really happened yet people who are reported and banned don't have a sufficient deterrent" Finally, some still found harassment impactful even when it was not real, such as P16 who experienced Physical Harassment: "yes none of it real but he is still a real person and he clearly wanted to cause discomfort and that is upsetting whether he is really present or not".

Emotional Responses to Harassment: Verbal harassment was considered by an average 62% of victims to make them feel unpleasant, and by 41% felt it made them feel 'intense' or not calm. Of these, Hate speech was considered unpleasant and intense by the highest proportions (77% and 59% respectively). One type, Voice Trolling was considered pleasant by more respondents (35%) than unpleasant (29%) and was only considered not calming by 18%, with several describing it as "funny" and P72 felt that "honestly it's others shaming them that are the worst. People goofing around trying to make people laugh with silly avatars", although they noted that "if it's pejorative and targeting one person based on prejudices now it's serious". On average visual harassment types were considered unpleasant by 71% of respondents and not calm by 43%. Virtual crashing (72%) and display sexual content (72%) were most often reported as unpleasant. Virtual Scaring was only unpleasant for 37%, with many downplaying its severity: "jump scares in VR are just like pranks in real life, for better and worse" (P83). Finally, an average of 62% or participants found physical harassment types unpleasant, while 48% found them intense. Sexual Assault was reported as particularly unpleasant by 73% of respondents and as intense by 64%. Movement Interruption was also unpleasant (74%) but far fewer respondents found it to be intense (37%), with qualitative feedback reflecting respondents felt "annoyed" (P7, P38, P88).

Feelings of Control During Harassment: 46% respondents on average felt they had control of their experience during verbal harassment, more than visual harassment (40%) or physical harassment (36%). This is likely due to the ubiquity and perceived effectiveness of tools like Mute and Block (see Sec.4.2) which users can deploy to curb verbal harassment, as highlighted by P48 "most platforms have a mute option so its a non-issue". Others felt a lack of agency, however, due to the overall prevalence of verbal abuse: "I can't really do anything about these situations. But discrimination against [the participant's country of origin] aren't rare in these virtual environments" (P26). The fewest respondents felt they had control during Virtual Crashing (78%), with P19 noting that they were "unable to adequately respond to situation as usually it's unknown which user is causing the crashes" and P51 writing ""I felt like I was not in control and that I could not get help quick enough".

How Users Respond to Harassment Qualitative feedback indicated that experiencing harassment impacted participants in one of four main ways. The most common approach, described a total of 227 times across participants (including multiple instances described by the same participant), was to use typical game safety features such Mute, Block, Vote-Kick and Reporting. Many framed this response as a natural, quick and effective default action, such as P18: "People are rude, nothing more or less to it, it's why the mute and block buttons exist", P33: "sometimes you just need to mute or leave to protect your sanity against this type of hate", or P93: "in public it's a instant report and block for me". Taking responding action also requires effort, however. After being subjected to Hate Speech P19 felt "saddened, emotionally hurt, and insanely uncomfortable" and responded by "muting myself and not engaging with harassers, but emotionally shutting down", noting that they only used safety features such as vote-kicking or blocking "if emotionally equipped to".

Second, some participants ignored harassment and attempted to continue their SVR usage unimpeded, described 72 times. This was sometimes because they did not find the harassment to be impactful, such as P4: "It didn't really matter that much to me. It just felt that a kid was trying to annoy my friend in the game. I just ignored the player and went on with my day"' As discussed in Section 4.1, some felt that harassment was inherent and needed to be adapted to: " You either deal with it or you don't" (P48). Others like P54 did feel impacted but

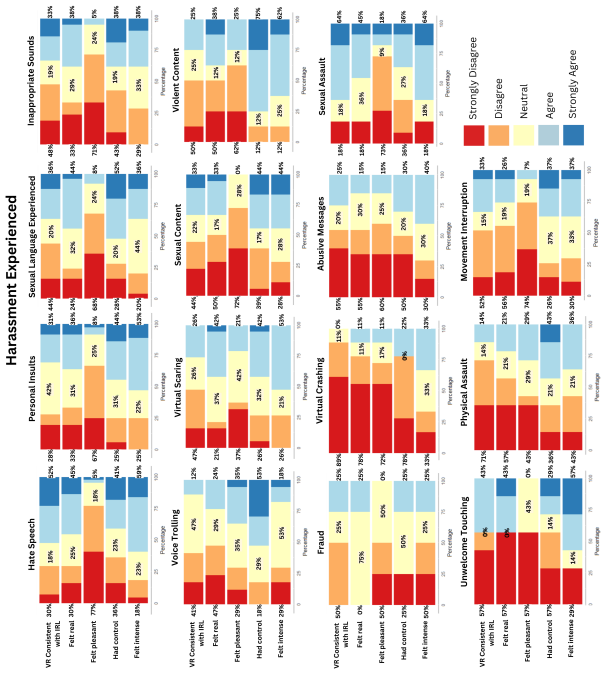

Fig. 1: Participants experiences as ***a victim*** of SVR with harassment, represented proportionally per type of harassment they experienced. Participants were asked if they felt the experience was real, if it was consistent with real life experience, if it felt pleasant (valence), if they felt that had control and if it felt emotionally intense (arousal).

