"*Pikachu would electrocute people who are misbehaving*": Expert, Guardian and Child Perspectives on Automated Embodied Moderators for Safeguarding Children in Social Virtual Reality

Cristina Fiani
c.fiani.1@research.gla.ac.uk
University of Glasgow
UK

Robin Bretin
r.bretin.1@research.gla.ac.uk
University of Glasgow
UK

Shaun MacDonald
Shaun.Macdonald@glasgow.ac.uk
University of Glasgow
UK

Mohamed Khamis
Mohamed.Khamis@glasgow.ac.uk
University of Glasgow
UK

Mark McGill
Mark.McGill@glasgow.ac.uk
University of Glasgow
UK

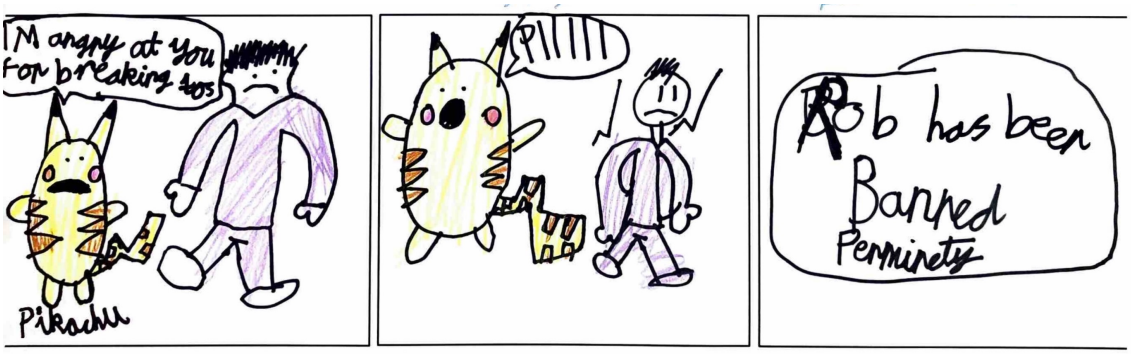


Figure 1: A child participant design of their embodied automated moderator in social VR, depicting a *Pikachu*-styled moderator banning an individual for their actions.

ABSTRACT

Automated embodied moderation has the potential to create safer spaces for children in social VR, providing a protective figure that takes action to mitigate harmful interactions. However, little is known about how such moderation should be employed in practice. Through interviews with 16 experts in online child safety and psychology, and workshops with 8 guardians and 13 children, we contribute a comprehensive overview of how Automated Embodied Moderators (AEMs) can safeguard children in social VR. We explore perceived concerns, benefits and preferences across the stakeholder groups and gather first-of-their-kind recommendations and reflections around AEM design. The results stress the need to adapt AEMs to children, whether victims or harassers, based on age and development, emphasising empowerment, psychological impact and humans/guardians-in-the-loop. Our work provokes new participatory design-led directions to consider in the development of AEMs for children in social VR taking child, guardian, and expert insights into account.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

social virtual reality, metaverse, child online safety, guardian, parent, grandparent, children, experts, design workshops, interviews

ACM Reference Format:
Cristina Fiani, Robin Bretin, Shaun MacDonald, Mohamed Khamis, and Mark McGill. 2024. "*Pikachu would electrocute people who are misbehaving*": Expert, Guardian and Child Perspectives on Automated Embodied Moderators for Safeguarding Children in Social Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3613904.3642144

1 INTRODUCTION

The use of social Virtual Reality (VR) in the metaverse, a concept that refers to a virtual world that is fully immersive and interactive [24], raises concerns regarding unsupervised interactions between children and other adult and child users worldwide [50]. As opposed to 2D social media in which users interact behind screens, in social VR, users interact via an embodied avatar synchronously in 3D immersive virtual environments, increasing the illusion of "being

there" [54] that may amplify virtual harm [18, 38]. The nature of VR being experienced through head-mounted devices, completely occluding reality and not supporting bystander awareness by default [60], makes it challenging for parents to oversee and comprehend the interactions taking place. And despite most VR devices and social VR platforms being designed for 13+ years old [3, 7, 8], young children have inevitably been drawn to these platforms due to the unique free-for-all rich social activities and games they offer [50].

Existing safety-enhancing features, such as such as blocking, personal space bubbles, muting, and reporting players [2] aim to keep users safe from nuisance users [5]. However, these measures suffer from significant limitations. First, they place the responsibility of moderation directly on users, such as children or guardians, who might not be well-equipped or familiar with the technology, or unaware of the most effective strategies [19, 43, 45, 46]. Second, they neither offer remote parental oversight nor inform parents of their children's involvement as bullies or as victims of bullying. Platforms have recruited human moderator volunteers to help mediate bad behaviour and enhance safety of public virtual rooms [18]. However, they are not able to tackle most of the incidents [63] and they cannot be present 24/7. Research shows that only 24% of incidents observed in social VR were addressed by human moderators, highlighting the need for new moderation tools [63]. Some platforms such as VRChat also offer "pay-to-trust" models, where you can pay a subscription and among other benefits become a more trusted member instantly which may be a questionable way of improving safety [6].

Recently, alternative solutions to better safeguard children in social VR have been investigated to tackle the issues described above. One promising approach is *embodied automated moderation* [33], where AI-moderators embody an interactive character that is present in the virtual space of the children, which has emerged as a potential solution that promises the advantages of automated approaches, such as complete awareness of the virtual environments and scalability [65], alongside the benefits of human moderators like relatability to the moderator, sense of safety, and improved explainability and engagement as opposed to voice or text messages [33]. There are further potential benefits to such embodied moderators. Children have displayed increased social responsiveness to VR characters, as opposed to their counterparts on 2D television [15], and research shows that individuals tend to alter their behaviour when they are aware of being observed due to the so-called "Audience Effect", resulting in actions for the betterment of others [20, 23]. However, to date there is no comprehensive understanding of the extent to which embodied automated moderators are suitable and applicable, what features should be customised to enhance child users engagement and enhance credibility, as well as intervention approaches that can be put in place by Automated Embodied Moderators (AEMs).

This paper explores the design space of embodied automated moderation in social VR through interviews with 16 experts (5+ years experience in child online/social VR safety and psychology) and 5 design workshops with 8 guardians (5 parents and 3 grandparents) and 13 children in total. By merging perspectives, we identify where there is consensus and different views around the suitability, role, presentation, interactions and actions of the automated moderator. Our findings inform the development of effective and child-centred AEMs, providing the foundations for the development of safer social VR through embodied moderation.

We investigate the following research questions from the perspectives of *experts*, *guardians* and *children*:

RQ1.1) What are the benefits and concerns associated with automated moderation for child in social VR from expert, guardian and child perspectives?

RQ1.2) What are the benefits and concerns associated with embodiment features of AEMs for children in social VR from expert and guardian perspectives?

RQ2) What are proposed intervention approaches of AEMs from expert, guardian and child perspectives?

RQ3) What are proposed embodiment features of AEMs from expert, guardian and child perspectives?

1.1 Contribution

The design space of AEMs for children in social VR has not been explored with any notable breadth, and no paper of our knowledge has holistically considered expert, guardian and child perspectives towards AEMs. The insights into the design space of AEMs uncovered by our work serve as groundwork for understanding their potential and applicability. The paper contributes insights into the design of AEMs to help children in harmful experiences in social VR that, for the first time, balance three stakeholders viewpoints: a) experts in child online/social VR safety and psychology, b) guardians and c) children. Driven from one-to-one interviews with experts and workshops with families, we synthesise their concerns, preferences and proposals for the design and use of AEMs. From these two phases, we produced design considerations towards eventual realisation of AEMs to safeguard children in social VR, around the lifecycle of a moderation incident with the emphasis on tailoring the AEM to each child, human/guardian-in-the-loop and the psychological impact on children.

1.2 Terminology

In this paper, "children" refers to minors under 16. "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. More details about the experts are in 3.1.1. Users causing social disruptions to child users are termed "harasser" or "wrongdoer", and the affected children as "victims."

2 RELATED WORK

2.1 Harassment in Social VR towards Children

Despite an age limit of 13 [3, 7, 8], social VR platforms have seen an increased uptake by children and teenagers, often sharing virtual spaces with adults [36, 50, 51]. The co-presence of adults and children, as well as unique VR affordances that mimic real face-to-face interactions while still being anonymous, have led to new forms of harm, e.g., physical and environmental [18, 38]. Drawn from 23 interviews [18], three main types of harassment have been identified in social VR: 1) *verbal* (i.e., voice or chat), 2) *physical* (e.g., avatars invading personal space or physical attack) and 3) *environmental* (e.g., displaying graphic content in a virtual space). As VR allows full embodiment and immersion leading to feeling present in the

virtual environment [54], users - especially children [14, 15] - are even more vulnerable when experiencing harm in social VR compared to social media [18]. Moreover, due to the anonymity and freedom of these new platforms, there are looser social and ethical norms that can result in increased hostile behaviours from users [37]. Through analysis of the videos and YouTube users' comments, a recent study reveals severe and unexpected safety risks in social VR, emphasizing the influence of varying understandings of community norms on reactions to these risks. [71]. This underscores the importance of understanding children's social VR experiences and perceptions and the need for suitable safety enhancing tools.

2.2 Automated Moderation in Social Media and Social VR

2.2.1 Automated Moderation in Social Media. With the increase of cyberbullying and online harassment on social media and its detrimental impact on mental health, the need for content moderation becomes stronger [41]. The immense amount of content posted everyday, however, makes the human moderation logistically challenging [40]. In response to the scale of content posted, automated moderation is used. It consists of deploying various artificial intelligence (AI) techniques to filter and process user generated content, applying pre-set rules to reject or approve the image or text posted online [68]. While AI-moderation shows some promise, it has key limitations including difficulties in parsing ambiguity, identifying false positives [68], reinforcing biases. Furthermore, there are ethical considerations such as the lack of transparency [42], visibility [59], privacy and the lack of user agency [22, 30]. A study involving 59 interviews with teenagers showed that these interventions may impact children's rights, and emphasised the need to involve children in AI design decisions for platforms [57]. The study further showed that content removal is insufficient.

2.2.2 Automated Moderation in Social VR. Existing safety tools like blocking, personal space bubbles, muting, and player reporting [2], aim to protect users from disruptive individuals [5]. However, these mechanisms place the responsibility on child users, who might lack the necessary understanding or familiarity with the technology [19, 43, 45, 46], including children and their guardians. Furthermore, they do not facilitate remote guardian supervision or inform parents about their children's negative encounters, whether as targets of bullying or as bullies themselves. While some platforms recruit volunteer human moderators to intervene in cases of misconduct [18] they cannot address the majority of incidents and their availability is limited. A recent study found that only 24% of disruptions within social VR environments were managed by human moderators [63]. Automated moderation that takes into account the unique affordances of VR has, therefore, been proposed as an alternative safety tool to combat harassment in social VR [33, 65]. While promising, research is needed to better understand the impact and implications of automated moderation on children, how it is perceived by child users, their guardians and what the design factors and actions are for an effective system.

2.3 Exploring the Role and Impact of Embodiment in Social VR for Child Safety

2.3.1 Enhanced Social Interaction. Embodiment, defined as *states of the body such as posture, arm movements, facial expressions* [17], can enhance and influence social interactions. In social VR, adult users have highlighted that engaging in shared activities in embodied and more physical ways (e.g., virtual hug) led to fostering extended relationships with other players and enhanced connectedness due to verbal and non-verbal communication modalities [37]. Research has shown the impact of embodiment in social relationships and interactions [48], in particular as one of the four social embodiment effects: "perceiving bodily states in other people actually results in bodily mimicry in oneself", the observer's embodied responses can mimic the perceived embodied stimuli (e.g., when someone smiles at you) [17].

Immersion and a feeling of presence in VR have been shown to have a significant impact on children psychologically [14, 15] and physically [55]. For example, immersive VR 3D characters have a stronger impact on decision-making than 2D TV characters [15]. Virtual embodiment and socially responsive virtual characters in VR could consequently have a notable impact on the way that children experience, and react to, social interventions [14] and AI. Embodied avatars could leverage an influential appearance and relationship: prior work shows that children's perception of such figures (e.g., Disney characters [27], Teachers VS Peers [44]) influences their behaviour, highlighting the need to carefully consider the avatar's representation as it may lead to negative (decreased agency) or positive (inspire good behaviours) impact [27, 44]. Few studies have examined social embodiment in VR among children [14–16, 33], leaving questions regarding what features are suitable, demanded and preferred for an embodied avatar safeguarding children from harassment in social VR spaces. The above research underscores the importance of embodiment in social interactions and VR, emphasising its potential significance in enhancing the effectiveness of automated moderation. This motivates the need of insights into the design and implementation of automated moderation systems that leverage the power of embodied experiences in a virtual space.

2.3.2 Embodiment of Moderator Agents to Safeguard Children in Social VR. As VR enables a myriad of avatar representations, from human-like to non-human-like, it raises questions about how this impacts user perception. Prior work found that children perceived and described the embodiment and personification similarly across three forms of embodied characters (human, animal and anthropomorphised creature) [16]. They treated these virtual bodies if they were physical bodies, regardless of their realism. It was concluded that the types of character does not affect how children perceive the realism of VR [16]. Additionally, children specifically remembered character body movements and facial expressions. An empirical study with a Wizard-of-OZ AEM prototype called "Big Buddy", showed children have various opinions on its physical and social features, falling into three preference clusters: 1) authoritarian, visible, humanised, teacher-like; 2) more friendly and indulgent; and 3) non-embodied, invisible [33].

