Highlights

Expert-Driven Design Considerations for Real-Time Safety Features in Social VR to Safeguard Children against Individual and Group Harassment

Cristina Fiani, Joseph O'Hagan, Mark McGill, Mohamed Khamis

- We report on results from interviews with 16 experts reflecting on social VR safety features.
- We produce seven expert-derived guidelines for child-friendly safety features in social VR.
- Current safety features are inaccurate, slow, misusable, circumventable & breach privacy.
- Safety features for children should be child-friendly, discreet, prompt, educative & empowering.

Expert-Driven Design Considerations for Real-Time Safety Features in Social VR to Safeguard Children against Individual and Group Harassment

Cristina Fiani[a,*]

London, UK

Joseph O'Hagan[a], Mark McGill[a], Mohamed Khamis[a]

[a] *University of Glasgow*

Abstract

Safety features in social VR vary across platforms in terms of what protections are offered and how they are enacted, with little publicised evidence validating such differences. With the increase in children using social VR, the need to develop safety features designed around their needs, rather than targeting adult users, becomes crucial. Accordingly, we conducted in-depth interviews (n=16) with experts in child online safety and psychology to explore child-targeted safety features. We examined expert perceptions of ten existing real-time social VR safety features, drawn from two popular social VR platforms. We then asked experts to propose recommendations for and to reflect on safeguarding children from harassment in social VR, giving specific consideration to a) individual versus group harassment, as the impact can be amplified and the social interaction dynamics differ; b) the perspectives of harasser versus victim. We contribute seven expert-derived guidelines for the design of safety features to safeguard children in social VR.

Keywords: child online safety, social virtual reality, virtual reality, metaverse, expert perspectives, safety-enhancing technologies

1. Introduction

Children[1] are growing up in an increasingly connected world, changing the way they interact, communicate and socialise [1, 2, 3]. Social Virtual Reality (VR) is transforming social interactions, from 2D social media to 3D digital experiences through the use of VR head-mounted devices [3, 2, 4], offering users full immersion and embodiment. *Immersion* is the extent to which a system accurately replicates real-world sensations and tracking, enhancing the user's sense of physical presence in a virtual environment [5]. Immersion can foster a strong sense of *presence*, often described as the feeling of "being there" [5], which, in turn, intensifies the user's emotional responses [6, 7]. For

*Corresponding author

Email addresses: cristina.fiani99@gmail.com (Cristina Fiani), joseph.ohagan@glasgow.ac.uk (Joseph O'Hagan), mark.mcgill@glasgow.ac.uk (Mark McGill), mohamed.khamis@glasgow.ac.uk (Mohamed Khamis)

[1] In this paper, "children" refers to minors under 16. We consider child "one-to-one" harassment as an individual targeting a child and "group" harassment as a group of users targeting a child.

children, highly immersive and captivating environments can feel more vivid compared to adults and can sometimes blur the line between virtual and real experiences [8]. *Embodiment* refers to the sense of having a physical body, allowing users to control a digital avatar through their real-world movements and engage socially [9]. For children, this sense of embodiment can be particularly influential, as they may perceive the embodied digital agent as a real person rather than a virtual entity. This perception can shape their decision-making processes and enhance the perceived realism of their experiences during interactions [8].

According to a UK study conducted in 2022, more than a fifth of 5–10-year-olds (21%) already have a VR headset of their own or have asked for one [10]. Initially designed for adults, social VR has become widely used by teenagers and children [2, 3]. A survey of adult VR users showed that amongst their children who are 13 years old or younger (N=50), almost half (43%) use social VR at least every two weeks [11].

Recent events and studies have shown the emergence of harassment in social VR not only towards adults [12, 11, 13, 14] but also children [2, 15, 13, 11, 16, 17]. Yet, there are gaps in the literature that we address in this paper: the lack of a) child-friendly interventions[2] and b) assessments of the efficacy of existing safety tools to protect from harassment in social VR. While platforms and developers have put in place codes of conduct and safety features that can be used in real-time[3] (e.g., blocking, muting, protective bubble), research has highlighted issues and the limitations of these tools [18, 2]. For instance, age restrictions are not followed or effectively enforced, and most tools, designed for teenagers and adults, rely on self-moderation [2]. Complex design features can lead to children unintentionally exposing themselves to safety risks, and the abuse or misuse of safety features in social VR can pose further challenges to managing safety risks [18]. Tools also vary across platforms [19, 20, 21], potentially leading to user unfamiliarity and not knowing how to effectively protect themselves during challenging situations.

As more children engage with social VR [11], and given that VR can blur the boundaries between the real and virtual worlds especially for children [8], the absence of proper safety mechanisms could expose them to inappropriate content, harmful social interactions, or harassment scenarios [2, 15, 13, 11, 16, 17] they may not be equipped to fully understand or navigate.

We address the lack of knowledge on how exactly to address child safety in social VR, considering the perceived effectiveness of real-time safety features for child users in the eyes of 16 expert interviews (5+ years experience in child online safety or psychology). Our experts reflected on the strengths and weaknesses of ten existing safety features (Section 4.2) from two of the most popular social VR platforms (VRChat [22] and Rec Room [23]). Following this, experts propose specific child-friendly solutions for two given scenarios to contextualise harassment towards children in social VR and gain practical insights. In Scenario 1, a child is harassed by an adult being provocative, while in Scenario 2, a child is harassed by a group of teenagers.

We address the following research questions:

RQ1 What are child online safety and psychology experts' perspectives and considerations regarding existing social VR safety features to ensure child safety in one-to-one or group harassment?

RQ2 What do child online safety and psychology experts recommend and propose (beyond the

[2]We consider safety features from two perspectives: 1) prevention (before an incident occurs) and 2) intervention (when and as an incident occurs).

[3]In this paper, we focus on real-time intervention, which can consist of *immediate* actions such as blocking but also *short-time-delayed* (e.g., within a few minutes) interventions (also described as "Boundaries Settings" and "Quick Reactions" in prior work [18])

existing ten safety features from RQ1) for addressing one-to-one or group harassment targeting a child in social VR?

This paper contributes to the field of human-computer interaction and child-computer interaction with insights into safety features' design considerations to safeguard children in harassment situations in social VR. Drawn from 16 experts' interviews, considering a child in a situation of "one-to-one" or "group" harassment, we: 1) report an expert-led evaluation of the suitability and efficacy of ten existing safety features in social VR derived from Rec Room and VRChat (Section 4.2), 2) provide insights into how social VR safety tools should be adapted to work across both individual and group harassment scenarios and 3) consider the impact of such interventions across both child victims and harassers, and the potential negative consequences of their use. We produce seven expert-derived guidelines for child-friendly safety features in social VR to better safeguard child users facing an incident of harassment by an individual or by a group.

2. Terminology

- "Experts" are individuals with at least 5 years of experience (research and/or industry) in child development/educational psychology, psychiatry and social media/social VR online safety.
- The concepts of "cyberbullying" and "online harassment" are still developing as technology evolves and are subjective (See subsection 3.1). We use "harassment in social VR" to describe offensive behaviour in social VR identified with three categories in recent research [12]: verbal, environmental, and physical. Although we acknowledge the nuanced difference between "intentional" and "unintentional" harassment, especially as highlighted by experts in study #2, the focus of this paper remains on the child's victim experience, irrespective of intent.
- In this paper, we focus on real-time intervention, which can consist of *immediate* actions such as blocking but also *short-time-delayed* (e.g., within a few minutes) interventions from human moderation (also described as "Boundaries Settings" and "Quick Reactions" in prior work [18]). However, when participants raise prevention or after-the-fact responses, we consider these contributions and address them in the discussions.

3. Related Work

3.1. Harassment Definitions

Various terminology exists to refer to offensive behaviours offline and online. The Protection from Harassment Act 1997 [24] considers behaviour to be *harassment* if it causes a person to feel threatened, distressed, humiliated, or fearful of further harm. *Bullying* was initially referred to as repeated, intentional acts of aggression carried out by an individual or group against a target repeatedly over time [25]. A definition was later proposed as: "an aggressive goal-directed behaviour that harms another individual within the context of a power imbalance", considering not only the frequency and intensity of bullying and victimisation, but also the victim perceptions of harm, to capture all bullying behaviours [26].

These terms were then expanded into the online world and are still evolving. *Cyberbullying* is a form of bullying via electronic mediums [27]. However, research pointed out the challenges with respect to applying the presented definition of *bullying* to *cyberbullying*, for measurement and classification, notably, whether and how the power imbalance criteria in the general definition of traditional bullying can be applied to cyberbullying [28, 27, 26]. Online, power imbalance may

not appear as physical differences, but rather in terms of social status or via anonymity [26]. *Cyberbullying* also encompasses the terms *online bullying* and *online harassment*.

Depending on the medium within which harassment occurs, group harassment has been depicted using various terms. On social networks, it is referred to as *networked harassment*, where an individual is "harassed by a group of people networked" through instant messaging or other forms of posting [29]. In the offline world, terms such as *group aggression* (collective physical or psychological harmful behaviour directed towards an individual [30]) and *mobbing* (form of group aggression characterised by repeated actions against the same victim [31]) are often used in relation to children in schools [31, 32], as forms of bullying.

3.2. Harassment in Social VR towards Children

The unique VR affordances that mimic real face-to-face interactions while still being anonymous have led to new forms of harm. Harassment in social VR can be categorised into three types: 1) verbal harassment (through voice or chat), which is also common in traditional media or 2D video games; 2) physical harassment (through avatars, such as intruding on personal space), and 3) environmental harassment (using virtual objects, such as displaying inappropriate content or throwing virtual items), both of which are unique to the immersive and embodied nature of VR environments [12].

While social VR was initially designed for adults, the number of child social VR users has been substantially increasing [3, 2, 11]. The presence of both children and adults in the same social VR spaces has led to significant risks of harm from adults towards children [18, 3, 33, 34, 35], including sexual harassment or self-disclosure of personal information, but also group pressure between children and teenagers [13, 36, 11]. Based on 30 interviews of social VR users, researchers investigated users' experiences interacting with children. Respondents highlighted the prevalence of young users, particularly on platforms like Rec Room [2]. They observed that children and adolescents frequently faced harassment from older users and were subjected to negative social environments and interactions. A noteworthy concern raised was that in social VR, children encountered social interactions far beyond their maturity level, while the current safety features were ineffective [2]. The impact of harm in social VR towards a child may be amplified compared to an adult. Using a head-mounted device, children aged 6–18 felt greater levels of presence and perceived "realness" in VR compared to adults aged 19-65 [37]. If young children perceive VR as more authentic than adults do, it raises the possibility that they could be more susceptible to the impact of content, which could be amplified by negative content and harm. Another study showed that children displayed increased social responsiveness to VR characters, as opposed to their counterparts on 2D television [38]. Therefore, there is a crucial need to develop child-friendly safety features in social VR, which can only be done after understanding what should be considered when developing those safety features for younger users.

3.3. Existing Safety Features in Social VR and Challenges

Unlike traditional social media platforms, the immersive nature of social VR demands a reevaluation and expansion of existing safety features. Social VR platforms have put in place codes of conduct [39, 40] that include behaviour rules such as "Sexist, racist, discriminatory or harassing language, behaviour, or content is not welcome" and "Players 12 or younger must use a Junior Account" [39]. Lessons learned from conventional social media platforms certainly carry weight for verbal harassment. However, other types of harassment specific to social VR, i.e., physical and

environmental, underscore the necessity for innovative safety features tailored specifically to the space of social VR.