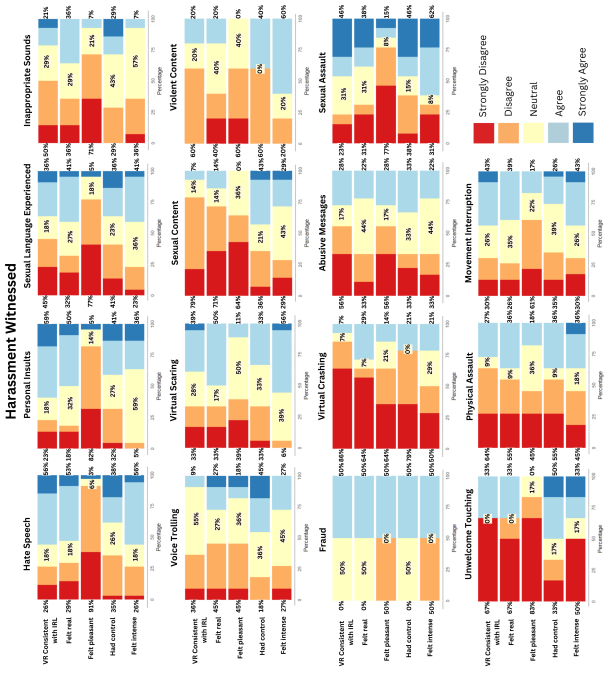

Fig. 2: Participants experiences ***witnessing*** SVR harassment, represented proportionally per type of harassment they witnessed. Participants were asked if they felt the experience was real, if it was consistent with real life experience, if it felt pleasant (valence), if they felt that had control and if it felt emotionally intense (arousal).

still chose to ignore harassment when possible: "it's disgusting and sometimes I just ignore it, but I wish I could just play without people like that".

Third, some participants described responding socially to harassers, which mentioned 79 times e.g. P15 tried to "...use humour to try and diffuse the situation.". Others reported that they attempted to speak out against harassers. This could be done to resolve the situation: "I'd tell off the harasser and advise the other(s) to block and report" (P19), but could also escalate it: "they shouted homophobic abuse at me so I told them to f*** off. I later received a report against me for bullying" (P16). The final social response some took was to recede from conversation, such as P1 after being subjected to Hate Speech: "the situation made me feel terrible about myself, all because I had a Filipino accent when speaking in English while the rest of the users on the server had clear western accents. I responded by just keeping quiet the entire time because I enjoyed the stuff happening in the server itself. I did not take any other steps afterward, but the experience made me understand that I should talk less while on VRChat when with other people."

The fourth approach taken in response to harassment, described 80 times, was to recede, leave the SVR apps or stop using VR entirely. This was often used by people who expressed being strongly affected by harassment, such as P86: "I felt panicky and I logged off and stayed off for a while". A lack of ability to proactively protect oneself makes guaranteeing one won't be subjected to harassment difficult, which can in turn motivate simply avoiding the platform. Regarding having been verbally abused, P25 wrote that "it made me feel unsafe knowing that there are people doing those that's why I rarely use those platforms". The implications of a lack of effective and preventative safety features could, thus, deter potential users from engaging with SVR.

4.1.3 Experiences of Witnesses versus Victims

Interestingly, Likert scale responses varied little based on whether the participant was victim or witness to harassment. To further investigate this, we used five linear mixed-effect models to observe if there was any main effect of being a victim or witness on responses to the five harassment experience Likert scales with the participant modeled as a random intercept, following an Aligned Rank Transform due to the data being non-parametric [27]. We found no effect on consistency with offline experiences ($F = 0.32, p = 0.57$), perceived realism ($F = 1.57, p = 0.21$), pleasantness ($F < 0.00, p = 0.99$), intensity ($F = 1.20, p = 0.27$) or control ($F = 2.60, p = 0.11$).