2.4 The Importance of Diverse Perspectives in Participatory Design of Social VR Safety Tools for Children

Collaboration among various perspectives is crucial to combat social VR child harassment and develop strategies that are easy for children to grasp and use [35], while also safeguard children and reassure guardians. Balancing the participation of guardians and children in the research is thus essential [39]. Guardians can shape and influence their children's media use and habits [61]. However, they can struggle with the unfamiliarity of technology, wanting greater transparency about their children's use of technology [19]. When designing AI-based interventions, it is important to understand children's needs and solicit children's feedback into the design of policies regarding online safety [57]. Professionals can bring an impartial perspective and with their knowledge and experience to offer important insights and influence the acceptance and application of safety measures designed to shield children from harm in virtual settings. However, conflicts between these perspectives might surface, leading to difficulties in reaching a consensus [35]. Collecting stakeholder perspectives from children, guardians and professional experts can provide enriching insights on the design of effective safety tools.

2.5 Gaps in Prior Work

As described above, there is an urgent need to develop suitable safety enhancing tools for children in social VR. While there is potential in AEMs, it is unclear how suitable these are and what the key factors and actions are for an effective system that does not only remove the source of harassment but also provides support. Moreover, further research is needed on how AEMs can be presented for effective safeguarding, what appearance and customisation features (e.g., personality, communication) children may change and personalise to feel safer without removing enjoyment and credibility of the moderator. We address these gaps in the following section.

3 METHODS - COLLECTING MULTI-STAKEHOLDER PERSPECTIVES ON AUTOMATED EMBODIED MODERATORS FOR CHILD SAFETY IN SOCIAL VR

We fill the aforementioned gaps by interviewing experts (Study 1) and running participatory workshops with families (Study 2), to get insight into recommendations and preferences of how AEMs should be presented and what interventions they should take. We developed activities based on design workshops featured in prior work which were used to develop mediation tools for 2D social media [25, 47], conversational agents [39, 67] or robots to foster anti-bullying [64]. All studies were approved by our ethics committee prior to commencing and lead researcher was a member of the national Protecting Vulnerable Groups scheme.

3.1 Study 1: One-to-one Interviews with Experts

3.1.1 Participants.

Recruitment. Experts with the criteria defined in section 1.2 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. We recruited 16 experts (10 female, five male, one non-binary/third gender), with a mean age of 43.3 years (σ=10.5). There were 13 White / Caucasian, one was Black / African, one was Asian and one Other. Six of the experts were parents. 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 Republic of Ireland. Experts had a mean Social VR Awareness (level of understanding and knowledge of what social VR is on a 1-5 scale, with 5 being extremely aware) of *4* (σ=1.5). All experts had at least *a little* experience in VR (e.g., engaged in VR gaming activities, tried VR in museums, tried a family's member VR headset or work-related), seven own a VR headset and 12 had at least *a little* experience in social VR. Four experts self-reported they were *extremely knowledgeable* in child cyberbullying, six self-reported being *very knowledgeable*, six as *moderately knowledgeable* and one as *not knowledgeable*. The demographics of experts who participated in the interviews are detailed in **Appendix A**. (Expert Participant are numbered P2-P17; P1 was excluded for lack of relevant input).

3.1.2 Procedure. We first ran pilot tests with researchers in Human-Computer Interaction, which aimed to refine the questions and ensure they are clear and understandable. The one-to-one interviews were conducted with 5 researchers from the HCI group of the university. Two were also authors of the paper. They were recruited after posting an advertisement on Slack. Some questions were revised for clarity and the slides were updated adding a timer for each question ensuring they could all be answered within the time limit.

We then conducted 16 semi-structured in-depth one-to-one interviews with experts that meet the criteria defined in section 1.2, via Zoom that lasted 45-60 minutes. We presented slides with timers (varying from 2min to 4min) for each question. The interviews all started with a 5-minute presentation about social VR regardless of 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 [1]. We then listed existing safety tools and a video of "Big Buddy", a current example of what an AEM could look like and actions it could take [33]. "Big Buddy" is introduced as a Wizard-of-Oz automated agent in a simulated social VR game and intervenes when a fictional player disrupts the child's game. It has a robotic voice (text-to-speech), a robotic/astronaut suit and appears

as a tall White/Caucasian man. When he intervened, he could take one of the three actions or a combination, spoken verbally and written in a speech bubble: reset points of the wrongdoer back to 0, notify parents of the wrongdoer, and/or exclude the wrongdoer for a day [33]. The three actions were chosen based on punishments on children's rating of school punishments: 'information being sent home', 'teacher explaining what is wrong with their behaviour in front of the class' and 'being stopped from going on a school trip' as top three of the most effective punishments [56].

Interview Questions. During the interview, we asked 10 questions related to "Big Buddy", with follow-up questions when further precision or clarification was required, or where points of interest were noted. These questions explored topics including:

- First impressions regarding the suitability of AEM to safeguard children from harassment in social VR;
- AEM's strengths and weaknesses, alternative actions, alternative features (physical, social);
- If and how interventions/features can be adapted for different groups, individual VS group harassment (i.e., if a group harasses a child).

The slides with the introduction and full list of questions can be found in the Supplemental Material.

3.1.3 Analysis. We used inductive thematic analysis [21, 26] to analyse the data. The analysis was completed in six phases [21]: 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 [4]. 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. This was accomplished in two meetings to compare code-by-code. 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. We did not seek inter-rater reliability between the coders because researchers may interpret the meaning of codes differently [49] and qualitative research acknowledges the researchers' influences [26]. 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. The themes were then developed based on the categories. 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.

3.2 Study 2: Workshops with Guardians and Children

3.2.1 Participants.

Recruitment. Families were recruited via a local library in the United Kingdom, in English, on a voluntary basis between May and July 2023. Parents' or legal guardians' consent was required. Once they signed up on Eventbrite, they were sent a Qualtrics form with an information sheet, a consent form to sign and demographics questions (parents' and children's social VR/VR/online games experience, age, gender, ethnicity, education). The latter was completed by guardians only. Children were all accompanied to the workshop by their legal guardian. Once at the workshop, both parents and children had to sign a consent form to make sure both agree to participate. At the start, we assured participants that no wrong answer can be given and assured them of the confidentiality of their responses. Children were offered a certificate of attendance at the end of the workshop as a token of appreciation.

Demographics. Guardians included five parents (four female, one male) and three grandparents (two female, one male). Parents had a mean age of 45.5 (σ=8.4); two identified as Black/African and three as White/Caucasian. Grandparents had a mean age of 63.5 (σ=2.8), all identifying as White/Caucasian. Guardians had a mean social VR awareness (level of understanding and knowledge of what social VR is on a 1-5 scale, with 5 being extremely aware) of 1.9 (σ=1.4), while regarding Attitudes towards social VR (1-5 scale, with 5 being a very positive attitude), they scored a mean of 2.6 (σ=1.1). Children (seven female, six male) had a mean age of 11.5 (σ=2.7). Eleven children had at least *a little* VR experience, two had *none at all* and one was not known. The summary of demographics can be found in tables in **Appendix B**.

3.2.2 Procedure. We conducted five workshops (approximately 75 minutes each), with a total of eight guardians (five parents, three grandparents) and 13 children. Table 1 and Table 2 provide detailed breakdowns of the workshop sessions. The first workshop was conducted in a Social Community Center with three child participants (13-16 years old). The workshop was led by a researcher who was assisted by a librarian and a social worker for logistics. The other four workshops were conducted in the local Library from which families were recruited on a voluntary basis. Guardians and children (8 to 16 years old) participated. Guardians and children were seated on opposite sides of the room to minimise influencing the children's responses. Two researchers led the sessions, each focusing on either the guardians or the children. Both tables were audio recorded with written consent. The librarian helped with logistics and primarily sat at the children's table to help them feel at ease, as many of the children were familiar with her.

Introduction and VR Demonstrations. The workshops started with a similar introduction as described in 3.1.2 before featuring live VR demonstrations to introduce the concept of social VR and the concept of AEMs. We used the same simulated virtual social game experience with Big Buddy described in prior work [32] (game access provided by authors), which was showcased in Study 1 through a video. The demonstrations aimed to provide a better understanding of the topic at hand.

Guardian Brainstorming. Guardians engaged in a collaborative group brainstorming session to identify pertinent aspects to consider for an effective AEM. The questions posed to guardians were similar to those posed to the experts regarding *suitability*,

strengths and weaknesses, alternative actions, alternative features (see section 3.1.2).

Children Brainstorming. Questions for the children focused on attitudes towards, and the anticipated interactions with, AEMs:

- *the appearance they would prefer for their AEM*;
- *when* and *where* in the scene they would want their moderator to appear;
- *actions and phrases* their AEM would employ.
- In the fourth workshop, older children (13 to 16 years old) mentioned not particularly wanting an AEM that is fully automated or a character. We therefore re-tailored their questions: *Why they did not want an AEM, its strengths and weaknesses, actions/features they would recommend.*

Each stakeholder group received an A3 paper per question, along with stationary and post-its for their responses, which we collected at the end of the workshop. The overall discussion covered both the embodiment and system of the moderator. This aimed to generate ideas and perspectives from multiple viewpoints.

Design Activity: Designing an AEM via Drawing and Storyboarding. The next step involved children designing their AEM, including its name, its appearance, and outlining its actions in a particular scenario, building upon the ideas they had previously brainstormed. Children were given a sheet of A3 paper for storyboarding and one of two child harassment scenarios of their choice based on existing possible harassment scenarios [18, 50, 51] (1. Group harassment scenario: a group of teenagers harassing a child in social VR, 2. One-to-one harassment: an adult being inappropriate to a child). They could draw their AEM, give it a name and actions it would to take to tackle the scenario chosen. The slides used to introduce the tasks are in the Supplemental Material.

Workshop Breakdown	Guardians	Children
Welcome and Social VR Introduction	10-15 min	
VR Big Buddy Demonstrations	20 min	
Brainstorming Post-its	40 min	20 min
Design activity (Children only)		20 min

Table 1: Workshops Schedule Breakdown for Guardian and Child Participants.

3.2.3 Analysis. The process of analysis and theme development mirrored the one described in Section 3.1.3 with the main experimenter and an additional researcher (different from the one who analysed the interviews) leading to four separate coding schemes (two separate ones for guardians and two separate ones for children). The researchers then collaborated to consolidate the coding schemes into one combined scheme for guardians and one for children, by collating or distinguishing between codes using the compare option on QCAmap. There were limited codes from one researcher that were not reflected on the second researcher's code scheme. This informed how we combined the schemes, by incorporating the missing codes and then revisiting the transcripts with the final code scheme. We created a set of higher-level codes and themes by bringing together related codes. The main author identified the common themes across experts' and families' results. Once the Findings Section was written, all authors reviewed themes.

Workshops	Children (participants)	Guardians (participants)
Workshop 0	3 (P1, P2, P3) (group: 13-16 yo)	0
Workshop 1	3 (P1C-1, P2aC-1, P2bC-1) (group: 8-12 yo)	2 (P1P-1, P2P-1) (2 parents)
Workshop 2	4 (P1C-2, P2C-2, P3aC-2, P3bC-2) (group: 8-12 yo)	3 (P1P-2, P2G-2, P3P-2) (2 parents, 1 grandparent)
Workshop 3	1 (P1C-3) (group: 8-12 yo)	1 (P1G-3) (1 grandparent)
Workshop 4	2 (P1C-4, P2C-4) (group: 13-16 yo)	2 (P1G-4, P2P-4) (1 parent, 1 grandparent)

Table 2: Table Summarising participant groups attendance of workshops. Whilst the presence of guardians varied, there was always a responsible adult accompanying children (the librarian was present in all workshops). P(Participant Number)(P parent or G grandparent)-(Workshop Number) (e.g., P3P-2) with corresponding child as P(Participant Number letter)(C)-(Workshop Number), a letter is added if more than one child per guardian came (e.g., P3aC-2, P3bC-2).

3.3 Limitations - Study 1 and Study 2

The sample size and necessity of conducting the studies in English in the United Kingdom may limit the external validity with respect to the global population and applicability to other cultures. The design activity of the workshops captured children's initial reactions in one session, rather than refining proposed designs over multiple sessions. Consequently, the proposed AEM designs by children serve as formative, imaginative insights that may not be practical or actionable, but reveal much about child attitudes towards, and perceptions of, AEMs. Further iterative co-design activities, engaging both adult and child users as well as experts, should be considered to refine these designs into those that could be implemented and deployed in practice. Experts' in-depth remarks resulted in significantly longer transcripts and more detailed answers when compared to those of the children. Future research should include a more balanced representation of families and may also require a more specific criterion for experts' recruitment if we need to focus on the psychological aspects or technical aspects for example. Still, the exploratory studies provide invaluable and novel insights into an emerging research field. The interviews were conducted online, which meant that experts experienced "Big Buddy" through videos, rather than in the immersive VR environment. This could lead to missing recommendations that could not have been noted solely from video-based observations. Nevertheless, experts reported finding the introductory presentation insightful and informative. All experts were already familiar with VR and had a high level of awareness of social VR, with a mean of 4 (1-5 scale, with 5 being extremely aware) (σ=1.5) and 75% had practical experience with social VR. Therefore, their perspectives still offer valuable contributions. Furthermore, we present findings around AEMs, but acknowledge that the only experience participants had of AEMs was through the "Big Buddy" example, potentially influencing the results. However, the use of such an example was also a strength as it offered a more practical and hands-on experience of what could be an AEM.