The main child-safety social VR mediation is age limits of 12+ or 13+ for most VR headsets and social VR platforms [41, 21, 23, 42], recently lowered to 10 years old [43]. However, most VR and social VR platforms do not have age-verification systems. This, combined with children's enthusiasm towards new technologies, and parents' lack of awareness, can lead to a gap in knowledge and difficulties for parents to safeguard their child [44].

Social VR platforms have incorporated existing safety social media tools, such as blocking and reporting a user, and developed new safety features to protect their users from physical and environmental harassment, such as personal space bubbles [19, 20, 21]. While little is known about their effectiveness and usage, a recent study involved researchers observing YouTube videos of social VR users' experiences and identified a list of major in-game real-time safety features across four social VR platforms, including VRChat and Rec Room [18]. They grouped social VR safety features into two categories: boundary settings (e.g., intimacy proxemics, trust reputation, social spaces, shield levels, and controls for voice and avatars) and quick-reaction tools (e.g., safety gestures, vote to kick, safety zones, and safety reports) [18]. The researchers highlighted that relying only on proxemics is inadequate since different safety threats require varied social cues due to the VR affordances. Additionally, current features lack intuitive reactive gestures (such as the "talk-to-the-hand gesture" used to block in Rec Room [20]), and adding these could enhance users' ability to respond to threats [18]. Most existing safety tools are reactive, designed to address issues after they occur. This puts pressure on victims to effectively use these tools while under distress, which may be especially challenging for children. The study advocates for a move towards preventive safety mechanisms, incorporating social cues to identify risks early, such as using signal processing data to detect gestures that signal safety concerns [18]. While the study examines the effectiveness of safety features through empirical observations, more research is needed to understand what should be considered for effective safety features, specifically for children, from expert perspectives.

Along with existing mitigation tools, there are human moderators to moderate and address incidents in real-time in the social virtual environments [12, 13]. Yet, only 24% of incidents observed in a study in social VR were addressed by human moderators, highlighting the lack of scalability for human moderators in a growing VR ecosystem and the need for new moderation tools [14]. Automated moderation can address this scalability issue and has, therefore, been proposed as an alternative safety tool to combat harassment in social VR [45, 46].

3.4. Gaps in Prior Work and Aims of this Study

As described above, prior work has shown that there are new forms of virtual harassment (e.g., physical and environmental) that require new safety features compared to 2D social media. This work aims to provide insights from the eyes of experts with relevant backgrounds in child online safety and psychology to better understand what safety tools are suitable and how they can be designed for children.

We explore safety-enhancing technologies for child users of social VR, specifically from expert viewpoints. This diverges from prior studies focusing heavily on children or adult social VR users [47, 46, 3, 15, 35]. We choose to interview experts as they can act as "crystallisation points" of insider knowledge and serve as proxies for broader stakeholder groups, especially in contexts where direct access is limited or sensitive [48].

4. Methods

4.1. Study Design

In May 2023, we identified five common (i.e., they have the same end-goal) and five distinct (i.e., they have a different end-goal) real-time safety features enacted by the users or human and automated tools, also described as "Boundaries Settings" and "Quick Reactions" in prior work [18], after extracting all the safety tools obtained from the Trust & Safety webpages of VRChat [49] and Rec Room [20] (See Table 1 and Table 2). VRChat and Rec Room are noted as top social VR platforms in January 2024 [50, 51], extensively studied in social VR research [18, 13, 52, 15] because of the capabilities of these platforms for enabling customisable, engaging experiences for children and adults, evident from the 2,530 papers found with the keyword "VRChat" on Google Scholar (in May 2023). Rec Room's design, in particular, has child-oriented design graphics that may be appealing to children and would require ensuring child safety. The five common tools available across both platforms also exist in other social VR platforms (e.g., Horizon Worlds [21] includes all five, Gorilla Tag [53] includes some of them). We chose ten for time constraints of the interview.

Table 1: Five common real-time safety features across VRChat [49] and Rec Room [20] (May 2023 - list is non-exhaustive for time and evaluation constraints).

Social VR Safety Feature	**Platform(s)**	**End-Goal [49], [20]**	**High-level Usage [54, 55, 56]**
Common Features			
Block	VRChat, Rec Room	Removes the user from the blocker's virtual environment.	In VRChat and Rec Room, you point at the user (or look up the user in a menu with the list of people), their profile opens, and you click a "Block" button.
Report	VRChat, Rec Room	Allows users to report inappropriate behaviour that violates the platform's code of conduct.	Similarly to blocking, you point at the user (or look up the user in a menu with the list of people), their profile opens, and you click a "Report" button. The user then fills in a form.
Mute	VRChat, Rec Room	Disables a user's voice.	Similarly to blocking, you point at the user (or look up the user in a menu with the list of people), their profile opens, and you click a "Mute" button.
Personal Space Bubble	VRChat, Rec Room	Creates an invisible virtual barrier that prevents other players from getting close.	The user can create a personal space bubble by going to Settings (through their virtual watch in Rec Room and through the "Quick Menu" in VR Chat). The size of the bubble can be adjusted.
Human Moderators	VRChat, Rec Room	Volunteers or staff monitor interactions, review reports and enforce rules [57, 58, 14].	In both platforms, there is an internal team of human moderators, but also community-based volunteer moderators in-game.

Table 2: Five distinct real-time safety features across VRChat [49] and Rec Room [20] (May 2023 - list is non-exhaustive for time and evaluation constraints).

Social VR Safety Feature	**Platform(s)**	**End-Goal [49], [20]**	**How [54, 55, 56]**
Platform-Specific Features			
Votekick	VRChat	Enables users to initiate a room-wide vote to decide on kicking disruptive members.	A user decides that another user is being disruptive and initiates a votekick. Other eligible users in the room are notified and vote "yes" or "no" in a limited time. If the majority votes yes, that player is removed.
Trust Rank	VRChat	Establishes user credibility based on behaviour and interactions within the platform.	The rank is based on activity time, content contributions and other indicators of participation. Higher ranks unlock more abilities. The rank appears on the user's profile, visible to everyone.
Gestures	Rec Room	Provides users with non-verbal communication options.	For example, a quick "talk to the hand" gesture that the user can perform on another player to block them directly.
Automated Voice Moderation	Rec Room	Limit abusive or discriminatory language in public areas.	Uses algorithms to detect and manage abusive language in public rooms.
Enable / Disable Animations	VRChat	Allows users to disable custom animations of other users, which can prevent seeing disruptive animations.	The user can toggle a button in the "Quick Menu" to disable or enable custom animations for specific trust ranks.

4.2. Procedure

We conducted 16 semi-structured in-depth one-to-one online interviews with experts who have at least 5 years of experience (research and/or industry) in child development/educational psychology, psychiatry and social media/social VR online safety. More details about the experts are in Section 4.5 and in Appendix A. Interviews lasted 20-30 minutes, from May to July 2023. We presented slides with the written questions and timers (varying from 2min to 4min) for each question. The interviews all started with a 5-minute presentation about social VR regardless of the experts' social VR knowledge and experience. The researcher introduced social VR opportunities and risks, and showed a 10-second video from YouTube of a child sharing their experience with another adult in social VR `https://www.youtube.com/watch?v=DBV1pTzsevM`. Experts were also shown a video of Rec Room demonstrating the safety features, after the main researcher presented each of the ten tools `https://www.youtube.com/watch?v=T5lw58uruCw`.

The protocol was approved by the university ethics committee. Each expert was compensated with an Amazon voucher of £8.

4.3. Interview Questions

After presenting the list of the ten safety features using illustrations on slides and videos from Rec Room (see Supplementary Material), we questioned experts about a) their perception of existing tools, b) their applicability to individual/group harassment.

- Rank your top 3 most useful tools in your opinion, that address one-to-one harassment towards a child and the least useful. Please explain your top 3 and your least effective: What do you think their strengths and limitations are?

- Repeat for group harassment towards a child.

We then introduced one of two possible scenarios in social VR to help experts understand what kind of incidents can happen in social VR. These scenarios were developed by the main researcher, based on reported events by participants in prior work [3, 2, 11] and include physical or environmental harassment, as these are specific to social VR [12] (Scenarios below).

> **Scenario 1 - Individual Harassment** - (Adult player inappropriate towards a child: Example of environmental harassment) *A 12-year-old child named Tom joins social VR because, as someone with autism, he struggles with the real world. One day, in social VR, Tom meets an adult player who starts asking him personal questions, making suggestive comments and displaying graphic content on a virtual screen when he is present.* [4]

> **Scenario 2 - Group Harassment** - (Group of teenagers bullying a child: Example of physical harassment) *Jane, 10 years old, joins social VR. Jane loves social VR because she is quite shy in the real world; social VR really helps her be more confident. Soon after joining, Jane becomes friends with a group of teenagers who seem friendly. One day, a group starts to turn on Jane, teasing her, surrounding her, getting very close to her avatar, following her in the virtual room.*

We asked experts the following questions:

- From your experience, can you think of a new solution that would be most effective to tackle the scenario?
- How do you think the solution impacts the victim and the harasser?
- Would you want to adapt your solution for different groups? How and why?
- Can you reflect on your solution (gaps, weaknesses, etc.)?

4.4. Analysis

The analysis was completed in six phases, broken down by the two research questions: data familiarisation, generation of initial codes, searching for themes, refining themes, defining and naming themes and writing the report. After transcribing the interviews, the main researcher reviewed the recordings and corrected any mistakes in the auto-generated transcripts. Both the main experimenter and another researcher read and familiarised themselves with the data. The pair of researchers then created individual coding schemes independently using the online qualitative analysis platform QCAmap [59]. The codes generated are words or short phrases that describe an idea. The researchers then collaborated to consolidate their two coding schemes into one combined scheme by collating or distinguishing between codes. A consensus via this process to adopt the final coding scheme based on the more detailed set of codes, adding missing codes or merging codes, before revisiting the transcripts with the final coding scheme.

[4] While prior work has identified situations in social VR settings where adults may be unaware that another user is a minor [3], we do not specify this here in order to keep the discussion open-ended.

As this paper is mainly exploratory, each idea even mentioned by one participant is considered important. The two coders then grouped codes into main categories by using a Miro Board during a meeting, where we organised individual codes on post-it notes and clustered them into relevant categories as patterns emerged. The themes were then developed based on the categories either as one full category developed or by combining one or more categories to reflect a higher-level aspect of the data (See Supplementary Material). All authors further refined both the names and descriptions of the themes, while also considering their grouping, such as combining or subdividing them into new thematic categories.

4.5. Participants

Recruitment. Experts were recruited via LinkedIn from May to July 2023, in English. Participants were found via search with keywords related to disciplines involved in child safety, such as "Child Psychologist", "Online Safety", "Child Online Safety" and "Trust & Safety". After voluntarily signing up via Calendly, they received a Qualtrics form with an information sheet, consent form to sign and demographics questions. We individually interviewed 16 experts, compensating each with an online shop voucher.

Demographics. The detailed demographics of the 16 experts are described in Appendix A, Table A.1 and Table A.2.

We recruited 16 experts (10 female, five male, one non-binary/third gender), with a mean age of 43.3 years (σ=10.5).