Despite this, witness-specific experiences did emerge from participants' qualitative responses. Most often participants reported feeling sympathy for the victim, such as P62 who described witnessing somebody experience "an inappropriate moment which should not of happened. I felt awful for her", P62 who found that witnessing hate speech "made me feel bad for the person", or P2 who saw another user experience sexualised language: "it made me feel bad for the victim. The victim left the VR room", although this was not universal: "I would just laugh if it happened to others" (P35). Being a witness could also be uncomfortable for more personal reasons: "a group was making jokes to a user saying that he was autistic. It was hurtful since I have a family member who is disabled" (P80). The most prevalent action reportedly taken by those witness harassment was to support victims and educate them about interventions they could take, or warn others. For example, after seeing somebody subjected to hate speech, P79 "told him to block that person and afterwards we reported him", P19 described "warning others and advising to report and block (especially if minors are present)". In response to witnessing virtual crashing, P92 wrote they "try to gather information about the situation and warn close by friends/moderators about the situation". Others did not directly support the victim, but still took action or intervened. P63 used built in safety features "the situation made me feel bad for the person it was happening to. I responded by reporting the person".

A subset of witnesses expressed specific concern about the presence of children in harassment scenarios, feeling that they were potentially more vulnerable and could be exposed to serious harassment that, while normalised in VR, would be considered very unusual offline. When discussing witnessing virtual violence, P29 commented that "things can sometimes escalate out of control which is also frightening especially to younger people". P94 highlighted a need for "adult only VR social spaces", advising that "kids should only be in human moderated spaces for their own safety while adults should have more freedom in their spaces". Concerns also arose about whether children would know how to respond in harassment scenarios, prompting respondents to try and

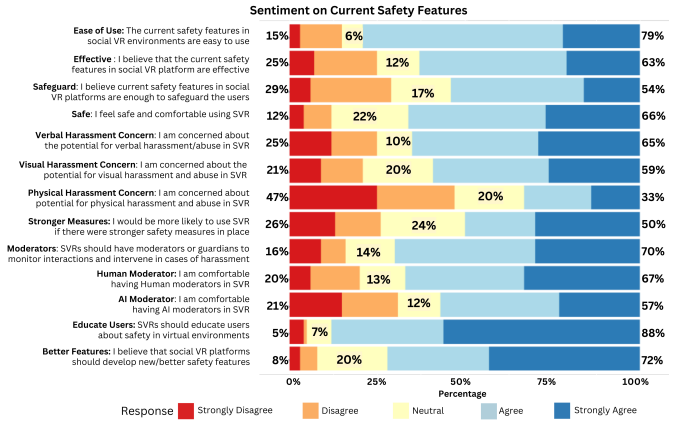

Fig. 3: Distribution of Likert scale responses to questions assessing participant general sentiment on current safety features in SVR.

educate them in advance: "I explained to my minor the dangers of playing online and that she should be cautious. I showed her how to block and report" (P51) Others also noted that it allowed children to act as harassers: "almost no one that moderates the users that are young" (P92). For some this harassment was viewed as less impactful: "threats from children over internet don't mean much" (P96), but others were still affected by it: "it did make me feel a bit sad [...] because these are mostly kids who say this kind of stuff. Parents need to supervise their kids more" (P50).

4.2 Social VR Safety Features

Before giving feedback about individual features, participants were surveyed about their general sentiment toward the state of safety features in SVR. Distribution of participant Likert responses to these questions are shown on Figure 3. General sentiment on safety features was positive, but far from universally so. 63% agreed that current safety features are effective, while 37% disagreed or were neutral. 79% of respondents also agreed that current features were generally easy to use, while 54% agreed they were enough to safeguard users and 29% disagreed. While 66% agreed the felt safe and comfortable in SVR, the many respondents were still concerned about verbal (65%), visual (59%) and physical (33%) harassment. Improving safety feature effectiveness may tackle these concerns, with 50% agreeing that stronger features would increase their usage, 72% agreeing that platforms should develop better features and 88% supporting better user education about safety features. Regarding moderation as a potential solution, many more respondents were more comfortable with human moderation (67%) than AI moderation (21%).

Following these general questions, participants gave individual Likert scale ratings for usefulness, effectiveness, and gave qualitative feedback for each safety feature they had used. These informed the construction of three key themes, discussed below.

4.2.1 Similarly Preferable, but not Similarly Accessed

Interestingly, we found that participants felt most safety features were similarly effective and useful. With the exception of AI Moderation, features were considered effective by 60% to 92% of respondents who had used them, and considered useful by 53% to 90% (see Figure 4). Despite similar positive perceptions, however, there are stark differences in the number of respondents who had experience with each feature. Only one feature had been used by the majority of respondents, Mute and Block (61%), while many have also used Reporting (48%), Social Spaces (40%) or Vote Kick (38%), all of which are interventions drawn from a long history of online games that precede SVR and are available across almost all SVR platforms. As already discussed in Section 4.1, many participants regarded these features as default practise and would even proactively teach them to others.