4 FINDINGS

Participants shared their thoughts on "Big Buddy" [33], an example of an AEM designed to intervene in child harassment situations

in social VR. Findings were grouped into two categories across all coding schemes: **1) Perceptions towards the automated moderation** i.e., the components involved in ensuring effective moderation through automation and **2) Perceptions towards the embodiment** i.e., the physical and social components that include the moderator's appearance, ways of communicating and interacting with child users. The structure of the findings' section is shown in Figure 2. Categories obtained from coding schemes are highlighted in grey. The following section describes themes across all three perspectives: experts, guardians and children.

4.1 Attitudes towards the Automated Moderation to Safeguard Children in Social VR (RQ1)

4.1.1 *Benefits of Automated Moderation.*

AEMs as a Guide and Exemplar. Experts found the AEM useful for guiding, educating and reminding users of rules in an easy and understandable way (N=5), addressing the lack of socialising norms in social VR. It could educate harassers and instil good behaviour: *"It's a great easy way to understand what someone is doing wrong, what the codes are, social codes for operating in the space"* [Expert P13], *"The idea is there because there are no real true safety measures in the metaverse at the moment"* [Expert P4]. Experts also valued that an automated system can deliver immediate action (N=9) in contrast to current methods like report delays and human moderation, and particularly in the case of group harassment as it can be easier to detect. It can also offer support (N=2): *"Having somebody to come and help"* [Expert P12]. Some noted that system could be particularly useful in supporting new users and adults.

Immediacy of Response. Guardians also found one of its strengths is the fact that it can take immediate actions (N=3) and it can enforce rules (N=5), there are consequences for actions, and it can restrict access to inappropriate content. It can also give helpful advice (e.g., take breaks) *"I'm just wondering if big buddy could also like see if a child has been logged on to game for a long time. They could say: How about taking a break?"* [Guardian P2P-4].

An Intermediary Between Parents and Children. Moreover, AEMs can potentially help with the parent-child relationship (N=3) as they can facilitate conversation between children using social VR and their parents [Guardian P1G-4] and help children disengage from prolonged VR usage. Some parents mentioned that as their children might be indifferent or not listen when they ask them to stop, Big Buddy could be an effective solution to stop the child from playing: *"The kids aren't going to really care about their parents being told"* [Guardian P3P-2]; *"A big buddy could just see your time and the game is finished [...] because I can't stop them. Especially when they're younger.* [Guardian P2P-4]. Finally, guardian P1G-4 felt it would be useful if the AEM is linked to guardians (N=1) via parental influence over the AEM (e.g., settings overriding the AI system) and parental insight into AEM actions and the events occurring (e.g., via real-time smartphone notifications).

Fair Moderators. Children from the 13-16 year-old group (workshop 4) (N=2) reflected on the AEM and noted that, unlike human moderators, the automated moderator would not be able to intentionally abuse their power and can act fairly if correctly trained. They also mentioned that automated moderators would lower cost and would not need to take breaks, allowing them to be being present in the virtual environment at any time.

4.1.2 *Concerns Around Automated Moderation.*

Technical Feasibility and Contextual Understanding. Experts expressed concern about an AI moderator's ability to address mistakes and technical challenges (N=7) with evolving online harms and social norms. To resolve this, they suggested updates to keep up with emerging and evolving harms and online social norms. It would also require the tool to gather contextual information to accurately detect disturbances. Experts mentioned the automated moderation needs to be context specific (N=4) for different games and environments.

Transparency and Trust. An expert highlighted a pitfall of moderators administering judgement without a visible reflection process: *"The weakness is the moment that any kind of automatic or human moderator comes in, enacts a judgement, and does not create a questioning phenomenological process for everyone to understand how they were all contributing to that. You destroy the group safety and that's going to be hard to train into a moderator"* [Expert P11].

Half of the guardians (n=6) would partially trust such a system. A guardian mentioned they would trust it as it would be impartial and not malicious [Guardian P2P-4] and another would trust it *"as long as there is open communication and I know what is going on"* [Guardian P1G-3]. However, the other half of guardians lacked trust due to concerns it would not be infallible.

Taking the Wrong Actions and Fallible AI. Some experts questioned how the system would address false positives: *"With any kind of automated tool I'm always worried about false positives"* [Expert P2], and that inaccurate judgements may feel intrusive: *"I think that your inaccuracy levels or anything automated is gonna make that feel intrusive a little bit so unless it's 100 accurate which really you can only do with the human still, I think that it's quite intrusive"* [Expert P14]. Some experts also mentioned that an automated moderator may be more suitable for group harassment, as it might be easier to detect, but interventions may be more effective in a one-to-one situations as the resolution can be individualised and personalised.

Guardians had very similar concerns regarding *technical challenges* (N=6). The child can potentially circumvent or outsmart the system (N=3) *"My point is it's not infallible. There's going to be workarounds. The kids can work out to get around it."* [Guardian P3P-2]. This susceptibility is a concern specifically for secondary school children who might exploit the AEM's limitations weaknesses and find ways to be unsupervised and bend rules, according to Guardian P1G-4. Moreover, mistakes and a lack of context recognition (N=4) by the AEM introduces risks, as it may not always detect problematic content or understand the nuances of different situations, potentially leading to false positives or negatives. Thus, some guardians mentioned that human-in-the-loop (N=2) is necessary and *"we must be able to shut it down"* [Guardian P2G-2]. The AI system requires continuous updates and management (N=2) to adapt to the changing dynamics of supervision as a child grows.

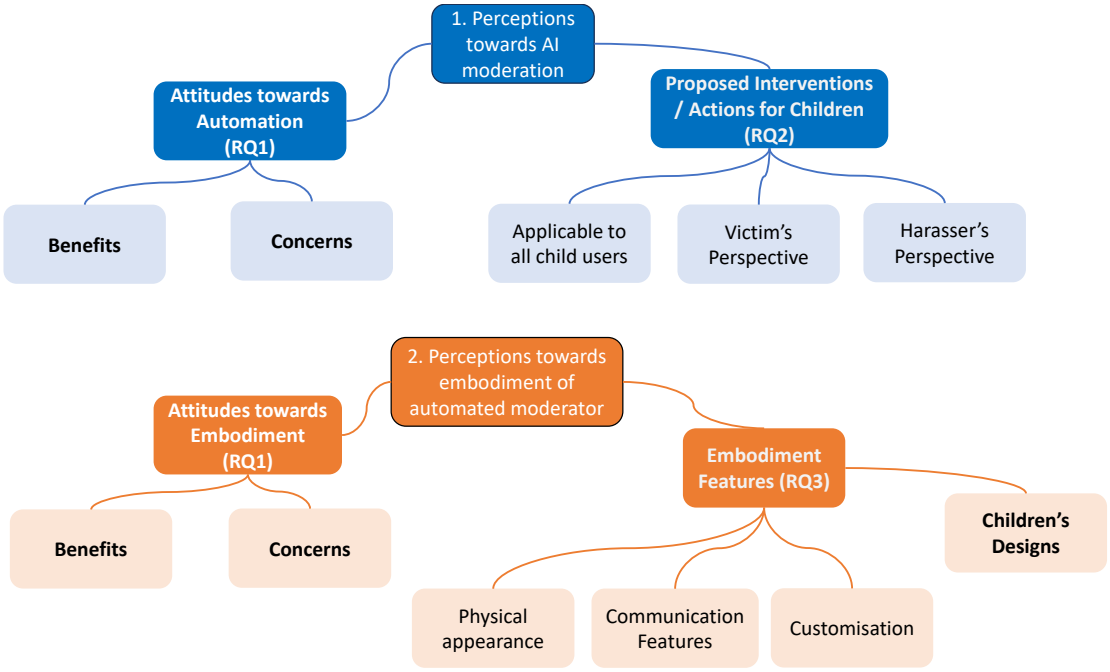


Figure 2: Findings structure across all three stakeholder groups (experts, guardians and children).

Teenagers mentioned concerns about the moderator being AI-based (N=2), as children may abuse the automated system and it might be too sensitive leading to false positives or not accurate enough with languages other than English [Child P1C-4]. They wished for AEMs to be carefully controlled [Child P2C-4] and suggested human involvement and, thus, a semi-automated approach. Child P1C-4 suggested different lines of moderation: the automated moderators detect events and intervene, the human moderators check and the admins finalise decision.

Impact on Children. Experts questioned the long-term impact on children (emotional and ethical issues) (N=9). Specifically, they questioned the emotional ramifications of harsh tactics, saying:

> *"I think it can actually be quite demeaning to the harasser [...] because I'm 99[%] certain that Big Buddy might make a mistake over something that wasn't deliberate or perceived [...] that's because we're still doing the learning in terms of what's acceptable behaviour"* [Expert P16].

Unintended consequences of AI-moderation may lead to risks of self-harm or extreme reactions from children *"My sense of unintended consequences would be massive meltdowns if there was a sense of injustice"*. [Expert P15]. Moreover, there are consequences of how the AEM is designed if it goes beyond just a moderator, for example acting as a tutor or friend: *"I suppose there is a question about how we feel about that child engaging with a tutor. [...] I wonder what Big Buddy's role is because what we've seen is Big Buddy comes when something's gone wrong. Does Big Buddy come in when things are going right as well?"* [Expert P8]. This highlights that the role and scope of involvement of the AEM are not yet defined. Additionally, experts discussed the potential for future and real-world consequences. While the example AEM ("Big Buddy") gives a punishment and removes the harasser, it does not support the victim on how to deal with future situations. Expert P11 also critiqued that setting a player's points back to zero could have a damaging and toxic effect.

Guardians expressed a concern about the negative impact on children (N=3), including their lack of skills to handle incidents, potential negative effects on mental health due to automated moderation children may feel devastated according to Guardian P1P-2, and possibly agitated resulting from warnings and restrictions [Guardian P2G-2]. Moreover, children (N=2) were worried about not being able to refute and resolve an argument.

Privacy. Child data protection (N=2) was mentioned by experts as an important consideration, as investigations and automated interventions would require accessing private content. Experts also had surveillance concerns (N=2) and how it may take away the child's agency. In contrast, children could also circumvent and find a way around it (misusing, provoking or using it as a harassment tool). Expert P17 mentioned the need for long-term research to see if behaviour changes and insights into how children interpret the automated moderator over time.

4.2 Proposing Suitable Interventions of the AEMs for Child Users, Victims, and Harassers (RQ2)

4.2.1 Interventions Applicable to All Child Users. Several approaches and actions for the AEM were suggested to safeguard children and promote safe social VR environments for all users:

Intervention Theme 1: **Rules and Fairness**

1.1) Set expectations and enforce rules. Experts (N=9) recommended the reinforcement of consequences for bad actions with visual and verbal reminders clearly indicating and explaining expectations and rules. Expert P2 proposes disturbance tests with dummy players to show new users what kind of sanctions can be taken for certain situations. Users should be made aware that a moderator is present in the room.

> *"To me the most important thing it does is just tell people more information about the expectations of the space and serve as that visible reminder [...]. So just serving as a reminder that this is still a social space and like any other social space there are expectations, there are rules and there are going to be consequences for violating those rules I think is important. "* [Expert P2]

Experts also suggested children set the rules at the start, either individually or collaboratively. This can include selective positive actions from a list to receive positive feedback from the AEM. Guardians emphasised the importance of enforcing rules (N=3), with sanctions and real-time notifications to guardians from both parties (victim and harasser). They highlighted the significance of reinforcing rules through verbal reminders (N=2). Child P2C-1 want the AEM to set the rules and expectations, it should make sure users follows the rules (N=1).

1.2) Applying Principles of Procedural Justice. Experts (N=2) note that consequences should be consistent and fairly adapted across all ages, applying principles of procedural justice:

> *"I could see a lot of issues surrounding feelings of injustice or unfairness.[...] Looking at the principles of procedural justice and how people perceive the fairness of offline governance systems and criminal justice systems, I would want to apply the same sort of principles to a system like this so that we can ensure that it's as fair as possible"* [Expert P2]

Intervention Theme 2: **Positive Support**

2.1) Positive reinforcement over negative punishments. Experts (N=5) discuss the benefit of positive reinforcement in behavioural change which can be accelerated when compared to negative punishments. This can be done by giving more attention to positive behaviours, shifting the focus from negative punishments, to promote behaviour changes, by giving positive feedback to young people engaging in appropriate behaviours, for example, *"The good thing about switching it so that it praises for positive behaviour is that it's likely to extinguish the behaviour of the negative player much faster and it's likely to model good behaviour"* [Expert P12]. Positive reinforcement is also universally accepted according to Expert P12, which reduces the need for adaptation between groups. Experts noted that negative punishments remove enjoyment, while promoting positive behaviour through behaviour can increase enjoyment. Statements should be predominantly positive and encouraging to emphasise the opportunity to try again and improve. A guardian also suggested positive reinforcement *"with positive rewards or points for those who follow the rules"* [Guardian P1G-3].

2.2) Reflection and learning experience. According to experts (N=7), the AEM should give the opportunity for reflection, repair and rehabilitation. This can be done by pausing the ongoing interaction to facilitate reflection or by notifying both parties (victim and harasser) with post-resolution reflection messages and explanation of wrongdoing to educate children about their actions. Child P2C-4 would want the AEM to give a tutorial on how to use the safety tool and AEM at the start of the game via a pop up.

2.3) Offer Psychological Support to all Child Users. Experts (N=2) suggested strategies that involve not only enforcing rules but also providing psychological support to both the victim and the harasser, by asking follow-up questions (e.g., "Are you okay?" [Expert P3]) or, for example, linking them to child helplines [Expert P10]. Expert P8 suggested monitoring self-harm as well, considering mental and physical health and safety concerns. However, some experts also raised doubts about the role of the AEM beyond a moderator enforcing rules as this could lead to ethical concerns.