Five experts held a Doctorate, eight had a College / University Degree, and three had a Professional Degree. Ten had professions involving child online safety (with five including social VR), seven had a background or profession in psychology / adolescent psychiatry, one in Policy and Public Affairs and one in Security, Trust and Criminology. Ten professionals practised within the United Kingdom, two in the USA, one in Spain, one in Italy, one based in Malta working Worldwide and one in the Republic of Ireland. Experts had a mean Social VR Awareness (1-5 scale, with 5: extremely aware; 1: not aware at all) of 4 (σ=1.5). All experts had at least *a little* experience in VR, seven owned a VR headset, and 12 had at least *a little* experience in social VR.

4.6. Limitations

The rapid evolution of the social VR gaming landscape, where new safety measures are constantly being proposed, means that safety measures released after our study are not covered. While the tools proposed still represent some of the common safety features of popular social VR platforms, new tools merit further evaluation and engagement with experts as well as child users.

Another limitation of this study is that the interview scenarios were intentionally kept broad, which may have led participants to interpret them through the lens of their own experiences and expertise. We did not focus on highly specific scenarios of harm (e.g., reactions to particular iconography or particular words or phrases). While examining more specific scenarios could have enabled more targeted interventions and greater depth, the resulting solutions would have been less generalisable. We therefore chose to maintain a broader framing. As a result, the findings emphasise a breadth of perspectives rather than an in-depth analysis of specific contexts.

Moreover, we acknowledge that experts were not developers or social VR designers, and some may lack hands-on experience with social VR systems. As a result, the findings should be interpreted with caution when generalising beyond the participants' perspectives.

We do not assess the feasibility of the solutions proposed by experts, as feasibility is often influenced by factors such as regulation and technological advancements like Artificial Intelligence (AI). If expert-recommended measures are mandated, platforms may need to invest in additional resources such as human moderation, parental mediation tools, or other safeguards. Therefore, expert suggestions should be viewed as aspirational design goals rooted in considerations of child safety, provoking consideration on the future of safety tools rather than to imply immediate implementation.

Guidelines provided still require validation and exploration through user studies. Users may not use these tools or may have additional needs that are not considered by experts, and longer reflections to tackle the broad range of topics covered would be needed. While some experts were limited in experience with social VR (although they had experience with VR), their expertise and background still offer informed insights, providing their vision of how to protect children in social VR, which future work can build upon.

Finally, the sample size and necessity of conducting the studies in English in the UK may limit the external validity with respect to the global population and applicability to other cultures.

Despite these limitations, the study proposes valuable groundwork for reflecting on social VR safety features for children through the lens of experts.

5. Findings

The first section presents expert perspectives and insights on the ten existing safety features. The second section explores additional solutions proposed by experts beyond these ten features. The third section offers a reflection on these proposed solutions for children in social VR.

5.1. A Comprehensive Review of Ten Existing Safety Features in Social VR for Ensuring Child Safety (RQ1)

Among the ten safety features presented, experts emphasised the usefulness for children's protection in social VR of **a personal space bubble, blocking and muting and human moderation**. Notably, the findings reveal the same top 3 among the list for one-to-one and group harassment (see Figure 1). However, some experts noted that group harassment leads to amplified harm and should be considered as a more serious event. According to experts, the least useful safety features identified are the "vote to kick" option in a one-to-one situation and "trust rank" in group harassment (see Figure 1) because of potential abuse against users. Experts' reflections and review regarding the reasons for choosing tools for one-to-one and group harassment are described in the following sections.

5.1.1. How Existing Tools Contribute to Safety in Social VR

Experts (N=6) noted that safety features in social VR can be effective when the child users have existing awareness from widespread usage in social media and gaming they may be familiar with, or by incorporating real-world metaphors, which can help increase intuitive usage (e.g., using the hand as a stop-sign gesture).

> *"Muting and blocking are tools that have been used in content moderation for ages. I think they are the foundations of everything." (P7 - Child Safety Intelligence Consultant)*

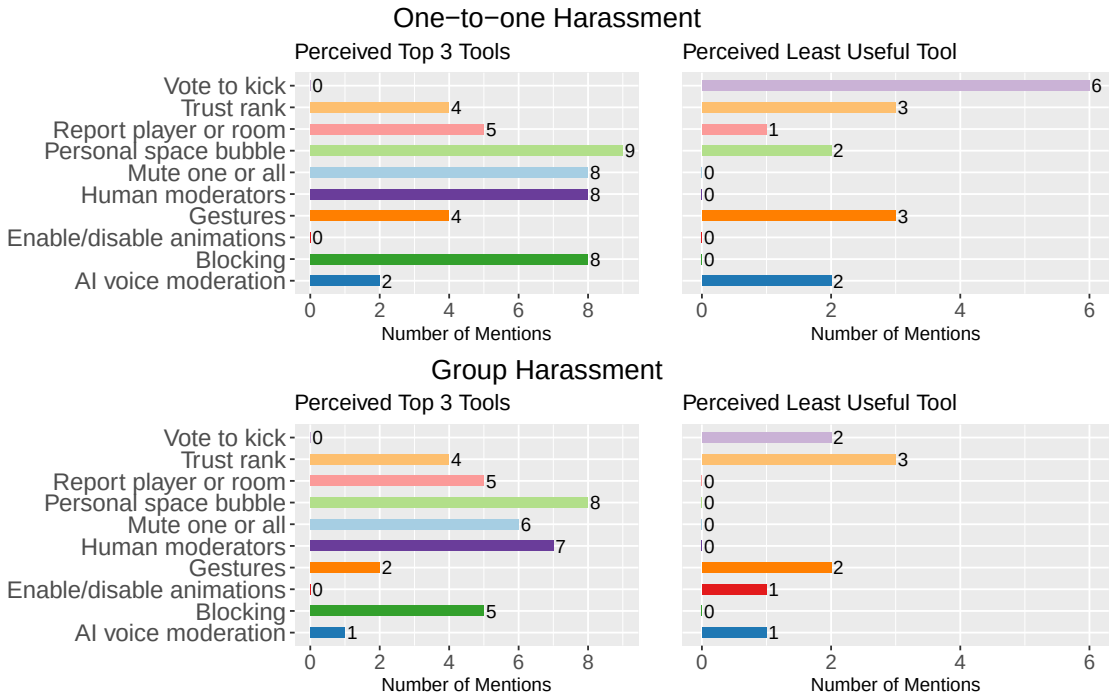

Figure 1: Counts of top 3 expert-selected tools for two harassment scenarios and the least useful tool perceived by experts: addressing one-to-one harassment and group harassment targeting a child.

Experts (N=4) also compared social VR environments to offline worlds and emphasised removing the difference between the online and offline world, where human moderators should be there to moderate children.

Finding 1: According to experts, safety features in social VR are most effective for child users when they leverage existing experience and awareness from other forms of social media they may be familiar with and incorporate real-world metaphors (i.e., drawing parallels with physical world concepts to make usability safety more tangible and relatable).

Seven experts added *human moderators* in their top 3 for both one-to-one and group harassment towards a child. They observed three noteworthy strengths of the presence of human moderators: first, the human presence amplifies the sense of protection and authority (N=7), giving children a dependable source to seek assistance from; second, a flexibility and nuance that allows for negotiation compared to more rigid software (N=4); and third, a deep understanding of the norms within the space (N=1). According to two experts, human moderators could also easily spot group harassment by scanning for group creation and ensuring safety.

Some experts suggested that automated moderation could serve as an alternative for scalability (N=2) or for handling minor tasks such as identifying instances of yelling or profanity (with the bot signalling to lower the volume) (N=1).

Finding 2: Human moderators within the game are an essential component, particularly for younger users, as they provide 1. presence and authority, 2. flexibility and nuance, 3. knowledge of social VR norms. However, automated moderation offers scalability and quick responses compared to human moderators.[a]

[a]Some experts lacked direct experience with social VR. The finding should therefore be interpreted with caution when considering their broader generalisability.

Experts (N=3) highlighted the importance of proactive and reactive tools like muting and blocking for the child users to **control others** in the space and the environment. However, in group harassment, the tools should allow users to moderate the group as a whole. Expert P2 highlighted the need to mute multiple users, and expert P6 suggested the development of blocking multiple users in a radius at once:

> *"Just having the potential to block one avatar at a time is something that is going to have to change in social VR." (P6 - Trainer in Child Online Safety)*

Experts (N=12) specifically positively noted the possibility for a child to **control their own visibility, boundaries and space** while staying in the VR space as a preventive method, for example via a protective bubble that can make avatars disappear if they get close. They also underscored the significance of a child feeling comfortable and acquiring an understanding of personal space. Expert P11 suggested that the protective bubble could have privacy settings for different groups or friends. The child would start with maximum boundaries and adjust those boundaries for certain friends.

While the tools "trust rank" and "vote to kick" were overall the least useful tools according to experts, two participants noted the strength of **sharing the problem with other users** who may have the same problem and resolving it as a group:

"For vote kick or trust rank, I would put them on the same level. A problem that may be just mine can be a problem for the whole group." (P9 - Online Safety Consultant, Clinical Psychologist)

Overall, experts (N=9) highlighted the need of proactive mitigation and instant, easy-to-learn-and-use tools (N=4) to give the child a sense of agency and control. Some experts noted gestures would allow faster actions, especially for beginners, compared to tools like blocking and muting complicated to find in the menus. In contrast to the finding above, three experts also noted the strength of tools when they inform the provider, such as reporting.

Finding 3: Experts considered safety features most effective for child users when they facilitate preventive and proactive user actions by **instantly and easily** giving back to the child the sense of agency: 1) restricting others, 2) controlling your own boundaries and space. For example, in group harassment situations, users should be able to mute and block multiple users at once.

5.1.2. Identified Child-Safety Risks and Weaknesses of Safety Features in Social VR

As children transition from game to game, experts (N=2) emphasised the need for consistent safety across platforms to ensure children are acquainted with the tools and can effortlessly use them.

Children might not be equipped to react and block users in a disruptive situation in social VR. Expert P5 highlighted that trauma in VR can be very real and close to physical; the brain perceives the harassment in VR as real, and the child can freeze:

"Oftentimes in VR, your brain perceives this as reality. When you're verbally harassed, even though the person is not there, people can have a reaction of just freezing, and the blocking would not come to their mind to protect themselves. [...] When it comes to children, they don't have the level of maturity to always block people". (P5 - Child Safety Intelligence Analyst/Consultant)

Social VR platforms should also ensure children can address their concerns and take actions with discretion to avoid escalation or retaliation, according to three experts:

"My only concern with gesture-based tools would be the lack of discretion. [...] I would be worried about retaliation and escalation, especially for children." (P2 - Researcher in social VR harassment)

Another concern (N=2) revolves around children inadvertently entering inappropriate virtual environments, such as strip clubs, where inappropriate content is streamed, and the risk of adults, potentially untrustworthy, interacting with children in virtual rooms.

Finding 4: Experts identified key child-safety risks in social VR, including inconsistent safety features across platforms, potential immaturity of children in handling overwhelming VR environments, the importance of discreet tools to prevent escalation, the concern of adults and children mixing in virtual rooms, and the need for the establishment of social norms.