Meanwhile, features that are more specific to SVR were still thought to be effective but were far less used, such as Intimacy Proxemics (25%),

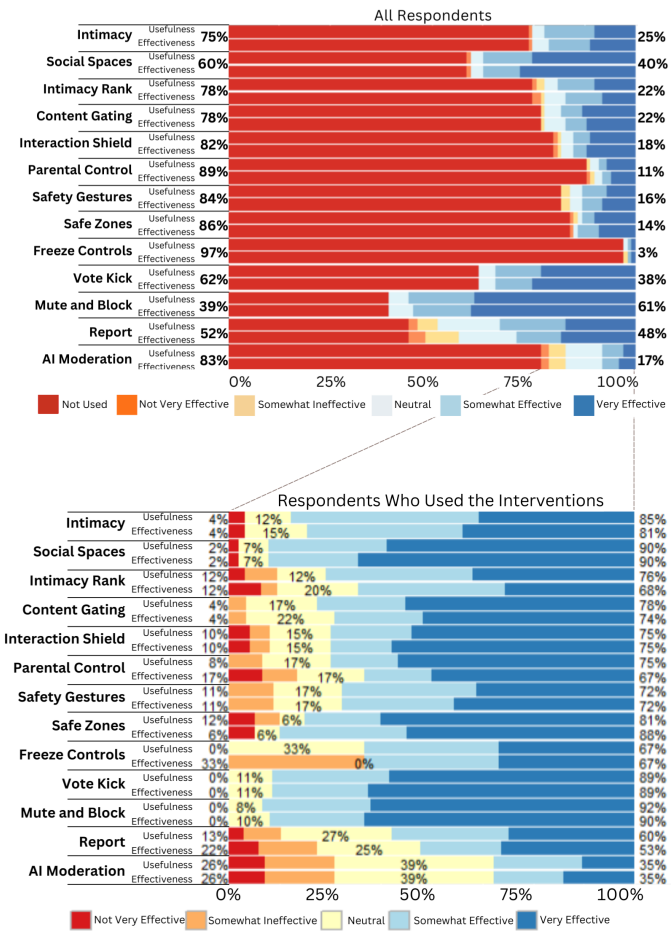

Fig. 4: TOP: Shows the proportion of participant usage of each safety feature (indicated by the left percentage and red bar), alongside their Likert scale ratings for effectiveness and usefulness for each feature. BOTTOM: Proportional Likert scale feedback for effectiveness and usefulness for each safety feature, where only responses from participants who had used each feature are shown.

Interaction Shields (18%) or Safety Gestures (16%). This may be due to features not being present across many platforms, such as Intimacy Rank which is only available on one major platform (see Table 2), while participants like P77 hypothesised that "I didn't feel like enough people understand" the Safety Gestures feature. Our findings indicate that, while the most used safety features are established features with inertia from online game and social media, there are many existing SVR safety features which current users find effective and useful. However, the average user may not be aware of, or have access to, them.

Participant feedback contextualised the lower effectiveness and usefulness ratings for Report and AI Moderation. While a commonly used and understood feature, many highlighted that the Report feature is too slow to take effect, such as P29 "can take a while to process which is not good when immediate action is needed", or P86 who said it was "never fast enough". The effectiveness of the feature was also questioned. While some felt it was " relatively effective" (P72) or "useful for extreme cases" (P13), this was not consistent perception, with P30 reporting "nothing changed" and P63 writing "I report a lot of people but nothing ever happens". Regarding AI Moderation, participants primarily had experience with expressed concern that it may underperform human moderators, particularly caused it to issue punishments for benign or non-harassment actions. P35 noted that "when chatting in a different language as it can flag some valid words (sometimes in English) as inappropriate and it gets really annoying", while P46 wrote "When it gets edge cases wrong its super painful".

4.2.2 Social Sculpting of User Experiences

Surprisingly, while many participants were motivated to use safety features to tackle harassment (mentioned 73 times), a similar number also used them to sculpt their user experience inside of SVR (62 times), selectively choosing which users they perceive and can interact with. Social Spaces provided an opportunity for this akin to offline social interaction in private spaces: "I invited my friends to use the social spaces feature in order to comfortably play with them" (P7). Other features allow a similar 'friends-only' approach in shared social spaces, such as P18 who used Intimacy Proxemics to "disallow users I don't know away from me while letting my friends get close". Similarly, the desire to exclude unwanted users was often mentioned. P38 used Interaction Shields so that they "don't have to interact with people I don't want to". Sometimes features were deployed on users deemed disruptive or annoying even if they have not enacted any specific harassment or broken platform rules, such as P10 who described using the Vote Kick feature "if someone is cringe" to exclude an unwanted user from a team activity. It is notable, however, that using these features also provided respondents with a preemptive protection against any future harassment. For example, P70 used Social Spaces to "mitigate any unpleasant situations by excluding people that are not known".