2.4) Encourage open discussion and conversations. Experts (N=2) mentioned that the AEM should facilitate open conversations, helping the child to discuss the incident.

Intervention Theme 3: **Adapting to Context and Dynamics**

3.1) Interventions proportional to severity and frequency. Experts (N=8) recommended reasonable and proportional outcomes based on behaviour, depending on frequency, context and the level of harassment and proportional support by the AEM (e.g., more check-ins). Sanctions should depend on the severity but also be balanced with educational intervention. More generally, some experts recommended to build on the history of interactions, for example:

> *"If it just came up with the same answer every time they'd start to sort of ignore it um and I think that's somewhat similar for the harasser I think you can you know when we talk about sanctions we talk about a matrix and an incremental path through a sanctions you know based on how many times you and what you've done the that the responses could be different and they would feel more in response to the actual situation".* [Expert P14]

To transition from warnings to more stringent enforcement, guardians (N=2) suggested employing strike warning systems. Children also suggested the use of a strike system: from warning to banning to increasing banning time (N=6) where the AEM gives a certain amount of warnings before taking the action such as banning the user. If the inappropriate behaviour is repeated, the ban gets longer. Additionally, some children would want the AEM to gather evidence when an incident occurs so that it can be re-watched to finalise the outcome. For instance, audio recorded within a radius of the child within the VR scene or a video clip summary sent to a human moderator. This would also allow the child to more effectively engage with their parent regarding the incident in specifics, rather than a parent just seeing the end result.

3.2) Interventions adapted to one-to-one harassment VS group harassment. In group harassment i.e., a group harassing a child, experts (N=10) discuss approaches that include encouraging groups,

and in particular those leading the groups, to reflect on their behaviour, for instance:

> *"Group dynamics are very different, there's always an alpha person that is leading the group and the others are followers so it should be customised depending on who you are dealing with."* [Expert P7].

Body language monitoring of the victim, harasser but also bystanders could be used to detect group harassment according to Expert P11:

> *"In group dynamics there's usually an aggressor and a victim, everyone else is often bystanders and their body language is going to shut down, freeze and get very awkward. I think there could be a great opportunity to be monitoring for that kind of body language" [Expert P11].*

Thus, taking more immediate actions in the case of group harassment was recommended. However, they also suggested focusing on individuals to break the group effect, balancing individual warnings and group approaches.

3.3) Private VS Public conflict resolution. Experts (N=2) question the effectiveness of public or private resolution. An expert suggested involving bystanders anonymously by posing questions that prompt them to reflect on the situation and contribute to a different outcome.

Child P2P-4 also mentioned not wanting punishments of actions that are acceptable between friends.

Intervention Theme 4: **Human and User Involvement in Decision-Making**

4) Human-in-the-loop. Experts (N=2) suggested the involvement of a human, with either the user being able to call the AEM in case it does not pick up the issue, or the moderator being automatically triggered and the intervention or outcome being decided by a human. They note a tension around how, and to what extent, humans work with AEMs and question if the AEM should be the first line of defence or if it should take decisions on interventions.

Intervention Theme 5: **Non-Verbal Interventions**

5) Non-verbal Interventions. Experts (N=2) suggested the AEM to act as a *"buffer"* [Expert P9] and separate the victim from the harasser, *"putting their avatars in opposite sides of the room"* [Expert P17]. Some children (N=3) also proposed non-verbal actions taken by the moderator: *"a virtual hug"* [Child P2aC-1] to comfort the victim or *"a virtual punch"* [Child P1C-2], *"slap 3 times"* [Child P1] to the bully.

Intervention Theme 6: **Immediate Safety**

6) Immediate Actions and Notifications to Guardians. Guardians proposed immediate actions (N=3) that include shutting down the source of harassment, blocking, bubble the harasser and the victim, immediate time out and debrief to parents. Children also suggested that the AEM takes direct action (N=8) in a child harassment situation. These direct actions include blocking, reporting and banning. Children also want notifications and messages (N=2) when an incident happens to be sent either to parents, to the victim saying the moderator went to the wrongdoer and then the outcome to the bully.

Additional approaches and actions proposed by Experts. An expert mentioned there should be the possibility to complain about the AEM and switch it off (N=1). There should be a balance between offering guidance and allowing user autonomy (N=1), light touch with occasional interventions. In contrast, one expert suggested maintaining a constant moderator presence (N=1), creating a situation where users cannot easily distinguish when they are being monitored, similar to how speed cameras are not always active. Experts also suggested age verification (N=1).

Although this research focused on child users, it was mentioned by experts that for adult harassers targeting a child, consequences should be more severe (N=3). Experts suggested adapting interventions to age groups. Other experts recommended to adapt interventions based on development stages and emotional skills rather than just age (N=4).

4.2.2 Interventions from Victim's Perspective. Several approaches and actions for the AEM were suggested specifically to victims.

Victim's Intervention Theme 1: **Victim-Specific Psychological Support**

1.1) Emotional Support, Empathy and Comfort. For the victim, experts proposed emotional support and empathy (N=9). For example, the AEM could offer emotional assistance and check-ins to the victim, asking how they are feeling, if they want to continue, communicating with personal messages and helping build back their confidence. The AEM could offer a reflection space and connect the victim to resources (e.g., child helplines) recognising potential trauma and ensuring the child's well-being. Overall, the moderator should be empathetic. Guardian P1G-4 suggested to comfort the victim, asking the victim how they felt and what they think should happen and administer a ban level based on responses. They suggested to bubble the victim.

1.2) Supporting Language. Children (N=4) emphasised the importance of supportive phrases being used with victims. This included informing the victim that the moderator is going to intervene in a situation involving a wrongdoer, according to child P3aC-2. There were also phrases designed to assist and provide advice to victims in handling difficult situations. Messages like *"The Player has been banned"* [Child P3aC-2] were introduced to reassure victims, and there were phrases that instructed victims on how to report an issue *"You can report, press this button to report"* [Child P1C-3]. Furthermore, phrases like *"Please wait while we talk to the player (bully)"* were suggested to maintain transparency during interventions [Child P1C-3].

Victim's Intervention Theme 2: **Empowering Victims through Voluntary and Discreet Active Engagement**

2) Empowering Victims. Experts also suggested empowering the victim with decision-making (N=8). On the one hand, the AEM could automatically be called to offer assistance with options of actions (N=4) to take such as blocking or protective bubble. On the other hand, it could be encouraged that the child reaches out to

the AEM (N=4) rather than the moderator stepping in, invoking it when needed. Experts also note that victims can become harassers, therefore, supporting the victim is important to avoid perpetuating cycles of aggression (N=2). Children also proposed to involve children in the decision-making with discretion (N=3). Child P1C-3 suggested:

> *"I think what you should do if somebody does that to you, a pop up that says "this is a request to kick this person or report this person" and if you request yes then you can report them that's only when you can report them, only if the computer senses that something went wrong.".* [Child P1C-3]

Another child suggested a button to report and call the AEM:

> *"Izzy [Made-up victim name] says: "I'm reporting and pressing the button". [...] Then there is this notification saying "Hey Izzy I went to tell Rob [Imaginary harasser] off, I will notify you when he is banned." "* [Child P3aC-2].

Child P2C-4 suggested, however, the possibility to change the outcome temporarily, for example unblocking someone when joining a private room with friends:

> *"If you have a friend who invites you in a room with the person you've blocked, you should be able to choose to unblock temporarily just for that room and then after immediately they are blocked again."* [Child P2C-4].

4.2.3 Interventions from Harasser's Perspective. Several approaches and actions for the AEM were suggested specifically to wrongdoers.

Harasser's Intervention Theme 1: **Stepwise Interventions from Warnings to Repair to Sanctions**

1.1) Stepwise approach. The use of a stepwise approach (warnings before sanctions, severity increased if repeated) was recommended by experts (N=4), with the opportunity for behaviour to improve before sanctions. If the inappropriate behaviours are repeated, the severity of sanctions can also be increased. Experts suggested that warnings should be private first, as children can become emotionally invested in these games and might react strongly in public. In fact, expert P11 mentions that there could be potential dangers of completely excluding the harasser directly, leading to triggering real-world negative consequences and inappropriate behaviours. Children suggested in cases where a victim had reported the bully, private warnings like "You have been reported; your fate will be decided" [Child P1C-3] or giving more chances before consequences such as being banned.

1.2) Notified with Reasons for their Consequences. Experts also suggested they should be notified of their actions with reasons for their consequences (N=3). Children proposed phrases that explicitly highlighted code of conduct violations *"You broke code of conduct / tos [terms of service]* [Child P3aC-2].

1.3) Opportunity for Reconciliation. Experts recommended an opportunity for the harasser to explain and apologise (N=5) before the outcome. This can be done with personal messages of the decision to the offender with the opportunity to explain themselves if they disagree with this decision. Similarly, children proposed messages that pointed out the negative impact of their actions on victims.

1.4) Considering victims of the past. Expert P9 reflected on the harassers' background, mentioning they are usually victims of the past so directly negatively punishing is not the best solution:

> *"If you just use the moderator to convey punishment, I don't think it's very effective for the harasser because there are some problems with the fact that you might be harassed in your family so you just repeat this chain so you don't know how to you know interact with others and you're just repeating"* [Expert P9].

Harasser's Intervention Theme 2: **Focusing on Reflection and Educating to Improve Behaviour**

2) Reflection and Education. The moderator could also provide a space for reflection, training and educating the harasser (N=8). The latter can be done through a test, game or any interactive learning process that add motivational opportunities (e.g., regaining points through participation of a test before re-entering the game). Gamifying the process of reflection for harassers could help them better understand and rectify their inappropriate behaviour. One suggestion was putting the harasser in a safe zone as a reflective zone education break with the AEM which they must complete or receive a time-ban. Another example of interactive learning process was suggested, where the AEM guides the harasser to step into the shoes of their victim promoting empathy:

> *"I think that interventions where the AEM asks the harasser to say "hey you did this thing before and I really want to know what that feels like" and they switch into that other person's avatar and they have that harasser do that instance with them and then they make an observation"* [Expert P11].

It can help harassers understand the impact of their actions, which could, for example, be of use for neurodivergent children: *"Particularly if they've got learning difficulties or autism there is a way that you could give an explanation"* [Expert P16]. Additionally, the harassers can be encouraged to reflect on their actions and the moderator can show empathy and explain reasons for the consequences. Experts anticipate that the latter intervention can have a positive impact on their real-world behaviour as well.

A child also proposed a follow-up before the harassers come back in the game *"They get banned for a day. Before joining the game they have to say they won't do it again. The fate will probably be a ban for one day. Then, they go back in the game but they have to say "I won't do that again"* [Child P1C-3].

Harasser's Intervention Theme 3: **Immediate Actions against Harasser Prior to Providing an Explanation**

3) Immediately Bubble Harasser. In contrast to experts suggesting reflection and explanation prior to the consequence of action, guardians and children suggested a direct time-out for the harasser before repairing. For example, guardian P1G-4 would want the AEM to *"blackout offenders straight away, then contact offender to debrief and ask them for an explanation of their actions"* [Guardian P1G-4]. Children also suggested time-out and putting wrongdoers in a bubble or restricted area, outlined in phrases, such as a 20-minute timeout *"You will be timed out for 20 min"* [Child P2] or a 30-minute cool-down period *"Calm down 30min then come back."* [Child P3].

4.3 Attitudes towards the Embodiment of the AEM for Children in Social VR (RQ1)

4.3.1 Benefits. Experts identified several strengths of an AI-embodied moderator for children. Experts refer the role of AEM ("Big Buddy") as modelling good behaviour (N=11) similar to a *"teacher", "referee", "football coach", "police officer", "brother" or "parent figure"*. It can model civility, kindness and can exemplify good behaviour via its communication and interaction. The "*embodied version has value over a just a voice*" [Expert P9], being beneficial as its visibility and presence offers a sense of protection, comfort and authority (N=5). Its physical presence can also be used for physical interaction dynamics, posture and position in the scene, such as a "buffer" between the victim and harasser.

Guardians noted that the embodiment of the AEM has notable strengths (N=4), including being friendly and hero-like, and instils a sense of protection and authority in users. This presence of an authority figure also contributes to feelings of security and comfort, *"similar to the reassurance provided by CCTV surveillance"* [Guardian P2P-1], leading to self-moderation among users.

4.3.2 Concerns. Experts had concerns related to the embodiment, such as potential systemic bias in design, where it may reinforce stereotypes, as well as worries that children might feel watched, not behave authentically and lack agency and control. They worried that the robotic voice and appearance of the moderator would lead to Credibility issues (N=5) and provoke skepticism from users about its effectiveness and seriousness. Finally, an expert suggested that the embodiment should blend more seamlessly into the social space and some experts mentioned it should engage more like a human to enhance credibility.

Guardians also raised weaknesses associated with the embodiment (N=3). Guardian P1G-4 mentioned the "Panopticon Effect" the sensation of being constantly watched can diminish the enjoyment of the child's experience. According to guardians, the effectiveness of embodiment would vary with age (N=2). Guardians find that it may not be suitable for teenagers, *"too childish"* [Guardians P1P-1, P1G-3], and guardians feel that it needs to be more depersonalised for older age groups. It was viewed as suitable for certain age groups, particularly for children from early ages up to 11 years old, where participants felt it would help maintain order and considered it appropriate. However, its perceived effectiveness diminishes for older children, as they may not respect it. There was also a concern about the potential reinforcement of stereotypes of behaviours and authority figures.