Experts reflected on the existing safety features and noted several factors that may lead to the ineffectiveness or weaknesses of those tools. The first factor is **Accuracy**. Experts (N=2) noted

that human moderators might not be able to notice a disruptive situation is taking place, due to noise, subtle situations difficult to notice or a lack of resources and the number of moderators. Moreover, human moderators should be trained by the companies, according to two experts. While automated moderation may be an alternative solution in terms of scalability, four experts noted that the current stage of automated moderation tools lacks the necessary nuance and detection accuracy to be deemed effective.

The second factor is **Response-time issues**. Experts are worried about the response time of some tools (N=2), where the gestures might not be quick enough or where technical issues arise with the protective bubble:

> *"Gestures make it a bit more age-friendly. I just think it's too visible and it's not prompt enough. If you're in a really unsafe situation, someone's really close to you and you put that gesture up to stop them from happening that person's already too close. It needs to be much faster." (P13 - Policy and Public Affairs Officer)*

The third factor is **Abuse and Unwanted Consequences**. Safety features may not remove the problem (N=1). In contrast, they could potentially have the opposite effect, as users may abuse these tools to inflict harm on other players (N=5), or it could backfire, putting the onus on a target (N=6). Examples include gestures, "vote-to-kick" and "trust rank" tools, or situations where a child seeks assistance from a human moderator:

> *"It could backfire [...] like when a child talks to a teacher or an adult about something, and they say they were a crybaby, they went to speak to someone." (P10 - Head of online safety research)*

> *"Vote to kick worries me a little bit because I wonder if that could be used as a method of harassment. What happens if a group has voted to kick out kids who are vulnerable." (P15 - Lecturer, Consultant and Adolescent Psychiatrist)*

The fourth factor is **Circumvention and Unawareness of Consequences**. Experts (N=3) highlighted the concern of children not fully grasping the outcomes following the use of certain tools, such as reporting, where the response to the player reacting and reporting is not always clear. The possible ways of circumventing should also be considered when developing a tool; for example, two experts (P13, P16) raised the concern of a player recreating an account after being blocked.

The fifth factor is **Privacy**. Experts (N=4) were concerned about privacy regarding human moderators actively listening or with automated voice moderation. VR can offer new forms of evidence that may not be only verbal compared to 2D social media or gaming, which also raises privacy concerns when there is the possibility of recording behaviours and gestures:

> *"There are privacy concerns with just recording everything that's happened in the whole space. Also, they don't have the operational capacity to just sit and watch hours of footage." (P2 - Researcher in social VR harassment)*

Finding 5: Experts highlighted critical weaknesses in current safety features for children due to: 1) accuracy detection issues with challenges of scale and resources [a], 2) response-time, 3) the risks of tools having the abuse and unwanted consequences, 4) circumvention and unawareness of consequences and 5) privacy.

[a]Some experts lacked direct experience with social VR. The finding should therefore be interpreted with caution when considering their broader generalisability.

5.2. Expert Propositions for Enhancing Child Safety in Social VR (RQ2)

Overall, experts proposed reactive technical solutions, preventive and educative solutions and human-involved solutions.

5.2.1. One-to-One Harassment Towards a Child

Experts proposed solutions (see Table 3) for scenario 1 (see Table 4.3).

For this scenario, experts (N=4) proposed to add **Age-Verification and Age Assurance** and to allow users to interact or to be grouped with users in the same age range (N=4).

> *"It should be suspicious if you have an older adult interacting so much with a younger account, and I'm sure that can be monitored." (P17)*

Experts (N=2) proposed buttons to help the child with a **Panic/Pause button** or with a **"I don't feel comfortable" button** to take action instantly and pause the experience. An expert also suggested educating children about social VR harms through real-world comparisons and metaphors. If safety features are shown at the setup, it might also already be a preventive method to dissuade potential offenders (N=2):

> *"If you set that up in a way where even in small print it says to any game participant "we have the parental invisible joining button". That in itself might police behaviour because there might even be an option where the parent can record and then that could be sent off." (P12 - Professor of Child and Adolescent Psychiatry)*

Experts (N=2) propose human-involved solutions, setting in-game friends that would help the child or the possibility of bringing in the parent as an invisible avatar when the child needs.

An expert also noted the need to focus on empowering the child, giving them the choice on what to do next.

5.2.2. Group Harassment Towards a Child

Experts proposed solutions (see Table 4) for scenario 2 (see Section 4.3).

Experts (N=3) proposed quick, instant and discreet actions as the child might freeze in a group harassment. Solutions include a panic button to pause the experience, an SOS button that calls a human moderator discreetly or pre-set suggested prompts to facilitate and ease the burden of conversation.

> *"In this scenario, if there was an option for Jane to instead of exiting the game, flag this situation to one of the moderators to intervene without pointing at users and blocking them individually, [...] an easy gesture that's very discreet, to call for help, bring in the moderator from a different space to then resolve the situation, like an SOS button." (P5 - Child Safety Intelligence Consultant)*

Other experts (N=3) also recommended solutions that involve pausing the experience and reflecting with a human, such as a safety advisor, to help the child. Two experts suggested combining the human moderator with AI or having a virtual online safety advisor.

Table 3: Solutions Proposed by Experts for Scenario 1 (Section 4.3): One-to-one harassment towards a child. (Half the participants had to propose a solution for scenario 1, while the other half had to propose a solution for scenario 2.)

Solutions Proposed for Scenario 1: One-to-one harassment	**Proposed by Experts**							
	P3	**P4**	**P6**	**P12**	**P13**	**P15**	**P16**	**P17**
Age-Verification and Assurance	✓	✓			✓			✓
Age-Matching and Grouping	✓	✓			✓			✓
Buttons to Pause the Experience				✓			✓	
Panic Button				✓			✓	
'I don't feel comfortable button'				✓				
Prevention and Education			✓			✓		
Automated Moderation Flagging and History Checking		✓						✓
Friends Assigned or Called (Bystanders, Friends or Parents) to Help Child	✓			✓				

"Human moderation combined with AI voice moderation that can detect if the terms they are using are harassment but also building knowledge databases." (P7 - Child Safety Intelligence Consultant)

Experts (N=3) proposed the use of biometrics to monitor the age, if the child is in distress and develop a system that learns from previous iterations and interactions of the users.

"Biometrics are all the data that the headset is giving you in order to interact with the environment. Monitor biometrics because if someone is claiming that they are 13 years old when in reality they are 47 that it's something that can be obtained thanks to biometrics."(P7 - Child Safety Intelligence Consultant)

Moreover, preventive and educative solutions were recommended (N=2) to ensure users understand the impact of harassment and know about the rules via immersive education or guidelines and tips sent via emails and ensuring the child understands when and how to take action.

"People could get emails with tips reminding them of the terms of service or perhaps the community guidelines about harassment, etc." (P10 - Head of online safety research)

Age-matching was also suggested (N=2). One expert proposed, in addition to the existing personal space bubble, the possibility of adding privacy settings to your safety bubbles, where only your friends will appear visible.

Finding 6: Overall, experts recommended and proposed safety features designed to be: 1. easy-to-use under pressure (e.g., Panic/SOS Button, Pre-set prompts) 2. discreet in use and in the outcome (e.g., Discreet gestures), 3. offering guidance (e.g., prevention and education, friends, human-involved), 4. providing immediate support (e.g., automated moderation), 5. tailored for VR physical, verbal and environmental harassment types (e.g., safety bubble based on privacy settings, biometrics), 6. enabling reflection and contemplation (e.g., pause button), and 7. delivering a sense of space, comfort, and empowerment (e.g., positive reinforcement).

5.3. Tailoring Child-Safety: Experts Advocate Personalised Social VR Tools for Children (RQ1 and RQ2)

While children are already a vulnerable group, experts (N=8) mentioned cases of neurodivergent children or children with autism who may need more help in tackling a harmful event and may not have the capabilities to take actions or to pick up appropriate social cues. Tools should be intuitive and familiar, and should be explained to the users according to two experts. They should also consider variations in social skills and resilience (N=3), accommodating shy children who may find it challenging to assert themselves, as well as addressing the needs of children with neurodivergence.

> *"It's really difficult because no two people with autism are the same, and I think often parents are the best judge of where the child's vulnerabilities lie. [...] The parent would need to negotiate with the child and help the child understand." (P12 - Professor of Child and Adolescent Psychiatry)*

> *"Some neurodivergent children might be more vulnerable to not picking up inappropriate social cues and people trying to harm them because they want to be friends, they want to fit in and that's what they often struggle to do in the real world so it might be a reason why they're here, and therefore they're vulnerable to using those tools." (P15 - Lecturer, Consultant and Adolescent Psychiatrist)*

Experts (N=3) suggested adapting safety features to the number of people co-present, with, for example, more immediate moderation tools and systems in the case of group harassment. It can be adapted depending on the configuration, i.e., a child harassed by a group of adults or a child harassed by children with easier language or more severe consequences for adults. Experts (N=2) suggested restricting the ability to communicate between unknown users via privacy settings, to only hear your friends, for example, and recommended the involvement of parents and guardianship, especially in cases of group harassment.

> *"I think when it comes to group harassment, I'm not entirely sure we've seen all of the variants of this just yet, where it might be adults against children or mobs of children against children [...] It will be different for each child, and that's what I'm noticing." (P16 - Digital Investigator Social VR threat researcher, Threat Intel analyst at social media companies)*

The proposed tools should be tailored to children's age (N=9), with a focus on increased restrictiveness and enhanced comfort for younger age groups.

> *"The wording should be one that introduces features when the user signs up. If they're 13, the safety feature should be explained using one language, but if they're 10, the language should be different." (P5 - Child Safety Intelligence Consultant)*

Additionally, the tools should be sensitive to the vulnerabilities present in various communities, including those related to gender-based concerns (N=1). Experts also noted that tools may need to be adapted according to diverse situations and severity of harm (N=3), personality traits (N=2), languages and countries (N=3), to ensure inclusivity and effectiveness.

> *"You could have some form of surveying on setup with what kind of things you enjoy, and there could be a logic that recommends your settings based on both age and personality. It's quite an interesting thing to dig into, but not something I've seen anyone do." (P14 - Online Safety/Child Online Safety Industry)*

"I think it's always good to put as much control as possible into the hands of the users." (P2 - Researcher in social VR harassment)

"In therapy practice, explaining something to a 50-year-old as I would a five-year-old really helps them grasp it. I think if we take a child-centred approach, it's applicable to everybody in that space." (P16 - Digital Investigator, Social VR threat researcher, Threat Intel analyst at social media companies)

Finding 7: Safety features should be adapted based on the child's age, social skills, neurodivergence or other disabilities, inclusivity, the number of individuals harassing and the severity and frequency of harm. This involves configuring pre-settings, parental involvement, granting control to older users, imposing more restrictions on younger users, and adjusting language settings to ensure clarity and facilitate translations.

5.4. Reflection of Solutions for Children in Social VR (RQ2)

5.4.1. Weaknesses and Concerns of Proposed Solutions

Experts noted limitations to the solutions proposed. For age verification and age grouping, as well as collecting biometrics, experts were concerned with the accuracy (N=2) and privacy issues (N=3). They also noted potential users bypassing the systems (N=2), as children want to be inclusive of older content or malicious adults want to appear as child users. Moreover, age-grouping might lead to children not learning how to interact with adults and handle potential issues (N=1).