Others used safety features to improve their experience with the VR environment or activity. Several participants used the Intimacy Proxemics feature to allow them to more clearly see and move within the VR environment, such as P1: "it has helped my experience become less jarring when interacting with users as it always makes sure they're at a certain distance before I can actually see their avatar", as well as P15: "used to prevent clashes of graphical meshes. If your avatar is overlaid on mine, it may be difficult to see/move". Similarly, P27 made use of private Social Spaces when they "want to have a noise-free environment" and P35 used Content Gating to combat that "some Roblox games have a lot of visual clutter that are really useless". P35 even described using the Safe Zone feature to gain a competitive advantage in a player-versus-player VR game "sometimes I can avoid encountering people who wants to kill me by just teleporting". Overall, respondents used these features to grant them increased agency over their SVR experiences in ways that would not be possible in offline social settings, highlighted by P85 who wrote "it was nice to have that control", and by P72: "being able to choose who can interact with you and the number of people that can do so simultaneously is pretty good".

4.2.3 Desirable Qualities for Social VR Safety Features

Transparency and Feedback: Respondents shared concerns about a lack of clarity or transparency in how specific features worked or if they were effective. The most opaque feature was Reporting and participants wished for the effectiveness to be more verifiable, echoing prior work [37]. For example, P17 wrote "I don't ever hear or see the result of the report so I have no idea if the person I reported got punished for their actions", while P81 compared the feature to its social media equivalent: "maybe more transparency? Some social media report back to you on whether or not they investigated the report and banned the person". The lack of transparency resulted in feature use based only on faith: "All that people within that instance can do is report and hope" (P19).

As discussed in Section 4.2.1, respondents expressed concerns about AI Moderation making erroneous judgements and P54 suggested that platforms should "provide insight into the moderation process to build trust". P19 highlighted further transparency issues, recalling that "recently I believe VRChat had gave update on new moderation features", but that they "haven't been able to find the original dev post on their official website". P72 offered suggestions to improve clarity for other features, such as changing the Intimacy Rank feature to "let good players get noticed by more people and to mitigate bad people", allowing players to sculpt their social environment more easily. They also suggested that features which alert users via sound cues, such as Vote Kick, should be "more apparent to users with sound off". In contrast, features like Mute and Block are instantly verifiable, improving perceived effectiveness.

Not Only Reactive: While legacy features like Mute and Block and are considered effective and easy to use, these are primarily used in response to a harassment incident and do not proactively prevent such incidents. This reactivity led to users like P25 feeling unsafe in SVR: "It made me feel unsafe knowing that there are people doing those that's why I rarely use those platforms". P50 felt that "usually all the damage has already been done" and P85 agreed that "the problem isn't solved by kicking one person. You are just as likely to run into someone else who is terrible around the corner". Even if a harasser is blocked, muted or punished, P54 noted that harassers could make "another account and another and another, so I guess it could be easy to break the rules". The lack of prominent preventative interventions may contribute to respondents' perception that harassment in SVR is inevitable and inherent (see Sec. 4.1.1). This means, however, that VR safety features that allow users to specify who they can interact with provide unique utility, meaning participants like P38 "don't have to interact with people I don't want to". This in turn may motivate the promotion and deployment of SVR features that allow for preemptive social sculpting, such as Social Spaces and Interaction Shields.

Quick and Easy: A key quality which motivated participant preference for prevalent and established features was how quickly and easily they could be deployed. Mute and Block was described by P19 as "simple and effective" and by P64 as "quick and effective", while P8 wrote it was "easy to use since you only have to click a few buttons". Speed of deployment was also important. For example, while the effectiveness of Reporting was questioned, the speed of use allowed respondents like P69 and P57 to "immediately" report harassers, making it easy to utilise regardless of provable effectiveness. Awareness of these features was also very high, with many considering Muting, Blocking or Reporting harassers to be a default response (see Sec. 4.1.2). This suggests that VR-specific safety features which are considered effective but are under-used should aim to match the visibility, convenience and usability of established features. For example, Safety Gestures are only available on one major platform (see Table 2), leading to low access and awareness, while some respondents who used it complained that "it's too slow, it's quicker to pull up a users profile and blocking them"(P18) and needed to be "less clunky and faster" (P86).

Customisation: Finally, participants expressed the desire for more customisation options from features to more finely tune their experiences. P87 wished for more personalised Social Spaces: "it would be nice to have a highly customisable creator for your own unique spaces" and P86 similarly asked for "more customisation" of Content Gating. Participants also floated the idea of features having the option to set custom presets, whereby users could switch between different levels of experience sculpting or content/user filtering based on their current context or mood. P19 wrote: "custom safety setting loadouts would be wonderful. Sometimes I would like to at least have 3 variations of custom safety settings", and P24 felt that " it is better if user can select by themselves the content they want to ignore/ gate". Others wished for similar but automatic context-sensitive enforcement, such as P8 who suggested having "a better AI modelling where the AI could measure the magnitude of how offensive (in a sexual sense) a certain conversation would be despite using non-sexual words".