4.4 Proposing Embodiment and Design Features of the AEMs (RQ3)

4.4.1 Embodiment and Design Features: Experts and Guardians.

Physical Appearance Considerations. Experts' opinions on general appearance considerations for AEMs highlighted several key points. Neutrality (N=4) was emphasised to build trust and avoid reinforcing stereotypes. To achieve this, using non-human representations like wise owls or fantasy creatures is suggested. Additionally, neutral features resembling emojis or familiar cartoon characters are recommended for broad cultural acceptance. The idea of an authority figure was mentioned, stressing, however, that it should not reinforce stereotypical traits *"It does not have to be a man or a human"*. Approachability and relatability (N=3) are also mentioned to be important for effective engagement with the moderator. Striking a balance between approachability and authority was advised. An expert raised a question between robot-like and human-like moderators, considering emotional transference and user perceptions, as well as the impact of the moderator's height on user perception. Visual feedback (N=1) also plays a role in appearance considerations. Suggestions include using different emotional faces as visual cues, such as empathetic faces for victims and sad faces for harassers.

Communication Considerations. Experts highlighted communication considerations of the AEM. A friendly, warm and encouraging (N=6) moderator was recommended rather than an authoritarian enforcer to enhance enjoyment, foster inclusivity and unity. It should also be more empathetic and can use playful moderation: *"It could separate users in a fun way if it's detecting a lot of yelling between two different people it would be kind of funny if he could just immediately put people to the opposite sides of the room."* [Expert P17]; or whimsical communication of rules: *"[...] kind of funny, kind of whimsical but actually communicates the rules in a way that you don't dismiss it but it maybe it could make you laugh at the same time so that you actually see it as a piece of the environment and a part of the game um and it's not annoying"* [Expert P17].

Moreover, a natural and human-like, less robotic approach (N=6) was suggested, focusing on decision-making like a human and giving the opportunity to reply, as well as using understanding language and slang to connect with users. For a more natural approach, it should also be able to adapt responses to avoid redundancy. An expert suggested the use of a voice with personality and a natural conversational tone to improve user engagement. Based on the uncanny valley theory [58], "Uncanny aspects" that include features or behaviours that seem too artificial or not quite human-like can make users, especially children, feel uncomfortable or reluctant to trust and interact with the system, according to Expert P9. Considerations also include consistency of communication characteristics across victims and harassers, and tailored approaches for vulnerable children to improve trust towards the moderator. However, Expert P17 suggested a stern tone when engaging with the harasser.

A guardian suggested incorporating playfulness and positivity (N=1) into moderation to encourage rule adherence: *"Adding playfulness (Fun theory) to follow rules Again just positive"* [Guardian P1G-3]. Another guardian advocates for communication features (N=1) like using a soft tone and recognising the presence of adults, which would enable the moderator to adapt its communication style based on the audience, whether they are adults or children.

Customisation. Customising moderators based on age groups (N=15) is a prominent theme. Experts stressed the need to treat older children and teens (13-16 years old) differently from younger ones (early ages to 12 years old), while maintaining the concept of a protector. Customisation of appearance and speech can enable age-appropriate communication: *"If it's a childlike environment you would speak and act in one particular way. If it's a slightly older environment it's speaking and acting in a different manner as well so it's understanding the audience"* [Expert P4]. In terms of age-dependent appearance (N=4), this includes age-appropriate customisable appearance from fantasy characters to established ones such as

"cute little dog, cute monster, paperclip" [Expert P14], "little imaginary friend" [Expert P2], "Clippy" [Expert P17]. In terms of communication (N=11), they suggested adapting voices, languages and tones for based age groups, developmental stage and preferences. For example, users can opt for gentle versus blunt approaches. The moderator should adapt its behaviour to the environment; if there are children present, different cultures, languages and accents. Options could be given at the start for children to choose the way the moderator guides and interacts with them. In terms of approachability and trustworthiness (N=7), experts suggested non-intimidating non-verbal approaches, such as having the moderator meet the child's eye level to reduce intimidation, and using of gentler and softer ways of communicating across ages, specifically, with a more authoritative stance for teenagers. They also proposed customisable approaches based on trust, with a questionnaire at the start to help tailor the moderator's interactions and appearance based on the answers.

Additionally, guardians recommended options for customisation (N=2), allowing users to personalise the moderator's features for a tailored experience (colours, animals, gender, non-human etc.).

4.4.2 Embodiment and Design Features: Children. Findings mainly encompass children's ideas regarding: appearance of the moderator, when they would like their moderator to appear, where where they would like to appear. A portion of children's answers regarding physical appearance, when and where their moderator would appear are displayed in Figure 3, highlighting creative designs and drawings. Children's storyboarding designs can be found in Figure 4 and Figure 5.

Appearance and identity. Children suggested a Human-like appearance (N=1, Child P3bC-2) for the AEM. Some proposals leaned towards an Animal-like (N=5) form, including options such as *e.g., a snake [Child P1], their pet [Child P2, P3], a cute lion for the victim transitioning from cute to a scary lion for the bully [P2aC-1].* There were also suggestions for designs that did not fit neatly into either category, encompassing original creations like *"a cloud"* [Child P1C-1]. These diverse form factor ideas highlight the creativity and variety in envisioning the automated moderator's appearance.

Children proposed incorporating Pre-existing Media Characters into the moderator's appearance, including figures from cartoons and superhero comics *e.g., Wonder Woman [Child P1C-2] or Superheroes Comics [Child P2C-2],* and fictional game characters (N=3) *e.g., "Pikachu would electrocute people who are misbehaving" [Child P3aC-2], "Princess Peach" [Child P1C-2].* In contrast, there were suggestions for Original Characters that would be unique creations specifically designed for the role of a moderator. Some designs leaned towards a Naturalistic approach, with suggestions like a tall, realistic police officer figure [Child P1C-3]. In contrast, the Supernatural design category featured suggestions like embodied moderators for younger children and victims and creatures that are scary but not petrifying *"something scary but not petrifying and realistic: not a monster or a teacher"* [Child P1C-3].

Relationship to User. Participants proposed a Friend Figure design (e.g., their best friend [Child PC2-1] for the automated moderator, suggesting it could take on the role of a supportive companion or their own pet as mentioned above with the same name. Another child suggested a Family Figure [Child P2] to provide a sense of familial connection and comfort. Additionally, the concept of an Authority Figure emerged, with suggestions ranging from powerful animals (Lion) [Child P2aC-1] to police officers [Child P1C-3], all designed to enforce rules and deter inappropriate behaviour, scaring the bully. Child P1C-3 also suggested players can choose the name of their moderator to feel safer.

When the AEM should appear. Children expressed various preferences regarding when they wanted their automated moderator to be appear. Some children advocated for the moderator to be Always (N=4) there, particularly when there were multiple players involved, ensuring a constant presence to oversee interactions. Others favoured the moderator's intervention When something happens (N=7) with options for it to stay invisible during periods of inactivity and to be triggered by specific words or incidents. Additionally, there was a consensus that the moderator should only become active once a child approves or decides (N=3), with features like a button for calling the moderator or its appearance after the child clicks a report button. A child also suggested implementing the moderator during "Peak times based on the number of player" (N=1, Child P3bC-2) to manage higher activity periods.

Child P1C-3 preferred a more subtle or invisible presence for the moderator. This could involve it being *"Always there but invisible or behind you"* or having a small pop-up and a larger buddy on the side to avoid disrupting the gaming experience. There was also a suggestion for the moderator to become visible once it detects inappropriate behaviour.

5 DISCUSSION: TOWARDS AEM TO SAFEGUARD CHILDREN IN SOCIAL VR IN PRACTICE

5.1 Summary of Findings

Our work provides an important first step in reflection around, and recommendations for, the design and use of AEMs to safeguard children in social VR. We merged perspectives from experts, guardians and children, providing an enriched and holistic set of insights. Summary maps of findings from each perspective on *automated moderation* and *perceptions towards the embodiment of AEMs* are displayed respectively in Figure 6 and Figure 7.

While there is consensus across stakeholder groups regarding most recommendations and preferences around AEMs, there is a contrasting perspective regarding the extent of AEMs' involvement: children favour a more passive role for the AEM, intervening only when necessary, with discretion and immediate actions, whereas adults anticipate a broader range of engagement with discussions and activities beyond intervening when an incident happens. There is also a tension between making the AEM friendly without being intimidating, while still having authority and being credible. Research is needed to understand which recommendations to prioritise and which are complementary for increased effectiveness.

Children and adults choose to join social VR to enjoy fun immersive and embodied social experiences, to escape from reality [50, 52] and benefit mental health [28]. However, it is a space lacking social norms and effective moderation, especially child-centered

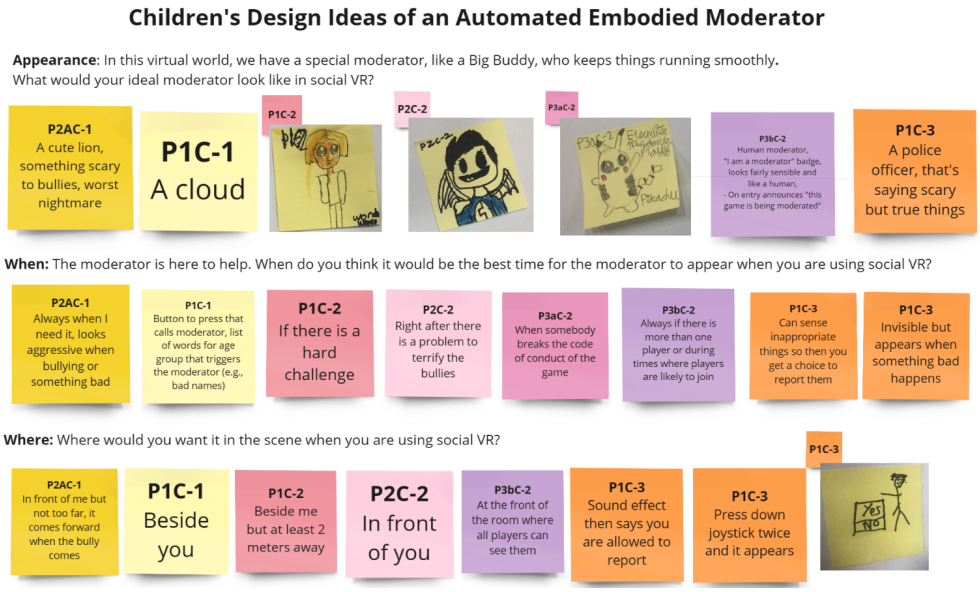


Figure 3: A selection of children's post-its transferred to a digital Miro board, regarding physical appearance, when and where their moderator would appear.

solutions. Assuming a problematic event occurs in VR and the automated moderation system can detect said problematic event (e.g., via the biometrics and social/contextual signals that VR headsets and platforms can provide [34, 71]), our work shows that AEMs appear to be suitable for children and desired by experts and guardians, potentially able to tackle most challenges current safety tools and human moderators have (see Section 2.2.2). The motivation behind AEMs is to create *safety-enhancing technology* that can react instantly to harassment or problematic incidents, offering help and a sense of protection [32]. And with the affordances VR offers, we can not only introduce embodied moderators, but also tailor their appearance and interactions for each user.

5.2 Specificity of Findings to VR versus Non-VR Social Media

Compared to non-VR social media (e.g., Instagram, Twitter/X), VR offers the feeling of presence, with users inhabiting social virtual environments with embodied interactions, replicating face-to-face experiences, where almost all senses are engaged (vision, hearing, phantom touch [12]). Therefore, *embodied* automated moderation is particularly unique to social VR, and brings with it notable benefits and challenges compared to how moderation is enacted on non-VR social media.

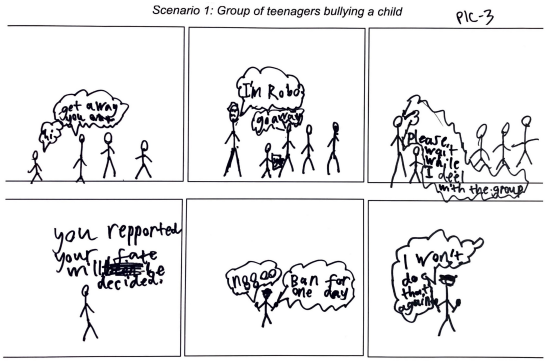
Differences of Our Findings in Social VR to Social Media. Firstly, the capacity of surveillance in VR, raised as a significant concern by our participants, is increased, given the additional movement, physiological and behavioural data available [10], and our participants speculated this could impact perceptions towards moderation over time through the "Panopticon Effect" [62] where users feel like they are being watched.

Challenges related to circumventing automated moderation also become more pronounced in social VR compared to social media where interactions are predominantly text- or image-based. In social VR, users may resort to forms of embodied, real-time harassment (e.g., physically, environmentally) which makes automated detection (previously relying on e.g., textual analysis [59]) more challenging. Unlike social media which retains users' preferences over time, AEMs in social VR may also depend on new sensitive biometric data related to a child's behaviour, emotions, and physical reactions. This includes biometric data, such as facial expressions or physiological responses, raising new privacy concerns.