"Now to have stronger Age-Assurance measures, you can come into questions about GDPR and how do you provide proof of identity. Is a platform allowed to hold this information? There are some challenges with that, but there's also no unique way to be able to identify whether someone is a child user or not." (P13 - Policy and Public Affairs Officer)

"Younger people don't want to be restricted to a younger content, they want to be inclusive so that could make users not want to do it, they would change their details to be over 18." (P4 - Child Online Safety Industry)

Solutions would need to be individualised based on the child's needs, but also generalised across countries, languages and platforms, which is a challenge (N=4). When solutions involve human moderators, friends or parents, experts noted the need to coordinate play time (N=2).

Overall, there is a tension between costs (N=1), safety and privacy (N=2), and avoiding putting off users from using the platforms (N=1) and children with disabilities wanting to make friends (N=1). As bad actors evolve with the policy and safety features or hurdles to overcome, solutions need to be constantly updated (N=3).

"For the automated elements we know how tech savvy offenders are, but we also know that oftentimes it's not as overt as the harmful behaviours. There could be codes, on the metaverse there's lots of anime language that's used to indicate something very different, like sexual themes and if you don't know the context, the moderation measures need to constantly be updated."

Finding 8: Safety features should aim to be accurate and constantly updated while considering privacy issues, children's engagement and potential ways of bypassing the tools. They should also aim to be generalised globally, taking into account different children's social skills and languages. While human-involved-based solutions are crucial, coordinating time play is a challenge.

5.4.2. Through the lens of the victim

Most solutions proposed are victim-based solutions where the victim has to take an action. Experts emphasised the need to teach the victim to self-moderate (N=3) by educating them on the safety features, teaching them to ask for help and remove themselves from a dangerous situation (N=2).

> *"(Solution safety advisor) For the victim, it would be an opportunity to learn some specific skills related to online safety and, through discussing with the virtual online safety advisor, what they could do differently." (P10 - Head of online safety research)*

> *"(Solution of prompts) It could be difficult for a child to initiate a confrontation with a person they know. If there is a list of pre-set phrases, that can be a way to help the child, especially if they're autistic." (P9 - Online Safety Consultant, Clinical Psychologist)*

Experts argued the tools should aim to support the victim quickly to feel more confident and relaxed (N=3). Children can lack maturity and self-awareness (N=4). They may be frightened of consequences if they tell someone (N=1), and they may freeze in a situation, so they need easy flag buttons or prompts to help them initiate a conversation to make the offender stop (N=2). Bystanders or friends assigned can help the children, as they may not be mature enough to take action (N=1).

> *"(Age-matching solution) I think it helps the victim in terms of empowerment they can relax more in in a situation where they know that adults are not in there." (P3 - Professor of Psychology, Security & Trust, Criminology)*

> *"(Solution with friends) With the buddy, it's kind of reassurance. [...] If two or three people go "yeah you're right, this isn't right" it really gives confidence that what you think is happening is actually happening." (P3 - Professor of Psychology, Security & Trust, Criminology)*

Finding 9: In a situation of harassment, the victim's control is usually taken away. The solutions (safety features and education) should consider empowering the child again, giving back control while allowing the victim to be discreet. Considerations through the lens of the victim include: quick action and tools to 1. Help the child to self-moderate 2. Help the child who lacks awareness or does not know what to do under pressure.

5.4.3. Through the lens of the harasser

While safety features allow to take direct actions, it does not allow offenders to reflect on their actions, and it does not give a chance to explain or apologise (N=4).

"(Human moderator pulled in bubble when nuclear button hit) The moderator can maybe help mediate and explain to these groups. [...] I want to allow ways for people to easily make amends, take accountability for their actions, apologise and be given the opportunity to step up and be a good community member." (P2 - Researcher in social VR harassment)

"(New version of protective bubble pushing back offenders) You are pushed back a meter that may give the harasser a moment to consider that their actions are causing the person distress." (P16 - Digital Investigator Social VR threat researcher, Threat Intel analyst at social media companies)

Experts made a distinction between intentional (N=2) and unintentional harassment (N=8). Not all harassment is intentional and may be due to a lack of self-awareness (N=8), leading to unwanted consequences (N=2).

"My concern is that a lot of these kids don't have a clue what's going on until it's way too late, and well past the point of damage being affected, and they sometimes don't know how to get out of it, that would be my experience from a professional point of view. Once kids are locked into that abusive cycle, it becomes quite hard for them to get back out of it. Harassers or abusers rely on the victim not being able to advocate for themselves, or they make it such that the victim cannot advocate for themselves." (P15 - Lecturer, Consultant and Adolescent Psychiatrist)

"In this case, we should really be careful to define the harasser, because even let's say an autistic person or a neurodivergent person is a person who sometimes can be doing things that are not considered normal, being very direct, for instance. So I think sometimes also the harassers should be protected in a way." (P9 - Online Safety Consultant, Clinical Psychologist)

Harassment may also be intentional and offenders may circumvent safety actions and consequences (N=8). Solutions should consider making offenders aware of appropriate social norms, the seriousness of their actions and consequences (N=2).

While the core reasons of harassment might also need investigation and questioning of the design of the social VR environment (N=4) that currently allows mixing adults and children (N=3), anonymity (N=1) and group dynamics that may bring the worst or inner child out of people (N=1), safety features should detect the harassment proactively regardless of the reasons (N=1).

Finding 10: Proposed solutions should prompt harassers to reflect on their actions. Considerations should include distinguishing between intentional and unintentional harassment. The solutions must detect and raise awareness of actions and their consequences, irrespective of the reasons behind the harassment.

6. Discussion

Based on our findings, we provide considerations and practical guidelines for designers to develop child-friendly safety features. These guidelines were derived based on the findings where safety features could be enhanced to better cater to children's needs in case of social VR harassment situations. We position these recommendations as forward-looking guidance to inform future design and research under real-world constraints (e.g., platform resources and scalability).

6.1. Design Guidelines Derived from Findings

Derived from the findings, we propose design guidelines grouped into four main categories: 1) Preventive and Restorative Harassment Interventions, 2) Child-friendly Easy-to-use Safety Features, 3) Child Comfort and Privacy and 4) Human Oversight and Involvement. Summary tables with the guidelines and corresponding findings can be found in Table 5.

6.1.1. Preventive and Restorative Harassment Interventions

As experts highlighted, harms in online and social VR environments may be either intentional or unintentional. Although experts distinguish between these forms of harassment, platforms may find it difficult to make this determination in real time when harassment occurs. Consequently, the proposed guidelines are designed to address both forms, before an incident happens, intending to avoid potential harassment, but also after the fact to support lasting behaviour changes, adopting a preventive and restorative approach.

Guideline 1 - Preventive Solutions*: enforce rules and community guidelines, consequences and stakeholders' responsibility.* As experts noted in the findings, intentional harassment can be facilitated in social VR due to anonymity, the lack of established social norms, and the mix of adults and children, as well as associated peer pressure. Users may differentiate the online world from the real world as a place of freedom without rules, potentially bringing out the worst in people. Therefore, experts emphasised the need to diminish that difference between the offline world and social VR interactions by integrating preventive solutions in addition to reactive solutions, establishing similar social norms and expectations around safety.

[Based on Findings 4, 10] While most safety features in social VR rely on the child to enact reactive protection measures, preventive and educational solutions are crucial. Research would be required to better understand the root causes of harassment in social VR environments, i.e., whether harassers themselves were subject to abuse, if they have insecurities or other reasons for their behaviour. This would enable designers and developers to create safer social VR experiences for children and adults, potentially reducing harassment. The responsibility also lies with the industry in establishing standards and social norms in social VR, as well as supporting meaningful parental involvement.

As preventive solutions, tips and in-game training/tutorials to use the safety features and abide by community guidelines could be shared via, e.g., in-game nudges, to dissuade users from harassing and set appropriate expectations for behaviour within these communities. Such efforts could, for example, draw from Instagram's approach to protecting users [60] via warnings to discourage harassment or prevent unwanted interactions that would automatically hide users that are not in the child's list of friends. In light of the recent OSA in the UK [61] and age verification requirement, age-matching would also help reduce unwanted interactions between minors and adults, but would require robust age-verification, which may use biometrics (itself a notable challenge and open to age ambiguity). Even if effectively applied, the latter may lead to privacy issues and identity leakage, whilst age-grouping may lead to the children not learning how to interact with adults or children of other ages. Challenges are faced by the industry in balancing engagement, costs, and safety. Therefore, there needs to be regulations to enforce safety (e.g., OSA [61]) and a collaboration between multiple stakeholders, including the platforms, user communities, and families, as shown in prior work investigating approaches to address online harm [62].

Guideline 2 - Restorative Solutions*: encourage harasser reflection, discuss with moderators, opportunities to explain and apologise and positive reinforcement.* Unintentional harassment was described by experts in the context of a neurodivergent child not understanding social cues, referring to situations where the child inadvertently engages in behaviours that may make others feel uncomfortable, offended, or harassed. Neurodivergent individuals, such as those with autism spectrum disorder, may face challenges in interpreting and responding to social signals in a typical manner. For example, a neurodivergent child might not recognise personal boundaries or understand non-verbal cues, leading to unintentional actions that others perceive as intrusive or inappropriate. Consequently, the child on the receiving end of the interaction might resort to blocking or reporting. In such cases, the neurodivergent child may miss the opportunity to learn, apologise, or adjust their behaviours.

[Based on Findings 2, 6, 10] Therefore, experts recommended restorative strategies that, instead of focusing solely on sanctioning harassers, foster a positive environment by rewarding respectful behaviour and offering opportunities for change.

Harassers should be allowed to reflect, learn from their actions and apologise, especially in mild cases. This could be done via a human moderator in a safe space. Social skills training in VR could also be put in place to help children or even adult users understand the impact of their actions on others and develop empathy [63]. A prior study investigated ways of designing social media platforms for people with intellectual disabilities [64]. One of the features that was included is education strategies for cyberbullying awareness, integrating structured teaching and video modelling into platforms [64].

Users could also be allowed to change their behaviour and receive positive reinforcement in the form of rewards. This approach could not only set a standard for desirable conduct but also motivate users to adopt positive behaviours and social cues. This follows Skinner's Reinforcement Theory [65, 66] whereby behaviour is strengthened by rewarding. Other forms of positive reinforcement, such as the Social Learning Theory [67] that focuses on imitation via attention, retention, reproduction and motivation, may be further researched as a form of improving behaviour in social VR.

Recently, according to OSA in the UK [61], companies are mandated to have the opportunity to appeal before content undergoes removal or moderation. This provision enables harassers to offer explanations and engage in reflection.

While these restorative approaches, such as reflection and discussion with safety advisors, were strongly supported by experts, they would require companies to invest in such measures, with varying scalability concerns.

6.1.2. Child-Friendly Easy-to-use Safety Tools

The guidelines below aim to safeguard children while also giving them a sense of agency and control, directly reflecting the design tension between enabling free social interaction and ensuring safety. Experts have emphasised the importance of empowering children by providing safety tools that are easy for them to understand (with real-world metaphors, for example) and enact independently, allowing protection to be maintained without unnecessarily restricting children's ability to engage in the social VR worlds.

Guideline 3 - Implement quick, easy, familiar and visible-grabbing-attention tools to be used when under pressure*. [Based on Findings 1, 3, 6, 8, 9]* The safety features should be quick, easy, familiar (e.g., derived from existing social media tools or real-world metaphors) and available and able to be used by a child when under pressure. This could be done via a visible

Panic Button that the child could quickly identify and use without having to go through multiple menus. Pre-prompts may also be available, such as "I don't feel comfortable" to help the child initiate confrontation or alert a moderator.