5 Discussion and Future Work

5.1 Harassment Currently an Inherent Reality of SVR

Our survey highlighted how common and prevalent harassment in SVR is. Many participants were habituated to this situation, reporting incidents as an inherent and unavoidable part of SVR in the same way that it is rife on social media and on traditional multiplayer computer games. Of these incidents, our survey confirms an overwhelming majority of verbal abuses, ranging from hate speech and insults, but also confirms and quantifies the existence of VR-specific highly distressing abuses [12, 24]. Some participants commented on the inconsistency between the consequences of harassment between the real-world and SVR, leading to higher prevalence. In what could be seen as a virtual *broken window* effect, the lack of enforcement of social norms and shunning of harassers leads to an impression of harassment having no consequences, which may beget further harassment. We note however

that all participants did not share the expectation that verbal harassment would be normalised in SVR. Some participants still expressed shock and surprise (e.g. P45, P78) at the severity of harassment, inherently treating them as akin to real-life experience. This disparity in experiences questions the boundary of cyber-experiences, with participants appearing to interpret and rationalise their SVR experiences with analogies to either online social media or online gaming experiences, where harassment is seen as rife and expected, or to real embodied social experiences. The difference in expectations warrants further study.

5.2 (Under-)Use of Safety Features and Social Sculpting

From participants' reports on their use of safety features in SVR environments, such as blocking, muting, or preemptively excluding others, emerges the impression that these safety features are double-edged swords in shaping the virtual social landscape. On the one hand, they provide essential - and importantly, preventive - protection against harassment, allowing users to curate their experiences and avoid inappropriate or harmful interactions. However, such preemptive exclusions may lead to fragmented social experiences where users are denied the opportunity to engage and interact due to factors like in-group dynamics, cliques, or even demographic biases. For example, when a user is preemptively blocked from entering a private VR room simply because their chosen avatar or accent differs from the established group's norms. We characterise this use SVR safety tools is as a form of "individual social sculpting", where users' perceptions of shared spaces can vary significantly based on their personal choices to exclude or interact with others. While this system empowers users to maintain a sense of safety, it also risks creating fragmented communities where the potential for meaningful interaction is diminished.

Ideally, we would argue that SVR should evolve toward a model of "community-based social sculpting", where consequences for antisocial behaviour are more tangible, explicit and accountable, and harassers are deterred from continuing such behaviour. In this model, the collective actions of the community would foster a more cohesive social environment, disincentivising harassment and promoting positive interactions. Without such a shift, the risk is that SVR would continue to allow harmful behaviours to persist unchecked, especially for new users who have yet to engage and setup their individualised preferences for social engagement, while longer-term users split in fragmented communities, undermining the potential of SVR to truly emulate the richness of real-world social interaction as an open and inclusive medium.

5.3 Leveraging Witness' Perception to Improve Safety

A surprising insight of our survey was the observation that witnesses of SVR harassment often perceive the severity of incidents in a manner similar to how victims experience them. We believe that this makes witnesses as valuable allies in addressing and assessing harassment. Encouraging bystanders to report these incidents can play a crucial role in victim support and advocacy, reducing the burden on victims - especially newer and younger users, less familiar with safety tools and reporting features - to prove their experiences. Platforms could support this by implementing systems that allow witnesses to anonymously offer support or testify in harassment cases, thus helping to create a safer environment. Additionally, real-time alerts could be used to involve bystanders in immediate interventions, potentially preventing situations from escalating further. An example of this approach can be seen prior online game safety features such as the League of Legends Tribunal system [28, 30], which engaged players in reviewing and responding to reports of misconduct, and provides a fascinating real-world case study of community-led judgment and social norm enforcement in a platform-supported structured setting.

However, the implementation of such tools comes with potential limitations. First, concerns remain that groups of harassers could manipulate these systems, falsely portraying a victim as the perpetrator. Additionally, some users respond to harassment by attempting to defuse situations through humor or by directly confronting the harassers, demonstrating the varied ways in which individuals react to and manage harassment in SVR settings. Second, in our sample, there were less witnesses of physical and visual harassment compared to verbal harassment. We suspect this is due to the ephemeral nature and context of these types of interactions. Verbal harassment tends to occur in open environments where multiple users can overhear or witness the interaction, while physical and visual harassment may take place in private or less visible spaces, or might be more difficult to detect as they are tied to individual perception and embodiment.

The experiences of victims and bystanders prompt us to consider the harassers' perspective as well. Survey participants, including witnesses and bystanders, show a pervasive expectation of harassment. If harassers similarly perceive harassment as an expected or acceptable behaviour, it can reinforce a toxic culture where they feel empowered to act as such without fear of consequences. This underlines the need for a cultural shift where harassment is not tolerated.