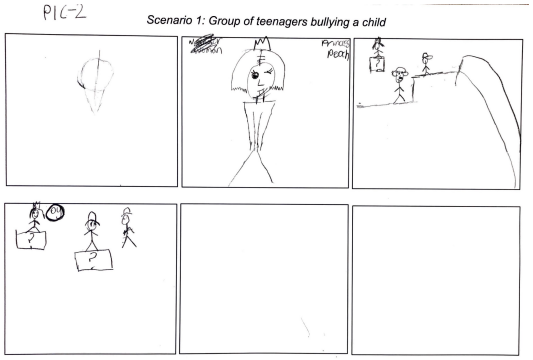
Additionally, VR offers the unique advantage that AEMs can be perceived as being *socially present* in similar ways to real authority figures, which our participants noted could give a sense of protection and potentially support AEMs being 'role models' to children. This characteristic draws a more direct parallel to traditional figures such as teachers, coaches, or parent figures and adds a layer of symbolism and authority that is challenging to replicate in 2D social media moderation.

With respect to interventions, while some proposed from our findings (Intervention Themes 1-4, Victim's Interventions and Harasser's Interventions) may also be relevant to social media, the way these are presented differs in VR due using embodiment, immersion and continuous real-time interactions, which lead to an increased importance of interventions by the AEMs being synchronous rather than asynchronous (Intervention Themes 5,6, Victim's Intervention Theme 2, Harasser's Intervention Theme 3 and Section 4.4). Non-verbal interventions stand out as specific interventions to VR, with the AEMs acting as a 'buffer' to resolve a conflict, or offering 'virtual hugs' to comfort the victim. Dealing with harassment and abuse in social VR also requires immediate actions, due to it being more confronting and continuous compared to 2D social

(a) Design Child P3aC-2 "*This guy, Rob, is bullying Izzy*: "give me your address so we can date" *and because it's against the rules to try and date minors, Izzy says* "I'm reporting and pressing the button", *here the button is red but I don't really want it red. Then there is this notification saying* "Hey Izzy I went to tell off Rob, I will notify you when he is banned". *Then there is Pikachu moderator saying to Rob* "I'm angry you broke TOS" *then* "Rob has been banned permanently". *I chose Pikachu because he would electrocute people who are misbehaving*"

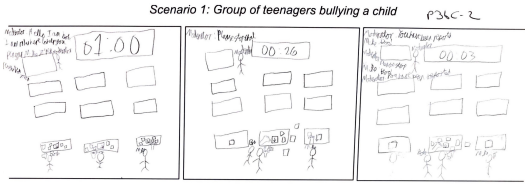


(b) Design Child P1C-3 "*As soon as they log in, children, to progress, they have to give it a name. The victim being harassed says* "Go away" *then the moderator appears, it's taller than everyone, says* "I'm (NAME), here Robo". "You" *is the victim. Robo says*: "Please wait until I talk to the bully". *You have to press the button for the group to be able to see the moderator saying* "You've been reported your fate will be decided". *The fate will be a ban for one day. They go back in the game but they have to say* "I won't do that again". *If it happens again, they get 3 warnings then a ban for much longer.*"



(c) Design Child P1C-2 "*My moderator from the Mario (Princess Peach). When something happens she says* "Oy, why are you (Luigi) hurting (Mario)" *and then Luigi runs away*".

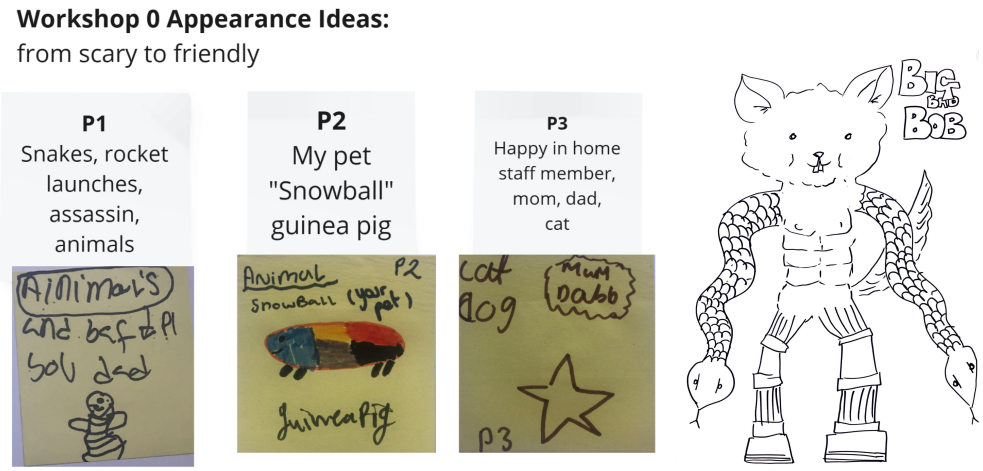
(d) Design Child P2aC-1 "*Name of the moderator: Kyle, it's from my favorite cartoon, the Lion King. It looks like a Lion. If the person becomes mean it becomes scary.*"



(e) Design Child P3bC-2 "*The moderator announces they are here. Player Mike joins and then player Bob joins. Bob's knocking Mike's tower over. The moderator says* "Can you please stop that" *and then he does it again, the moderator* "you've been reported"."

(f) Design Child P2C-2 "*Name of moderator is Monitor (from Dark Monitor Superman comics) a guy with wings and asks the bully to stop mocking the other child.*"

Figure 4: Examples of children's storyboarding designs with their created automated embodied moderator in a scenario of harassment.



(a) Physical Appearance Post-its Workshop 0.

(b) Combined Design from Children in workshop 0 (cat, snake, assassin)

Figure 5: Designs from workshop 0 combining children's moderators' ideas into one final design (cat, snake, assassin).

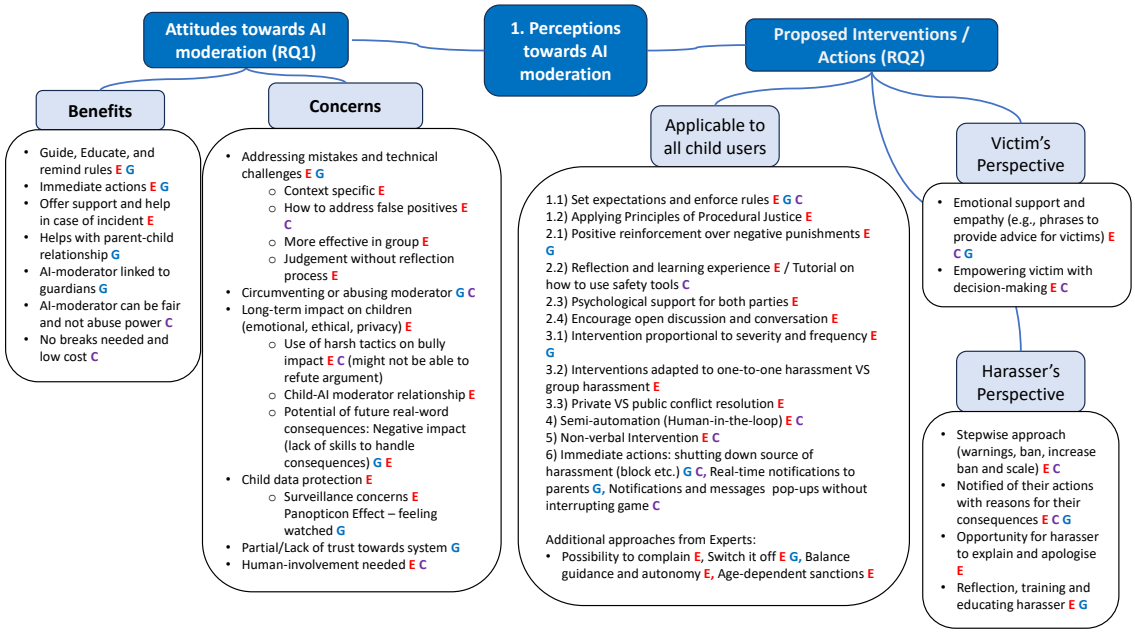


Figure 6: Summary Map of Findings regarding Automated Moderation with corresponding perspectives (E for Experts, G for Guardians and C for Children).

media. Such actions may include promptly shutting down sources of harassment or sending real-time notifications to parents.

Finally, the distinction between resolving conflicts privately and publicly is harder to manage in social VR, due to the complications of real-time face-to-face interactions in virtual rooms. In contrast to social media, which primarily handles issues privately, the dynamics in social VR can mirror a classroom-like scenario where a teacher might address an issue in front of all classmates or opt for a private discussion with a student during recess. This underscores the need for nuanced conflict resolution strategies tailored to the different social dynamics in VR environments.

Similarities of our Findings in Social VR to Social Media. Studies have revealed children's openness to automated and semi-automated monitoring approaches [13, 53] and emphasise the importance of incorporating risk-coping strategies, including victim-oriented and quick interventions to improve their well-being, such as watching

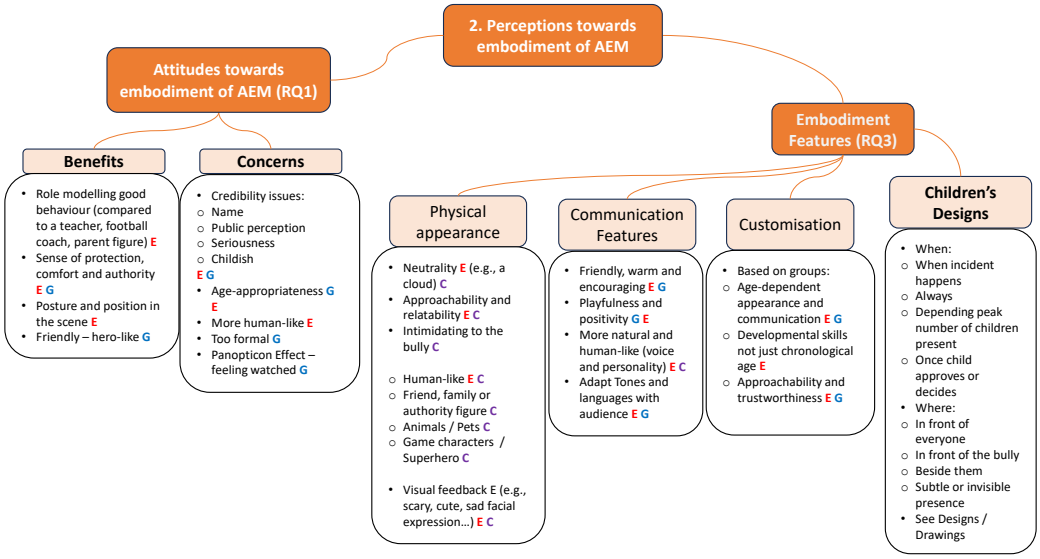


Figure 7: Summary Map of Findings regarding Embodiment of Automated Moderators with corresponding perspectives (E for Experts, G for Guardians and C for Children).

cat videos or suggesting activities like playing with a sibling [53]. Interestingly, these prior investigations also raised concerns similar to those in our study regarding the feasibility of automation. Children expressed worries about the app potentially misclassifying conversations as risky and unnecessarily notifying parents, thus escalating situations [13].

Our findings also highlight the importance of designing proactive solutions to safeguard against risk exposure not just after-the-fact but at every stage of the risk (before, during and after-the-fact) and designing for guidance for the teenagers with live assistance to help them respond after the risk [11]. Researchers similarly identified the potential of restorative justice in understanding and addressing adolescents' needs in online harm [69]. However, in social VR, restorative justice can be enacted differently, via non-verbal communication, 3D interactive simulations and spatial engagement.

5.3 The Role of an AEM Pre/During/Post-Event

If we consider the lifecycle of a moderation incident, and informed by our findings, we envision how an AEM could work in practice. We highlight key considerations around the role of AEMs before the incident; what actions the AEM should take during the incident; and in the long-term what impact AEMs have on children, and how their role may evolve.

5.4 Before the Incident

5.4.1 The AEM's role. When children join social VR, the AEM should introduce itself and explain its role and scope. When present in a room it should transparently notify users. According to the Information Commissioner's Office (ICO) [9], children must have the right to know if they are being monitored, with an "obvious sign to the child" [9], and AEMs offer an embodied, relatable route to providing such transparency. Children suggested a verbal announcement and introduction of the AEM if it is present in the social VR space.

Instead of relying solely on negative punishments when incidents occur that are similar to punishments in schools [33], the AEM should establish clear expectations about appropriate behaviour and consequences for actions from the start and most notably and importantly, employ positive reinforcement and restorative justice. The latter is an approach highly recommended by experts as it may be more effective for long-term behavioural change [70]. Considerations regarding privacy and child data protection should also be taken into account, including what data can be accessed for the AEMs to initiate actions ensuring transparency for users, for example.

5.4.2 Child-Centred Solution. Children should be able to customise their AEM, thereby enhancing their sense of control, and potentially also trust. Indeed, our experts emphasised the AEM should avoid reinforcing stereotypes and consider cultural factors, promoting social cohesion and avoiding labels by suggesting neutral appearances or letting the child design its moderator's appearance features (from non-human to human-like to existing fictional characters) and communication features (from language to accents and tones). The child can also set rules, decide when they would want help and where it would appear and if they would want to be guided (fully automated moderator) or have a say in the decision-making (semi-automated). These settings would be useful as they would allow crucial adaptation for children of different ages, development skills and personal preferences.

5.5 During the Incident

5.5.1 Automated Detection. In case of an AEM that automatically detects incidents, it is important to note that incidents can be subjective. While the goal is to achieve high accuracy in detection, procedures must be in place to address errors, which could be false positives resulting in unintended consequences by misinterpreting a situation, or false negatives where the system fails to recognise an incident because of ambiguous or subjective harassment. This procedure may involve a human (either the user or trusted third party like a human moderator or guardian) to verify outcomes, assist with decision-making, and explore the possibility of reconciliation. We propose the following workflow: the AEMs detect events, then discreetly ask the victim for confirmation; if confirmed, the intervention occurs. This approach aligns with adult user preferences for Human-User-AI collaboration in social VR settings [65].