Social VR platforms are constantly implementing new tools, such as Meta, adding a 'Pause' in Horizon Worlds social VR platform (June 2023) that lets the user take a break from a situation or space [68].

Guideline 4 - Customisation of safety features based on child's individual needs. Safety features should allow child users to control their own boundaries based on their needs and preferences.

[Based on Findings 7, 8] At the setup, a questionnaire may be filled in by the child and/or the parent, which could adapt settings based on the responses. The questionnaire could consider age, emotional maturity, personality traits, or needs, for example. The language, the visibility and accessibility of the safety tools, as well as the UX/UI interface, may also need to be adapted depending on those preferences.

Recently, Rec Room has offered Junior Accounts that users can set up initially and alternatively (October 2023), if their moderation system detects users under the age of 12, their account is automatically converted to a Junior Account (no information is given on how these detection systems work) [69]. Junior accounts limit communication features such as voice audio, in-game text messages, as well as attending events or joining clubs [69].

Guideline 5 - Tools taking into account the number of harassers. Exploring the social VR environment in a harassment scenario involving a child reveals various entities: the number of offenders (group VS one-to-one harassment), demographic factors (e.g., age or gender). Safety features should allow users to control the child user's surroundings.

Scenario: In a virtual world like VRChat, a child enters a public space and is quickly surrounded by a small group of users who begin shouting slurs and making inappropriate comments.

[Based on Findings 3, 7] Existing safety features should enable the capacity to take simultaneous actions on multiple users at once, quickly. If the design of the tool is geared towards "one-user-at-a-time" moderation, the delays experienced will expose children to harm further. The implementation of automated moderation can also contribute to addressing situations efficiently, considering both scalability and the potential volume of cases.

6.1.3. Child Comfort and Privacy

The guideline below aims to safeguard children while also giving them a sense of agency, addressing the design tension between protecting children and preserving their freedom and comfort in social VR. Experts have emphasised the need to empower children by ensuring their comfort and privacy, so that safety measures do not become intrusive, embarrassing or create unwanted consequences that could inhibit children's willingness to participate in social spaces.

Guideline 6 - Tackling unwanted consequences: discreet ways of using safety features and enacting outcomes. Unwanted consequences for using a safety feature by a child would need to be considered when developing safety features. This could arise when a child uses a safety feature visible to the other users, such as the gestures introduced in Rec Room, when asking a human moderator for help in front of others or when appearing as blocked or reported. Design and implementation considerations should ensure that safety features do not unintentionally lead to visibly exposing the victims' actions in a way that might lead to further victimisation. In some

instances, the child might simply be shy and prefer not to draw attention to themselves, resulting in them choosing not to use safety features

[Based on Findings 3, 4, 5, 6, 9] Discreet solutions should be developed, both in the way the child user would take action, and also in the resulting outcome. While social VR is synchronous, what the other users see or perceive around the use of safety features could be invisible or be portrayed differently between victims and harassers. For instance, employing a gesture-based blocking feature could remain effective without revealing the hand movements of the user facing harassment to the other users, and consequently exposing the use of this feature.

6.1.4. Human Oversight and Involvement

Guideline 7 - Involving human moderators, friends and parents*. [Based on Findings 2, 5]* Consistent with prior work [45], experts emphasised the importance of human involvement in moderation. At the same time, echoing a study showing that 24% of incidents observed in social VR were addressed by human moderators [14], they acknowledged that fully human, round-the-clock moderation is impractical at scale, leading to valuing automated moderation. Experts, therefore, framed human efforts as complementary to automated systems. For example, they recommended involving supportive friends, parents and bystanders to help children navigate harassment situations in social VR.

Further research should consider the role of bystanders and friends, for example, finding ways the platform can encourage them to become active upstanders or setting friends as "buddies" in case of a disruptive event, as experts suggested.

Table 5: Guidelines Derived from Findings

Guideline	**Corresponding Findings**
Guideline 1: Preventive Solutions: enforce rules and community guidelines, consequences and stakeholders' responsibility	**Finding 4:** Stresses the need for consistent safety measures and establishing social norms to prevent intentional harm. **Finding 10**: Highlights the importance of consequences for intentional harassment, raising awareness of actions.
Guideline 2: Restorative Solutions: reflect, discuss with safety advisors, opportunities to explain and apologise and positive reinforcement	**Finding 2**: Human moderators can guide children through reflective processes to handle unintentional harassment. **Finding 6**: Recommends offering guidance, helping users reflect on unintended consequences. **Finding 10**: Emphasises reflection and raising awareness about unintentional actions.
Guideline 3: Implementing quick, easy, familiar and visible-grabbing-attention tools to be used when under pressure	**Finding 1**: Emphasises familiar safety features using real-world or social media concepts for easier use by children under pressure. **Finding 3**: Calls for quick tools like muting/blocking multiple users in real-time, important during group harassment. **Finding 6**: Recommends user-friendly tools for immediate support, essential under pressure. **Finding 8**: Stresses the need for accurate, updated tools that are reliable in urgent situations.

Continued on next page

Guideline	Corresponding Findings
	Finding 9: Supports the use of tools to quickly empower the victim during a crisis.
Guideline 4: Customisation of safety features based on the child's individual needs (age, neurodivergence, social skills, etc.)	**Finding 7**: Recommends adapting safety features based on the child's age, social skills, and vulnerabilities, ensuring appropriate levels of restriction and control. **Finding 8**: Suggests tools should be generalised but adaptable to diverse needs, including language and skills, ensuring inclusivity for all children.
Guideline 5: Tools taking into account the number of offenders.	**Finding 3**: Highlights the need to block or mute multiple offenders at once, key for group harassment. **Finding 7**: Recommends adapting tools to consider the number of harassers and varying harm severity.
Guideline 6: Tackling unwanted consequences: discreet ways of using safety features and enacting outcomes	**Finding 3**: Supports discreet tools to control space and mute/block harassers, minimising escalation. **Finding 4**: Suggests discreet tools to prevent further escalation during interactions. **Finding 5**: Warns about the risk of safety tools being abused, making discreet use critical. **Finding 6**: Highlights the need for discreet, effective features to avoid public attention during intervention. **Finding 9**: Focuses on the victim's need for quick, discreet actions to regain control.
Guideline 7: Involve human moderators, friends, and parents.	**Finding 2**: Human moderators provide flexibility and authority, aiding in real-time interventions. **Finding 5**: Suggests human moderation is crucial to handle accuracy and privacy issues.

6.2. Specificity of Guidelines to VR versus Non-VR Social Media

While guidelines 1, 2 and 4 provided may overlap with existing proposed design recommendations for online safety in social media [70, 71], the presentation of tools needs to be tailored to the unique characteristics of VR, which leverage embodiment, immersion, and continuous real-time interactions.

Guidelines 3, 5 and 6 are specific to VR because child users are fully immersed, engaging multiple senses—touch, sight, and sound. Unlike 2D social platforms, where delayed responses are possible, or 2D games, where the impact is less immersive and easier to distinguish from reality [37, 38], VR requires immediate reactions, especially for children under pressure who would need to enact the tools as the incident happens. Thus, the recommendations for tools required are also to be discreet as other users can see actions in real time, which could lead to unintended consequences or backlash (guideline 6), quick, familiar, and easy to use in case of an overwhelming situation for the child (Guideline 3), considering users present in the virtual room either participating in or witnessing the situation (Guidelines 5, 7).

6.3. Comparing Expert Recommendations and Proposals with Prior Work and Emerging Social VR Safety Solutions

Our findings align with points from prior work, such as the challenges presented by safety feature misuse and confusing tools in social VR [18]. The authors highlighted that confusing design elements can lead to safety risks and unintended consequences. Our findings similarly advocate for familiar and real-world child-aware concepts for the design of safety features. Moreover, experts noted the potential of "vote to kick" and "trust rank" to be abused or misused, especially in group harassment scenarios. Another study that investigated trust systems in social VR also suggested that such safety features can create hierarchies or stereotypes that can foster discrimination [72].

Prior work has called for quick-to-activate safety measures, such as gesture-based features (e.g., covering ears or pushing away to trigger protective actions) [18]. While experts' recommendations echo this need by advocating for

tools that are quick and easy to use under pressure (Guideline 3), experts also recommend making safety features discreet to prevent escalation (Guideline 6). The latter is a novel consideration not prominently addressed in previous literature.

Tools later introduced by platforms, such as Horizon Worlds' wristband "Pause" feature [21] and VRChat's "Hiding All Avatars" option [49], resonate with our findings about empowering users to regain control, especially in group harassment scenarios. Advancements in automated voice moderation from Rec Room, improving in accuracy [73], also align with experts' emphasis on continuous updates and accuracy in safety features.

Recent work [74] explored blocking strategies with 40 children and showed, as experts mentioned, that tools should enable handling the challenges of large-scale group harassment while maintaining a balance between engagement and protection. Researchers also investigated the child user's *intervention perspective (IP)*, positioning children above the harassment situation. They found that children felt significantly safer when "elevated" compared to when they were directly intervening at the ground level.

Both our findings and previous work [45] emphasise the need for human involvement in moderation to create safer online spaces. Authors highlight the importance of user-human-AI collaboration [45]. This aligns with Guideline 7, which advocates for the involvement of human moderators, but also friends and parents to complement automated solutions.

6.4. Future Work

Child Perspectives. Further research remains necessary to assess and evaluate empirically the effectiveness of social VR safety features with child users who may encounter disruptions or harassment in social VR.

From Expert-Derived Guidelines to Practice: The Need for Further Research into the Design of Safety Features. We have proposed promising new safety features and key design considerations that should be taken into account in their development. However, further research is required to turn these expert-derived guidelines into real implementations that could be evaluated by users in practice.

The Need for New Methodologies for Ethically Evaluating Child-Oriented Safety Feature Effectiveness. There is a need to ethically evaluate the proposed safety features with child users. While assessing these features in controlled environments is feasible, there is a limitation in evaluating their effectiveness when children are under pressure or facing situations of harassment.

The Need for Advocacy to Encourage Adoption. After the validation of the novel safety features, the challenge lies in promoting and facilitating their adoption to maintain consistency across all social VR platforms. Given the decentralised and heterogeneous nature of the social VR platforms and VR headset brands, achieving uniformity and consistency in safety protocols necessitates concerted advocacy efforts directed towards platform stakeholders, industry leaders, and end-users.

Maintaining Freedom while Maximising Safety. While experts have recommended safety features designed to empower and protect children rather than fully restrict their access to social VR platforms or certain features of the platforms, future research could explore how to maximise safety without removing freedom and children's sense of agency.

Safety features considering Adults VS Children. Future work may explore how safety features are perceived differently by children compared to adult offenders. Additionally, although some issues identified in this study may be more general or applicable to other vulnerable groups, the findings were derived from expert discussions explicitly framed around children. Distinguishing which aspects are unique to children and which generalise to adults remains a direction for future research.

Implications of Developmental Stages. While this paper refers to children under the age of 16, the scenarios specifically feature a 10- and a 12-year-old as illustrative examples. However, developmental differences across this age range may influence needs and responses, and should be explored in future work.