5.4 Transparency and Feedback in Safety Tools

Our findings reveal a significant demand from users for transparency and feedback after reporting incidents of harassment. Participants frequently expressed dissatisfaction with the lack of follow-up, echoing concerns raised in previous research that emphasises the importance of visible outcomes to build trust in reporting mechanisms as they encourage users to engage with reporting systems [20], and increase their sense of agency and trust in the reporting system [31]. This highlights the need for "immediate user protection" and, when not feasible, promoting "transparency through timely, clear feedback" on report status and outcomes.

5.5 Best Practices for Effective SVR Safety Tools

Building on our findings, we recommend the following practices for SVR safety tools: a) Ensure *immediate user protection*, offering swift responses to threats or harassment, b) Promote *transparency* by providing feedback on the outcomes of reported incidents and tool usage, c) Support versatile use through *proactive, real-time, and reactive functions*, enabling users to address incidents before, during, and after they occur, d) *Standardise features across platforms* to create a seamless, consistent experience for users across platforms, and e) *Encourage community involvement* to foster trust and cooperation in moderation and reporting processes.

6 Conclusion

SVR promises to expand the boundaries of social interaction, but also raises concerns about harassment and abuse, especially given the heightened psychological and emotional impact of immersive experiences. While safety tools have been introduced to address these risks, their effectiveness and user experience had not been fully evaluated. Through an online survey, we provided the first evidence on how these tools operate in practice. Our findings confirm the prevalence of harassment in SVR, offer insights into the use of reactive and proactive safety measures, and show how preventive tools are used to "socially sculpt" spaces—sometimes at the cost of excluding others. We also highlight the important role of bystanders in fostering community-based social norms. We concluded by a set of best practices for SVR safety tools, which will ultimately reinforce safety measures, and improve user retention in SVR.

Acknowledgments

This work was supported by REPHRAIN: The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online under UKRI (grant number EP/V011189/1), and partly by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents (grant number EP/S02266X/1), and a 2020 Meta Research Award on Responsible Innovation.

Dr Mark McGill's research time was supported by UK Research and Innovation (UKRI) under the UK Government's Horizon Europe funding guarantee (AUGSOC) [EP/Z000068/1].

References

[1] A girl was allegedly raped in the metaverse. is this the beginning of a dark new future? | nancy jo sales | the guardian. Last Accessed: 23-07-2024. 1, 2

[2] The metaverse has a groping problem already | mit technology review, 2021. Last Accessed: 23-07-2024. 1, 2

[3] Comfort and safety — rec room, 2023. Last Accessed: 30-08-2023. 1

[4] Vrchat safety and trust system, 2023. Last Accessed: 30-08-2023. 1

[5] Comfort and safety — rec room, 2024. Last Accessed: 24-01-2024. 1

[6] Playstation vr2, 2024. Last Accessed: 21-01-2024. 2

[7] Police investigate virtual sex assault on girl's avatar - bbc news, 2024. Last Accessed: 23-07-2024. 1, 2

[8] Y. Abdelrahman, F. Mathis, P. Knierim, A. Kettler, F. Alt, and M. Khamis. Cuevr: Studying the usability of cue-based authentication for virtual reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, AVI '22, New York, NY, USA, 2022. Association for Computing Machinery. 5

[9] M. Abraham, M. Mcgill, and M. Khamis. What you experience is what we collect: User experience based fine-grained permissions for everyday augmented reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. 2

[10] M. Abraham, P. Saeghe, M. Mcgill, and M. Khamis. Implications of xr on privacy, security and behaviour: Insights from experts. In *Nordic Human-Computer Interaction Conference*, NordiCHI '22, New York, NY, USA, 2022. Association for Computing Machinery. 2

[11] D. Acena and G. Freeman. "in my safe space": Social support for lgbtq users in social virtual reality. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–6, 2021. 2

[12] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz. Harassment in social virtual reality: Challenges for platform governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. 1, 2, 3, 8

[13] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, (3):77–101, 2006. 3

[14] V. Braun and V. Clarke. *Successful Qualitative Research: A Practical guide for Beginners*. SAGE Publications, London, 2013. 3

[15] Q. Chen, J. Cai, and G. Jacucci. "people are way too obsessed with rank": Trust system in social virtual reality. *Computer Supported Cooperative Work (CSCW)*, pages 1–33, 2024. 2

[16] C. Fiani, R. Bretin, S. A. Macdonald, M. Khamis, and M. Mcgill. "pikachu would electrocute people who are misbehaving": Expert, guardian and child perspectives on automated embodied moderators for safeguarding children in social virtual reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2