The role of bystanders should also be considered as a means of improving the reliability to detect said events. Bystanders may explicitly witness and report incidents, or may implicitly (e.g., via body language) point to their occurrence. Automated detection of harassment in social VR will require further multidisciplinary research, from AI to psychology, to HCI, for accurate harm detection and interpretation, and for developing strategies to address mistakes and system failures.

5.5.2 Taking Action. Once the incident is detected, tailored interventions to harassment type, the environment, interaction types (e.g., one-on-one vs group interactions) should take place. Responses should be proportionate to severity and frequency, ranging from warnings to immediate actions. In contrast to experts, there was a particularly high prevalence of guardians and children recommending immediate actions (sanctions) prior to reflection and repair. Some immediate proposed intervention approaches also seemed to align with the concept of altering the physical environment, such as creating greater separation between individuals, limiting their mobility near one another, or relocating to entirely different spaces. There are tensions between proposed intervention methods between the groups we talked to, and research is needed to build on our findings and understand which recommendations to prioritise and which solutions are complementary for increased effectiveness.

5.6 After the Incident

5.6.1 Psychological Impact. Immediate actions can stop the harassment, but the AEM's intervention and embodiment must be carefully considered for their impact on children, whether harasser or victim. Empathy and understanding the emotional impact of actions are important for both victims and harassers as harassers could be victims of the past and victims may become harassers [29, 31]. Therefore, interventions must support the mental wellbeing of child users. As agreed by our experts and prior work [66], harsh tactics like negative punishments without explanation, reflection and support, may lead to self-harm or extreme reactions from children. The actions the AEM takes must avoid causing damage and toxicity and instead focus on positive reinforcement and facilitating conversations and open discussions. They should also involve the children in the decision-making process, with discretion if preferred, to empower them. From our findings, it is also suggested that the AEM serves as a link, connecting the child users to guardians or other resources like child helplines.

5.6.2 The AEM's Role VS Guardian's Role. By imbuing AEMs with customisable, human-like personality, it can potentially become akin to a tutor or a friend. However, this ambiguity of scope raises ethical concerns, necessitating the clear definition of an AEM's role and continuous evolution within the system with adaption of rules over time and ensuring transparent explanations of system functionality and limitations to child users.

Guardians may come to rely on the AEM which could lead to either improving guardian-child relationship in the case of children becoming indifferent to rules set by the guardians or in contrast. However it might also hinder the responsibility of the guardians, relying solely on the AEM and neglecting parenting duties. Children may also miss out on valuable opportunities for self-development and building social resilience, diminishing their ability to stand up for themselves. Therefore, control and agency, reflection and learning should be accounted for as part of the design of the system.

6 CONCLUSION

We have presented two empirical studies to address the gap in understanding AEMs to help children in harmful experiences in social VR. The studies serve as groundwork to gain initial insights into the design space of AEMs. Driven by individual interviews with 16 experts in child online safety and psychology and workshops with 13 children and 8 guardians, we collected recommendations and reflections regarding the automation and the embodiment of AEMs to safeguard children in social VR. Experts, guardians and children see benefits of AEMs for children such as enforcing rules and 24/7 availability in a virtual space that lacks social norms. Intervention approaches should focus on clear communication of expectations, positive reinforcement, and adaptive responses to harassment. Victims need emotional support and decision-making capabilities, while harassers should be held accountable with educational and repair opportunities. For AEM embodiment, neutral or player-customised forms are suggested, with adaptable communication and natural voices. Future interdisciplinary research remains to tackle design challenges raised. This paper proposes design considerations towards AEMs in practice, around the lifecycle of a moderation incident with the emphasis on tailoring the AEM to each child, human/guardian-in-the-loop and the psychological impact on children.

ACKNOWLEDGMENTS

We would like to thank Claire Quigley for her assistance in coordinating the workshops at Mitchell Library and Bellcraig Community Centre in Glasgow. We also thank experts and families for their participation. 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] 2019. Little Kid Shows Me the Dark Side of VR Chat - YouTube. https://www.youtube.com/watch?v=DBV1pTzsevM Last Accessed 28-08-2023.

[2] 2022. Comfort and Safety — Rec Room. https://recroom.com/safety Last Accessed: 30-08-2022.

[3] 2022. Community Guidelines — VRChat. https://hello.vrchat.com/community-guidelines Last Accessed: 04-08-2023.

[4] 2022. QCAmap. https://www.qcamap.org/ui/en/home Last Accessed 05-09-2023.

[5] 2022. VRChat Safety and Trust System. https://docs.vrchat.com/docs/vrchat-safety-and-trust-system Last Accessed: 30-08-2022.

[6] 2023. FAQs — VRChatPlus. https://hello.vrchat.com/vrchat-plus-faq Last Accessed 05-09-2023.

[7] n.d.. Oculus Safety Centre | Oculus. https://www.oculus.com/safety-center/ Last Accessed: 22-07-2023.

[8] n.d.. A Parent's Guide to Rec Room — Rec Room. https://recroom.com/parents-guide Last Accessed: 22-07-2022.

[9] n.d.. Use of parental controls | ICO. https://ico.org.uk/for-organisations/childrens-code-hub/how-to-use-our-guidance-for-standard-one-best-interests-of-the-child/children-s-code-best-interests-framework/use-of-parental-controls/#recommendations Last Accessed: 09-08-2022.

[10] Melvin Abraham, Pejman Saeghe, Mark Mcgill, and Mohamed Khamis. 2022. Implications of XR on Privacy, Security and Behaviour: Insights from Experts. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) *(NordiCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 30, 12 pages. https://doi.org/10.1145/3546155.3546691

[11] Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2023. "Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 149 (apr 2023), 32 pages. https://doi.org/10.1145/3579625

[12] Sasha Alexdottir and Xiaosong Yang. 2022. Phantom Touch phenomenon as a manifestation of the Visual-Auditory-Tactile Synaesthesia and its impact on the users in virtual reality. In *Proceedings - 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR-Adjunct 2022*. Institute of Electrical and Electronics Engineers Inc., 727–732. https://doi.org/10.1109/ISMAR-ADJUNCT57072.2022.00218

[13] Karla A. Badillo-Urquiola, Diva Smriti, Brenna McNally, Evan Golub, Elizabeth M. Bonsignore, and Pamela J. Wisniewski. 2019. Stranger Danger!: Social Media App Features Co-designed with Children to Keep Them Safe Online. *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (2019). https://api.semanticscholar.org/CorpusID:174802028

[14] Jakki O. Bailey and Jeremy N. Bailenson. 2017. Considering virtual reality in children's lives. *http://dx.doi.org/10.1080/17482798.2016.1268779* 11 (1 2017), 107–113. Issue 1. https://doi.org/10.1080/17482798.2016.1268779

[15] Jakki O. Bailey, Jeremy N. Bailenson, Jelena Obradović, and Naomi R. Aguiar. 2019. Virtual reality's effect on children's inhibitory control, social compliance, and sharing. *Journal of Applied Developmental Psychology* 64 (2019), 101052. https://doi.org/10.1016/j.appdev.2019.101052

[16] Jakki O. Bailey and Isabella Schloss. 2023. "Awesomely Freaky!" The Impact of Type on Children's Social-Emotional Perceptions of Virtual Reality Characters. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 396, 10 pages. https://doi.org/10.1145/3544548.3581501

[17] Lawrence W Barsalou, Paula M Niedenthal, Aron K Barbey, and Jennifer A Ruppert. 2003. Social Embodiment. *THE PSYCHOLOGY OF LEARNING AND MOTIVATION Advances in Research and Theory* (2003).

[18] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 100 (nov 2019), 25 pages. https://doi.org/10.1145/3359202

[19] Lindsay Blackwell, Emma Gardiner, and Sarita Schoenebeck. 2016. Managing expectations: Technology tensions among parents and teens. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 27 (2 2016), 1390–1401. https://doi.org/10.1145/2818048.2819928

[20] Alex Bradley, Claire Lawrence, and Eamonn Ferguson. 2018. Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences* 285 (3 2018). Issue 1875. https://doi.org/10.1098/RSPB.2018.0116

[21] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (2006), 77–101. Issue 2. https://doi.org/10.1191/1478088706QP063OA

[22] Robyn Caplan and Tarleton Gillespie. 2020. Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society* 6, 2 (2020), 2056305120936636. https://doi.org/10.1177/2056305120936636 arXiv:https://doi.org/10.1177/2056305120936636

[23] Roser Cañigueral and Antonia F.de C. Hamilton. 2019. Being watched: Effects of an audience on eye gaze and prosocial behaviour. *Acta Psychologica* 195 (4 2019), 50–63. https://doi.org/10.1016/J.ACTPSY.2019.02.002

[24] Shenghui Cheng. 2023. *Metaverse*. Springer Nature Switzerland, Cham, 1–23. https://doi.org/10.1007/978-3-031-24359-2_1

[25] Ananta Chowdhury and Andrea Bunt. 2023. Co-Designing with Early Adolescents: Understanding Perceptions of and Design Considerations for Tech-Based Mediation Strategies that Promote Technology Disengagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 198, 16 pages. https://doi.org/10.1145/3544548.3581134

[26] Victoria Clarke and Virginia Braun. 2013. *Successful Qualitative Research: A Practical Guide for Beginners*. https://www.researchgate.net/publication/256089360_Successful_Qualitative_Research_A_Practical_Guide_for_Beginners

[27] Rebecca N.H. de Leeuw and Christa A. van der Laan. 2017. Helping behavior in Disney animated movies and children's helping behavior in the Netherlands. *https://doi.org/10.1080/17482798.2017.1409245* 12 (4 2017), 159–174. Issue 2. https://doi.org/10.1080/17482798.2017.1409245

[28] Mairi Therese Deighan, Amid Ayobi, and Aisling Ann O'Kane. 2023. Social Virtual Reality as a Mental Health Tool: How People Use VRChat to Support Social Connectedness and Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 100, 13 pages. https://doi.org/10.1145/3544548.3581103

[29] Catherine N. Dulmus, Karen M. Sowers, and Matthew T. Theriot. 2006. Prevalence and Bullying Experiences of Victims and Victims Who Become Bullies (Bully-Victims) at Rural Schools. *http://dx.doi.org/10.1080/15564880500498945* 1 (4 2006), 15–31. Issue 1. https://doi.org/10.1080/15564880500498945

[30] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[31] Daniel Falla, Rosario Ortega-Ruiz, Kevin Runions, and Eva M. Romera. 2022. Why do Victims become Perpetrators of Peer Bullying? Moral Disengagement in the Cycle of Violence. *Youth and Society* 54 (4 2022), 397–418. Issue 3. https://doi.org/10.1177/0044118X20973702/ASSET/IMAGES/LARGE/10.1177_0044118X20973702-FIG1.JPEG

[32] Cristina Fiani, Robin Bretin, Mark McGill, and Mohamed Khamis. 2023. 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)* (Hamburg, Germany). ACM, New York, NY, USA, 7. https://doi.org/10.1145/3544549.3585840

[33] Cristina Fiani, Robin Bretin, Mark Mcgill, and Mohamed Khamis. 2023. 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* (Chicago, IL, USA) *(IDC '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3585088.3589374

[34] Cristina Fiani and Stacy Marsella. 2022. Investigating the Non-verbal Behavior Features of Bullying for the Development of an Automatic Recognition System in Social Virtual Reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces* (Frascati, Rome, Italy) *(AVI 2022)*. Association for Computing Machinery, New York, NY, USA, Article 67, 3 pages. https://doi.org/10.1145/3531073.3534492

[35] Cristina Fiani, Mark Mcgill, and Mohamed Khamis. 2023. Ensuring Child Safety in Social VR: Navigating Different Perspectives and Merging Viewpoints. http://mkhamis.com/data/papers/fiani2023chiworkshop.pdf

[36] Cristina Fiani, Pejman Saeghe, Mark McGill, and Mohamed Khamis. 2024. 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, Article 54 (April 2024), 54 pages. https://doi.org/10.1145/3637331 25 pages.

[37] Guo Freeman and Dane Acena. 2021. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 84–95. https://doi.org/10.1145/3452918.3458805

[38] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 85 (apr 2022), 30 pages. https://doi.org/10.1145/3512932

[39] Radhika Garg and Subhasree Sengupta. 2020. Conversational Technologies for In-home Learning: Using Co-Design to Understand Children's and Parents' Perspectives. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, Hawaii) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376631

[40] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234. https://doi.org/10.1177/2053951720943234 arXiv:https://doi.org/10.1177/2053951720943234

[41] Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining 2022 12:1* 12 (9 2022), 1–41. Issue 1. https://doi.org/10.1007/S13278-022-00951-3

[42] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945. https://doi.org/10.1177/2053951719897945 arXiv:https://doi.org/10.1177/2053951719897945

[43] Patricia M. Greenfield. 2004. Developmental considerations for determining appropriate Internet use guidelines for children and adolescents. *Journal of Applied Developmental Psychology* 25, 6 (2004), 751–762. https://doi.org/10.1016/j.appdev.2004.09.008 Developing Children, Developing Media - Research from Television to the Internet from the Children's Digital Media Center: A Special Issue Dedicated to the Memory of Rodney R. Cocking.

[44] Sevtap Gurdal and Emma Sorbring. 2018. Children's agency in parent–child, teacher–pupil and peer relationship contexts. *International Journal of Qualitative Studies on Health and Well-being* 13, sup1 (2018), 1565239. https://doi.org/10.1080/17482631.2019.1565239 arXiv:https://doi.org/10.1080/17482631.2019.1565239 PMID: 30709328.