7. Conclusion

Drawing from interviews with 16 experts in child online safety and psychology, our study informs the factors contributing to the effectiveness and weaknesses of ten existing safety features from two of the most popular social VR platforms (VRChat [22] and Rec Room [23]). Experts propose specific child-friendly solutions for two given scenarios to aid experts in considering the potential context of harassment situations towards a child and gain practical and concrete insights. We contribute seven expert-driven guidelines for the design of real-time safety features to effectively safeguard children in social VR.

8. Selection and Participation

No children participated in this work.

9. Acknowledgements

We would like to thank the experts for their participation. The first author led the research, including interviews, analysis, and writing the paper. The second author assisted with coding and proofreading. The third and fourth authors supervised the main researcher and contributed to meetings and paper reviewing. This work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents (grant number EP/S02266X/1), by REPHRAIN: The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online under UKRI (grant number EP/V011189/1) and partly sponsored by a 2020 Meta Research Award on Responsible Innovation.

References

[1] M. P. Leggett-James, B. Laursen, The consequences of social media use across the transition into adolescence: Body image and physical activity, Journal of Early Adolescence 43 (2023) 947–964. doi:10.1177/02724316221136043.
URL https://journals.sagepub.com/doi/full/10.1177/02724316221136043

[2] D. Maloney, G. Freeman, A. Robb, It is complicated: Interacting with children in social virtual reality, Proceedings - 2020 IEEE Conference on Virtual Reality and 3D User Interfaces, VRW 2020 (2020) 343–347doi:10.1109/VRW50115.2020.00075.

[3] D. Maloney, G. Freeman, 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 (2020) 472–483doi:10.1145/3410404.3414268.
URL https://dl.acm.org/doi/10.1145/3410404.3414268

[4] J. R. Williamson, J. O'Hagan, J. A. Guerra-Gomez, J. H. Williamson, P. Cesar, D. A. Shamma, Digital proxemics: Designing social and collaborative interaction in virtual environments, in: Proceedings of the 2022 CHI conference on human factors in computing systems, 2022, pp. 1–12.

[5] M. Slater, B. Lotto, M. M. Arnold, M. V. Sanchez-Vives, How we experience immersive virtual environments: the concept of presence and its measurement, Anuario de psicología 40 (2) (2009) 193–210.

[6] A. Felnhofer, O. D. Kothgassner, M. Schmidt, A. K. Heinzle, L. Beutl, H. Hlavacs, I. Kryspin-Exner, Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios, International Journal of Human-Computer Studies 82 (2015) 48–56. doi:10.1016/J.IJHCS.2015.05.004.

[7] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, A. Mühlberger, The impact of perception and presence on emotional reactions: A review of research in virtual reality, Frontiers in Psychology 6 (2015) 111605. doi:10.3389/fpsyg.2015.00026.
URL www.frontiersin.org

[8] J. O. Bailey, J. N. Bailenson, Chapter 9 - immersive virtual reality and the developing child, in: F. C. Blumberg, P. J. Brooks (Eds.), Cognitive Development in Digital Contexts, Academic Press, San Diego, 2017, pp. 181–200. doi:https://doi.org/10.1016/B978-0-12-809481-5.00009-2.
URL https://www.sciencedirect.com/science/article/pii/B9780128094815000092

[9] D. Maloney, G. Freeman, A. Robb, Stay connected in an immersive world: Why teenagers engage in social virtual reality, Proceedings of Interaction Design and Children, IDC 2021 (2021) 69–79doi:10.1145/3459990.3460703.

[10] Generation vr - the institution of engineering and technology, last Accessed: 04-08-2023 (2022).
URL https://www.theiet.org/media/press-releases/press-releases-2022/press-releases-2022-april-june/19-april-2022-generation-vr/

[11] C. Fiani, P. Saeghe, M. McGill, 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) (2024) 54, 25 pages.
URL http://mkhamis.com/data/papers/fiani2024cscw.pdf

[12] L. Blackwell, N. Ellison, N. Elliott-Deflo, R. Schwartz, Harassment in social virtual reality: Challenges for platform governance, Proc. ACM Hum.-Comput. Interact. 3 (CSCW) (nov 2019). doi:10.1145/3359202.
URL https://doi.org/10.1145/3359202

[13] G. Freeman, S. Zamanifard, D. Maloney, D. Acena, Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality, Proc. ACM Hum.-Comput. Interact. 6 (CSCW1) (apr 2022). doi:10.1145/3512932.
URL https://doi.org/10.1145/3512932

[14] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, M. Elsherief, Challenges of moderating social virtual reality, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–20. doi:10.1145/3544548.3581329.
URL https://doi.org/10.1145/3544548.3581329

[15] D. Maloney, G. Freeman, D. Y. Wohn, "talking without a voice": Understanding non-verbal communication in social virtual reality, Proc. ACM Hum.-Comput. Interact. 4 (CSCW2) (oct 2020). doi:10.1145/3415246.
URL https://doi.org/10.1145/3415246

[16] Sex offenders using virtual reality to abuse children as victims describe feeling 'phantom touch' — science & tech news — sky news (2023).
URL https://news.sky.com/story/sex-offenders-using-virtual-reality-to-abuse-children-as-victims-describe-feeling-phantom-touch-12954906

[17] Police 'investigate sexual abuse of young girl's avatar in the metaverse' - prompting nspcc warning — science & tech news — sky news (2024).
URL https://news.sky.com/story/amp/police-investigate-sexual-abuse-of-young-girls-avatar-in-the-metaverse-prompting-nspcc-warning-13041003

[18] Q. Zheng, S. Xu, L. Wang, Y. Tang, R. C. Salvi, G. Freeman, Y. Huang, Understanding safety risks and safety design in social vr environments, Proc. ACM Hum.-Comput. Interact. 7 (CSCW1) (apr 2023). doi:10.1145/3579630.
URL https://doi.org/10.1145/3579630

[19] Vrchat safety and trust system, last Accessed: 30-08-2023 (2023).
URL https://docs.vrchat.com/docs/vrchat-safety-and-trust-system

[20] Comfort and safety — rec room, last Accessed: 30-08-2023 (2023).
URL https://recroom.com/safety

[21] Safety and privacy tools available for your teen in meta horizon worlds — meta store, last Accessed: 10-10-2024 (2024).
URL https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safety-tools-for-your-teen-horizon/

[22] Community guidelines — vrchat, last Accessed: 22-07-2022 (2022).
URL https://hello.vrchat.com/community-guidelines

[23] A parent's guide to rec room — rec room, last Accessed: 22-07-2022 (n.d.).
URL https://recroom.com/parents-guide

[24] Protection from harassment act 1997, last Accessed: 13-09-2025 (2011).
URL https://www.legislation.gov.uk/ukpga/1997/40/section/8

[25] D. Olweus, Bullying at school: What we know and what we can do, Blackwell Publishers, 1993.

[26] A. A. Volk, A. V. Dane, Z. A. Marini, What is bullying? a theoretical redefinition, Developmental Review 34 (2014) 327–343. doi:10.1016/j.dr.2014.09.001.
URL https://www.scopus.com/pages/publications/84926125579

[27] R. Slonje, P. K. Smith, Cyberbullying: Another main type of bullying?, Scandinavian Journal of Psychology 49 (2008) 147–154. doi:10.1111/J.1467-9450.2007.00611.X.
URL /doi/pdf/10.1111/j.1467-9450.2007.00611.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9450.2007.00611.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1467-9450.2007.00611.x

[28] D. Olweus, School bullying: Development and some important challenges 34 (2013) 758. doi:10.1146/annurev-clinpsy-050212-185516.
URL http://clinpsy.annualreviews.org

[29] A. E. Marwick, Morally motivated networked harassment as normative reinforcement, Social Media and Society 7 (6 2021). doi:10.1177/20563051211021378.

[30] J. M. Rabbie, H. Lodewijkx, Individual and group aggression, Current Research on Peace and Violence 10 (2/3) (1987) 91–101, pSYCHOLOGICAL ISSUES OF WAR AND PEACE.
URL https://www.jstor.org/stable/40725065

[31] K. M. LAGERSPETZ, K. BJÖRKQVIST, M. BERTS, E. KING, Group aggression among school children in three schools, Scandinavian Journal of Psychology 23 (1982) 45–52. doi:10.1111/J.1467-9450.1982.TB00412.X.
URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9450.1982.tb00412.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9450.1982.tb00412.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1467-9450.1982.tb00412.x

[32] R. Faris, S. Ennett, Adolescent aggression: The role of peer group status motives, peer aggression, and group characteristics, Social Networks 34 (2012) 371–378. doi:10.1016/J.SOCNET.2010.06.003.

[33] J. O'Hagan, F. Mathis, 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, Association for Computing Machinery, New York, NY, USA, 2023, p. 236–243. doi:10.1145/3626705.3627780.
URL https://doi.org/10.1145/3626705.3627780

[34] A. Singh, 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).

[35] E. Deldari, D. Freed, J. Poveda, Y. Yao, An Investigation of Teenager Experiences in Social Virtual Reality from Teenagers', Parents', and Bystanders' Perspectives, 2023.
URL https://www.usenix.org/conference/soups2023/presentation/deldari

[36] P. Sykownik, D. Maloney, G. Freeman, M. Masuch, Something personal from the metaverse: Goals, topics, and contextual factors of self-disclosure in commercial social vr, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1–17. doi:10.1145/3491102.3502008.
URL https://doi.org/10.1145/3491102.3502008

[37] S. R. Sharar, G. J. Carrougher, D. Nakamura, H. G. Hoffman, D. K. Blough, D. R. Patterson, Factors influencing the efficacy of virtual reality distraction analgesia during postburn physical therapy: Preliminary results from 3 ongoing studies, Archives of Physical Medicine and Rehabilitation 88 (12, Supplement 2) (2007) S43–S49, the NIDRR Burn Injury Rehabilitation Model System Program: Selected Findings. doi:https://doi.org/10.1016/j.apmr.2007.09.004.
URL https://www.sciencedirect.com/science/article/pii/S0003999307015547

[38] J. O. Bailey, J. N. Bailenson, J. Obradović, N. R. Aguiar, Virtual reality's effect on children's inhibitory control, social compliance, and sharing, Journal of Applied Developmental Psychology 64 (2019) 101052. doi:https://doi.org/10.1016/j.appdev.2019.101052.
URL https://www.sciencedirect.com/science/article/pii/S0193397318300315

[39] Code of conduct, last Accessed: 07-05-2024 (n.d.).
URL https://recroom.com/code-of-conduct

[40] Code of conduct for virtual experiences, last Accessed: 07-05-2024 (n.d.).
URL https://www.meta.com/en-gb/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/

[41] Vrchat community guidelines, last Accessed: 21-01-2024 (2024).
URL https://hello.vrchat.com/community-guidelines

[42] Playstation vr2, last Accessed: 21-01-2024 (2024).
URL https://www.playstation.com/en-us/ps-vr2/

[43] Meta quest safety center, last Accessed: 21-01-2024 (2024).
URL https://www.meta.com/gb/quest/parent-info/

[44] H. Hartikainen, N. Iivari, M. Kinnula, Should we design for control, trust or involvement? a discourses survey about children's online safety, in: Proceedings of the The 15th International Conference on Interaction Design and Children, IDC '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 367–378. doi:10.1145/2930674.2930680.
URL https://doi.org/10.1145/2930674.2930680

[45] K. Schulenberg, L. Li, G. Freeman, S. Zamanifard, N. J. McNeese, Towards leveraging ai-based moderation to address emergent harassment in social virtual reality, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–17. doi:10.1145/3544548.3581090.
URL https://doi.org/10.1145/3544548.3581090

[46] C. Fiani, R. Bretin, M. Mcgill, 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, Association for Computing Machinery, New York, NY, USA, 2023, p. 1–13. doi:10.1145/3585088.3589374.
URL https://doi.org/10.1145/3585088.3589374

[47] C. Fiani, R. Bretin, M. McGill, M. Khamis, Big buddy: A simulated embodied moderating system to mitigate children's reaction to provocative situations within social virtual reality, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), ACM, New York, NY, USA, 2023, p. 7. doi:10.1145/3544549.3585840.
URL https://doi.org/10.1145/3544549.3585840

[48] A. Bogner, B. Littig, W. Menz, Introduction: Expert interviews - an introduction to a new methodological debate, in: Interviewing Experts, Research Methods Series, Palgrave Macmillan, London, 2009, pp. 1–13. doi:10.1057/9780230244276_1.