[17] C. Fiani, R. Bretin, M. Mcgill, and M. Khamis. Big buddy: Exploring child reactions and parental perceptions towards a simulated embodied moderating system for social virtual reality. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, IDC '23, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2

[18] C. Fiani and S. Marsella. Investigating the non-verbal behavior features of bullying for the development of an automatic recognition system in social virtual reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, AVI 2022, New York, NY, USA, 2022. Association for Computing Machinery. 2

[19] C. Fiani, P. Saeghe, M. McGill, and M. Khamis. Exploring the perspectives of social vr-aware non-parent adults and parents on children's use of social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):54, April 2024. 25 pages. 1, 2, 3

[20] 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*, 19(8):1290–1307, 2017. 9

[21] G. Freeman, L. Li, and K. Schulenberg. "i have abused someone who abused me": Understanding people who have experienced both sides of harassment accusations in social vr. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW):Article, 2025. Preprint. 2

[22] G. Freeman and D. Maloney. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on human-computer interaction*, 4(CSCW3):1–27, 2021. 2

[23] G. Freeman, D. Maloney, D. Acena, and C. Barwulor. (re) discovering the physical body online: Strategies and challenges to approach non-cisgender identity in social virtual reality. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–15, 2022. 2

[24] G. Freeman, S. Zamanifard, D. Maloney, and D. Acena. Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022. 1, 2, 3, 8

[25] C. George, M. Khamis, D. Buschek, and H. Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 277–285, 2019. 5

[26] A. Hobson. Phantoms, crashers, and harassers: Emergent governance of social spaces in virtual reality. Last Accessed: 25-11-2024, 2020. 2, 3

[27] D. G. James Higgins Jacok Wobbrock, Leah Findlater. The aligned rank transform for nonparametric factorial analyses using only anova procedures. *CHI 2011*, pages 1–5, 2011. 6

[28] M. Johansson, H. Verhagen, and Y. Kou. I am being watched by the tribunal: Trust and control in multiplayer online battle arena games. In *FDG*, 2015. 9

[29] kentbye. 1057: What parents should know about social vr, understanding social vr harassment, & parental guidance for the metaverse with lance g. powell, jr. – voices of vr podcast, 2022. Last Accessed: 25-11-2024. 2, 3

[30] Y. Kou and B. A. Nardi. Governance in league of legends: A hybrid system. *FDG*, 7:1, 2014. 9

[31] C. Lackey and S. H. Taylor. Algorithmic folk theories of online harassment: How social media algorithms enable online harassment and prevent intervention. *AoIR Selected Papers of Internet Research*, 2023. 9

[32] D. Maloney and G. Freeman. Falling asleep together: What makes activities in social virtual reality meaningful to users. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, page 510–521, New York, NY, USA, 2020. Association for Computing Machinery. 2

[33] D. Maloney, G. Freeman, and A. Robb. It is complicated: Interacting with children in social virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 343–347, Atlanta, GA, USA, 2020. IEEE. 1, 2, 3

[34] D. Maloney, G. Freeman, and A. Robb. A virtual space for all: Exploring children's experience in social virtual reality. *CHI PLAY 2020 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 472–483, 11 2020. 3

[35] D. Maloney, G. Freeman, and A. Robb. Stay connected in an immersive world: Why teenagers engage in social virtual reality. IDC '21, pages 69–79. Association for Computing Machinery, Inc, June 2021. 1, 2, 3

[36] D. Maloney, G. Freeman, and D. Y. Wohn. "talking without a voice": Understanding non-verbal communication in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 4, 10 2020. 2, 3

[37] K. Marky, S. Macdonald, Y. Abdrabou, and M. Khamis. In the quest to protect users from side-channel attacks–a user-centred design space to mitigate thermal attacks on public payment terminals. In *32nd usenix security symposium (usenix security 23)*, pages 5235–5252, 2023. 8

[38] F. Mathis, J. H. Williamson, K. Vaniea, and M. Khamis. Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing. *ACM Trans. Comput.-Hum. Interact.*, 28(1), Jan. 2021. 5

[39] J. O'Hagan, F. Mathis, and M. McGill. User reviews as a reporting mechanism for emergent issues within social vr communities. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, MUM '23, page 236–243, New York, NY, USA, 2023. Association for Computing Machinery. 1

[40] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, and M. Elsherief. Challenges of moderating social virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3

[41] K. Schulenberg, G. Freeman, L. Li, and C. Barwulor. " creepy towards my avatar body, creepy towards my body": How women experience and manage harassment risks in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29, 2023. 2, 3

[42] 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. page 17, 2023. 2

[43] A. Singh and J. O'Hagan. Exploring topic modelling of user reviews as a monitoring mechanism for emergent issues within social vr communities. *arXiv preprint arXiv:2406.03994*, 2024. 1

[44] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009. 1

[45] 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. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023. 1, 2, 3