[45] Heidi Hartikainen, Netta Iivari, and Marianne Kinnula. 2016. 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* (Manchester, United Kingdom) *(IDC '16)*. Association for Computing Machinery, New York, NY, USA, 367–378. https://doi.org/10.1145/2930674.2930680

[46] Alexis Hiniker, Sarita Y. Schoenebeck, and Julie A. Kientz. 2016. Not at the dinner table: Parents' and children's perspectives on family technology rules. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 27 (2 2016), 1376–1389. https://doi.org/10.1145/2818048.2819940

[47] TTC Labs. Accessed on 2023-04-09. How we've used co-design to develop parental supervision tools at Meta. https://www.ttclabs.net/news/how-weve-used-co-design-to-develop-parental-supervision-tools-at-meta

[48] Jessica Lindblom. 2015. Embodiment and social interaction. *Cognitive Systems Monographs* 26 (2015), 115–159. https://doi.org/10.1007/978-3-319-20315-7_4/TABLES/1

[49] Shaun Alexander Macdonald, Euan Freeman, Stephen Brewster, and Frank Pollick. 2021. User Preferences for Calming Affective Haptic Stimuli in Social Settings. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 387–396. https://doi.org/10.1145/3462244.3479903

[50] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. It Is Complicated: Interacting with Children in Social Virtual Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Atlanta, GA, USA, 343–347. https://doi.org/10.1109/VRW50115.2020.00075

[51] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 472–483. https://doi.org/10.1145/3410404.3414268

[52] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (Athens, Greece) *(IDC '21)*. Association for Computing Machinery, New York, NY, USA, 69–79. https://doi.org/10.1145/3459990.3460703

[53] Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. 2018. Co-designing Mobile Online Safety Applications with Children. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3174097

[54] Joshua McVeigh-Schultz, Elena Márquez Segura, Nick Merrill, and Katherine Isbister. 2018. What's It Mean to "Be Social" in VR? Mapping the Social VR Design Ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (Hong Kong, China) *(DIS '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 289–294. https://doi.org/10.1145/3197391.3205451

[55] Jenifer Miehlbradt, Luigi F. Cuturi, Silvia Zanchi, Monica Gori, and Silvestro Micera. 2021. Immersive virtual reality interferes with default head–trunk coordination strategies in young children. *Scientific Reports* 11 (12 2021), 17959. Issue 1. https://doi.org/10.1038/s41598-021-96866-8

[56] Andy Miller, Eamonn Ferguson, and Rachel Simpson. 1998. The Perceived Effectiveness of Rewards and Sanctions in Primary Schools: adding in the parental perspective. *Educational Psychology* 18, 1 (1998), 55–64. https://doi.org/10.1080/0144341980180104 arXiv:https://doi.org/10.1080/0144341980180104

[57] Tijana Milosevic, Kanishk Verma, Michael Carter, Samantha Vigil, Derek Laffan, Brian Davis, and James O'Higgins Norman. 2023. Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From Youth Perspective. *Social Media + Society* 9, 1 (2023), 20563051221147325. https://doi.org/10.1177/20563051221147325 arXiv:https://doi.org/10.1177/20563051221147325

[58] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The uncanny valley. *IEEE Robotics and Automation Magazine* 19 (2012), 98–100. Issue 2. https://doi.org/10.1109/MRA.2012.2192811

[59] Marie Ozanne, Aparajita Bhandari, Natalya N Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* 9, 2 (2022), 20539517221115666. https://doi.org/10.1177/20539517221115666 arXiv:https://doi.org/10.1177/20539517221115666

[60] Joseph O'Hagan, Julie R. Williamson, Mark McGill, and Mohamed Khamis. 2021. Safety, Power Imbalances, Ethics and Proxy Sex: Surveying In-The-Wild Interactions Between VR Users and Bystanders. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Bari, Italy). IEEE, 211–220. https://doi.org/10.1109/ISMAR52148.2021.00036

[61] Jenny S. Radesky, Caroline Kistin, Staci Eisenberg, Jamie Gross, Gabrielle Block, Barry Zuckerman, and Michael Silverstein. 2016. Parent Perspectives on Their Mobile Technology Use: The Excitement and Exhaustion of Parenting while Connected. *Journal of Developmental and Behavioral Pediatrics* 37 (2016), 694–701. Issue 9. https://doi.org/10.1097/DBP.0000000000000357

[62] Jeffrey H. Reiman. 2017. *Driving to the panopticon: A philosophical exploration of the risks to Privacy Posed by the Highway Technology of the future.* Taylor and Francis, 159–176. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315246024-8/driving-panopticon-philosophical-exploration-risks-privacy-posed-highway-technology-future-jeffrey-reiman

[63] Nazanin Sabri, Bella Chen, Annabelle Teoh, Steven P. Dow, Kristen Vaccaro, and Mai Elsherief. 2023. Challenges of Moderating Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 384, 20 pages. https://doi.org/10.1145/3544548.3581329

[64] Elaheh Sanoubari, John Edison Muñoz Cardona, Hamza Mahdi, James E. Young, Andrew Houston, and Kerstin Dautenhahn. 2021. Robots, Bullies and Stories: A Remote Co-design Study with Children. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (Athens, Greece) *(IDC '21)*. Association for Computing Machinery, New York, NY, USA, 171–182. https://doi.org/10.1145/3459990.3460725

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

[66] Petr Slovak, Katie Salen, Stephanie Ta, and Geraldine Fitzpatrick. 2018. Mediating Conflicts in Minecraft: Empowering Learning in Online Multiplayer Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174169

[67] Jessica Van Brummelen, Maura Kelleher, Mingyan Claire Tian, and Nghi Nguyen. 2023. What Do Children and Parents Want and Perceive in Conversational Agents? Towards Transparent, Trustworthy, Democratized Agents. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago, IL, USA) *(IDC '23)*. Association for Computing Machinery, New York, NY, USA, 187–197. https://doi.org/10.1145/3585088.3589353

[68] Andreas Veglis. 2014. Moderation techniques for social media content. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8531 LNCS (2014), 137–148. https://doi.org/10.1007/978-3-319-07632-4_13/COVER

[69] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 146, 15 pages. https://doi.org/10.1145/3491102.3517614

[70] Howard Zehr. 2015. *The little book of restorative justice: Revised and updated.* Simon and Schuster.

[71] Qingxiao Zheng, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan C. Salvi, Guo Freeman, and Yun Huang. 2023. Understanding Safety Risks and Safety Design in Social VR Environments. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 154 (apr 2023), 37 pages. https://doi.org/10.1145/3579630

A EXPERTS DEMOGRAPHICS TABLES

P number	Age	Gender	Ethnicity	Parent?	Education	Profession	Profession involving children?	Country Profession
P02	33	Female	White / Caucasian	No	Doctorate (PhD, Information Science)	Researcher in social VR harassment	No	USA
P03	54	Female	White / Caucasian	No	Doctorate (Psychology)	Academic Professor of Psychology, Security & Trust, Criminology	No	UK
P04	50	Male	White / Caucasian	Yes	College / University degree (Sport and business management)	Head of Child Safety Online - Online safety profession involved in CSE related content inc VR	No	UK
P05	30	Female	White / Caucasian	No	University degree (Marketing and management)	Child Safety Intelligence Analyst/Consultant	Yes	UK
P06	62	Male	White / Caucasian	Yes	Professional degree (Masters in Cybercrime Investigations and Forensic Computing)	Trainer with Cybersafekids	Yes	Republic of Ireland
P07	48	Male	White / Caucasian	No	College / University degree (Social Psychology)	Child Safety Intelligence Analyst/Consultant (Social Media, Content Moderation, Trust & Safety, Online child sexual exploitation, Grooming, Sextortion, Self-Generated content, Suicide and Self-Harm, Generative AI, Metaverse)	Yes	Spain
P08	35	Female	White / Caucasian	No	College / University degree (Clinical psychology)	PhD student - Clinical Psychologist	No	Italy
P09	48	Female	White / Caucasian	Yes	College / University degree Linguistics and AI (and an MBA)	Online Safety Consultant Clinical Psychologist Online safety consultant (7 years online safety consulting)	Yes	UK
P10	42	Female	White / Caucasian	No	Doctorate (PhD)	Head of online safety research	Yes	Malta / Worldwide
P11	42	Non-binary / third gender	Other	No	Doctorate (Psy.D.)	Community Developer	No	USA
P12	59	Female	Black / African	Yes	Doctorate (PhD)	Professor of Child and Adolescent Psychiatry	Yes	UK
P13	27	Female	Asian	No	College / University degree (MSc Comparative Social Policy)	Policy and Public Affairs Officer	Yes	UK
P14	42	Male	White / Caucasian	No	College / University degree (BA Italian)	Director, Online Safety Child	Yes	UK
P15	46	Male	White / Caucasian	Yes	Professional degree (MA, PGCE:PSE, MBChB, MRCPsych)	Clinical Senior Lecturer and Honorary Consultant Child and Adolescent Psychiatrist	Yes	UK
P16	49	Female	White / Caucasian	Yes	Other: MSc and PhD candidate PhD (current) and MSc in Psychotherapy	Psychotherapist and cybertrauma Consultant	Yes	UK
P17	25	Female	White / Caucasian	No	Professional degree (Data & Society)	Digital Investigator Social VR threat researcher, Threat Intel analyst at social media companies	No	UK

Table 3: Demographics of experts who participated in one-to-one interviews: age, gender, ethnicity, parent, education, profession, profession involving children and country of profession. Experts had a Mean age of 43.3 years (σ=10.5).

Participant	Child harassment/bullying Knowledgeable	Child Cyberbullying Knowledgeable	Social VR Awareness (1-5 Likert scale)	Social VR Experience	Owner of VR headset	VR Experience
P02	Extremely	Extremely	5	A lot	Yes. (Oculus / Meta brand)	A lot
P03	Moderately	Moderately	5	A moderate amount	Yes. (Oculus / Meta brand)	A moderate amount
P04	Very	Very	5	A little	No	A little
P05	Very	Extremely	5	A little	No	A little
P06	Very	Very	4	None at all	No	A little
P07	Very	Very	4	A moderate amount	Yes. (Oculus / Meta brand)	A moderate amount
P08	Not at all	Not at all	1	None at all	No	A little
P09	Moderately	Very	5	A little	Yes. (Oculus / Meta brand)	A little
P10	Moderately	Very	2	None at all	No	A little
P11	Moderately	Moderately	5	A great deal	Yes. (Oculus / Meta brand)	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. (Oculus / Meta brand)	A lot
P17	Extremely	Moderately	5	A great deal	Yes. (Oculus / Meta brand)	A lot

Table 4: 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).

B FAMILIES DEMOGRAPHICS TABLES

We used a format where family's IDs were represented for guardians as P(Participant Number)(P or G)-(Workshop Number) (e.g., P3P-2) with corresponding child as P(Participant Number letter)(C)-(Workshop Number), a letter is added if more than one child per guardian came (e.g., P3aC-2, P3bC-2).

P number	Guardian	Age	Gender	Ethnicity	Education	Profession	Owner of VR headsets	VR Experience	Social VR Awareness (1-5 scale)	Social VR Experience	Attitude towards social VR (1-5 scale)
P1P-1	Parent	58	Male	White / Caucasian	Professional degree	Employed full time	No.	None at all	2	None at all	4
P2P-1	Parent	39	Female	Black / African	Professional degree	Student	No.	None at all	1	A little	2
P1P-2	Parent	37	Female	Black / African	Professional degree	Unemployed	No.	A little	1	A little	3
P2G-2	Grandparent	Unknown	Male	White / Caucasian	Prefer not to say	Retired	No.	None at all	1	None at all	3
P3P-2	Parent	44	Female	White / Caucasian	2 year degree	Employed full time	No.	A little	4	A little	1
P1G-3	Grandparent	65	Female	White / Caucasian	Some college	Retired	Yes. Oculus Quest 2	None at all	1	None at all	2
P1G-4	Grandparent	61	Female	White / Caucasian	Prefer not to say	Prefer not to say	Yes. Meta Quest 2	A little	4	None at all	4
P2P-4	Parent	49	Female	White / Caucasian	4 year degree	Other: Housewife	No.	A little	1	None at all	2

Table 5: Demographics of parents and grandparents who participated in the workshops. Parents had a mean age of 45.5 (σ=8.4). Grandparents had a mean age of 63.5 (σ=2.8). Guardians had a mean Social VR awareness of 1.9 (σ=1.4), while regarding Attitudes towards social VR, they scored a mean of 2.6 (σ=1.1).

P number	Age	Gender	Ethnicity	VR Experience	Social VR Experience
P1C-1	11	Female	White / Caucasian	unknown	unknown
P2Cb-1, P2Ca-1	12, 10	Female, Female	Black / African, Black / African	A little, None at all	None at all, None at all
P1C-2	8	Female	Black / African	None at all	None at all
P2C-2	8	Male	Other: African/scottish	A moderate amount	A little
P3Cb-2, P3Ca-2	11, 8	Male, Female	White / Caucasian, White/Caucasian	A little, A little	None at all, None at all
P1C-3	10	Male	White / Caucasian	A lot	A little
P2C-4	13	Male	White / Caucasian	A little	None at all
P2P-4	15	Male	White / Caucasian	A great deal	A moderate amount
P1	16	Male	Black / African	A little	None at all
P2	14	Female	White / Caucasian	A little	None at all
P3	13	Female	Prefer not to say	A little	None at all

Table 6: Demographics of Children who participated in the workshops. Children had a mean age of 11.5 (σ=2.7). These were reported by the guardians.