[49] Vrchat safety and trust system, last Accessed: 10-10-2024 (2024).
URL https://docs.vrchat.com/docs/vrchat-safety-and-trust-system

[50] Most played vr games steam charts · steamdb, last Accessed: 17-01-2024 (n.d.).
URL https://steamdb.info/charts/?tagid=21978

[51] 14 best virtual reality social games 2024, last Accessed: 17-01-2024 (2023).
URL https://rigorousthemes.com/blog/best-virtual-reality-social-games/

[52] M. Jonas, S. Said, D. Yu, C. Aiello, N. Furlo, D. Zytko, Towards a taxonomy of social vr application design, CHI PLAY 2019 - Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play 19 (2019) 437–444. doi:10.1145/3341215.3356271.
URL https://dl.acm.org/doi/10.1145/3341215.3356271

[53] Gorilla tag, last Accessed: 16-09-2025 (2021).
URL https://www.gorillatagvr.com/

[54] Basic controls — rec room wiki — fandom, last Accessed: 16-09-2025.
URL https://rec-room.fandom.com/wiki/Basic_Controls

[55] Watch menu — rec room wiki — fandom, last Accessed: 29-11-2025.
URL https://rec-room.fandom.com/wiki/Watch_Menu

[56] Settings - vrchat wiki, last Accessed: 29-11-2025.
URL https://wiki.vrchat.com/wiki/Settings

[57] Vrchat community guidelines, last Accessed: 21-01-2024 (2024).
URL https://hello.vrchat.com/community-guidelines

[58] Applying for volunteer moderator – rec room help center, last Accessed: 29-11-2025.
URL https://recroom.zendesk.com/hc/en-us/articles/4419902650135-Applying-for-Volunteer-Moderator

[59] Qcamap, last Accessed 05-09-2023 (2022).
URL https://www.qcamap.org/ui/en/home

[60] A. Mosseri, Introducing new ways to protect our community from abuse — instagram blog, last Accessed: 23-01-2024 (2021).
URL https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse

[61] Online safety act 2023, last Accessed: 07-05-2024 (2023).
URL https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted

[62] S. Xiao, C. Cheshire, N. Salehi, Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm, Conference on Human Factors in Computing Systems - Proceedings 15 (4 2022). doi:10.1145/3491102.3517614.
URL https://dl.acm.org/doi/10.1145/3491102.3517614

[63] P. C. Ferreira, A. M. V. Simão, A. Paiva, C. Martinho, R. Prada, A. Ferreira, F. Santos, Exploring empathy in cyberbullying with serious games, Computers & Education 166 (2021) 104155. doi:10.1016/J.COMPEDU.2021.104155.

[64] Y.-W. A. Chang, L. Sitbon, L. Simpson, Towards a secured and safe online social media design framework for people with intellectual disability (2021). doi:10.1145/3441852.3476540.
URL https://doi.org/10.1145/3441852.3476540

[65] B. F. Skinner, The Behavior of Organisms, B. F. Skinner Foundation, 1938.
URL https://books.google.co.uk/books?hl=en&lr=&id=S9WNCwAAQBAJ&oi=fnd&pg=PT20&ots=LnwqamwBH-&sig=SGAPVzfIx4ZZIo522nyl55uOP1A&redir_esc=y#v=onepage&q&f=false

[66] M. A. Adams, Reinforcement theory and behavior analysis., Behavioral Development Bulletin 9 (2000) 3–6. doi:10.1037/H0100529.

[67] A. Bandura, R. H. Walters, Social learning theory, Vol. 1, Prentice hall Englewood Cliffs, NJ, 1977.

[68] Safety and privacy tools available for your teen in meta horizon worlds, last Accessed: 07-05-2024 (2024).
URL https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safety-tools-for-your-teen-horizon/

[69] Junior accounts & age verification, last Accessed: 07-05-2024 (n.d.).
URL https://recroom.zendesk.com/hc/en-us/articles/4426900227735-Junior-Accounts-Age-Verification

[70] L. Blackwell, J. Dimond, S. Schoenebeck, Article 24. insights from heartmob, Proc. ACM Hum.-Comput. Interact 1 (2017). doi:10.1145/3134659.
URL https://doi.org/10.1145/3134659

[71] K. Badillo-Urquiola, D. Smriti, B. McNally, E. Golub, E. Bonsignore, P. J. Wisniewski, Stranger danger! social media app features co-designed with children to keep them safe online, in: Proceedings of the 18th ACM International Conference on Interaction Design and Children, IDC '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 394–406. doi:10.1145/3311927.3323133.
URL https://doi.org/10.1145/3311927.3323133

[72] Q. Chen, J. Cai, G. Jacucci, "people are way too obsessed with rank": Trust system in social virtual reality, Computer Supported Cooperative Work (CSCW) (123). doi:10.1007/s10606-024-09498-7.
URL https://doi.org/10.1007/s10606-024-09498-7

[73] State of voice moderation — rec room developer blog, last Accessed: 10-10-2024 (2024).
URL https://blog.recroom.com/posts/voice-moderation-updates

[74] C. Fiani, R. Bretin, M. McGill, M. Khamis, "abracadabra, block the bullies!": Child perceptions of manual, semi-automated and automated interventions against group harassment in social vr, in: Proceedings of the 24th Interaction Design and Children, Association for Computing Machinery, 2025, pp. 314–339. doi:10.1145/3713043.3728850.
URL https://doi.org/10.1145/3713043.3728850

Table 4: Solutions Proposed by Experts for Scenario 2 (Section 4.3): Group harassment towards a child. (Half the participants had to propose a solution for scenario 1, while the other half had to propose a solution for scenario 2.)

Solutions Proposed for Scenario 2: Group harassment	**Proposed by Experts**							
	P2	**P5**	**P7**	**P8**	**P9**	**P10**	**P11**	**P14**
Quick, Discreet and Instant: Button and Gesture								
Panic/Pause Button (to pause the experience)	✓	✓						
SOS Button (calls a moderator)		✓						
Discreet gesture to take action	✓	✓						
Pre-Set Suggested Prompts / Nudges					✓			
Automated Moderation Flagging / Biometrics and History Checking			✓	✓		✓		
Age-Verification and Group Age-Matching			✓					
Preventive Solution: Rules, Expectations, Tips Reminding Community Guidelines				✓	✓			
Human-Involved Solutions								
Safe Space / Timeout to Consult Safety Advisor						✓	✓	
Human Moderator Combined with AI			✓			✓		
Positive Reinforcement and Safe Space							✓	
Safety Bubble based on Privacy Settings								✓

Appendix A. Experts Demographics Tables

Table A.1: Demographics of experts who participated in one-to-one interviews: parent, education, profession, profession involving children and country of profession. Experts had a Mean age of 43.3 years (σ=10.5). 10 experts identified as female, 1 as non-binary/third gender and 5 as male.

P number	Parent?	Education	Profession	Profession involving children?	Country Profession
P2	No	Doctorate (PhD, Information Science)	Researcher in social VR harassment	No	USA
P3	No	Doctorate (Psychology)	Professor of Psychology, Security & Trust, Criminology	No	UK
P4	Yes	College / University degree (Sport and business management)	Child Online Safety Industry	No	UK
P5	No	University degree (Marketing and management)	Child Safety Intelligence Consultant	Yes	UK
P6	Yes	Professional degree (Masters in Cybercrime Investigations and Forensic Computing)	Trainer in Child Online Safety	Yes	Republic of Ireland
P7	No	College / University degree (Social Psychology)	Child Safety Intelligence Consultant	Yes	Spain
P8	No	College / University degree (Clinical psychology)	Clinical Psychologist	No	Italy
P9	Yes	College / University degree Linguistics and AI (and an MBA)	Online Safety Consultant Clinical Psychologist	Yes	UK
P10	No	Doctorate (PhD)	Head of Online Safety Research	Yes	Malta / Worldwide
P11	No	Doctorate (Psy.D.)	Community Developer	No	USA
P12	Yes	Doctorate (PhD)	Professor of Child and Adolescent Psychiatry	Yes	UK
P13	No	College / University degree (MSc Comparative Social Policy)	Policy and Public Affairs Officer	Yes	UK
P14	No	College / University degree (BA Italian)	Online Safety/Child Online Safety Industry	Yes	UK
P15	Yes	Professional degree (MA, PGCE:PSE, MBChB, MRCPsych)	Lecturer, Consultant and Adolescent Psychiatrist	Yes	UK
P16	Yes	Other: MSc and PhD candidate PhD (current) and MSc in Psychotherapy	Psychotherapist and cybertrauma Consultant	Yes	UK
P17	No	Professional degree (Data & Society)	Social VR/Media threat researcher	No	UK

Table A.2: Demographics of experts who participated in one-to-one interviews: child bullying knowledge, child cyberbullying knowledge, Social VR awareness, Social VR experience, owner of VR headset, VR experience. Experts had a mean of Social VR Awareness of 4 (σ=1.5). While some experts were limited on experience of social VR (although had experience with VR), their expertise and background still offer informed insights, providing their vision of how to protect children in social VR, future work can build upon.

P Number	**Child harassment/bullying Knowledgeable**	**Child Cyberbullying Knowledgeable**	**Social VR Awareness (1-5 Likert scale)**	**Social VR Experience**	**Owner of VR headset**	**VR Experience**
P2	Extremely	Extremely	5	A lot	Yes. (Meta)	A lot
P3	Moderately	Moderately	5	A moderate amount	Yes. (Meta)	A moderate amount
P4	Very	Very	5	A little	No	A little
P5	Very	Extremely	5	A little	No	A little
P6	Very	Very	4	None at all	No	A little
P7	Very	Very	4	A moderate amount	Yes. (Meta)	A moderate amount
P8	Not at all	Not at all	1	None at all	No	A little
P9	Moderately	Very	5	A little	Yes. (Meta)	A little
P10	Moderately	Very	2	None at all	No	A little
P11	Moderately	Moderately	5	A great deal	Yes. (Meta)	A great deal
P12	Moderately	Moderately	1	None at all	No	A little
P13	Very	Moderately	5	A moderate amount	No	A moderate amount
P14	Extremely	Extremely	5	A little	No	A little
P15	Very	Very	2	A little	No	A little
P16	Extremely	Extremely	5	A lot	Yes. (Meta)	A lot
P17	Extremely	Moderately	5	A great deal	Yes. (Meta)	A lot