Big Buddy: Exploring Child Reactions and Parental Perceptions towards a Simulated Embodied Moderating System for 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

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

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

ABSTRACT

Children experience new forms of harassment in Social Virtual Reality (VR), often inaccessible to parental oversight. We aimed to understand how an artificial intelligent moderator safeguarding children from harassment in social VR is perceived by children and parents, by introducing “Big Buddy”, a Wizard-of-Oz embodied AI-moderator. 43 children (aged 8-16) played a tower-block-construction game in a simulated Social VR classroom where fictitious competitors disrupted their game and, in experimental conditions where present, Big Buddy intervened. We measured children’s perceptions after the disruptions, towards Big Buddy, and the moderation actions it took. Children felt significantly less sad and safer when Big Buddy suspended the saboteur. Parents (n=17) noted Big Buddy’s usefulness and felt reassured but would remain in the supervision loop. We present the first empirical research of a VR-embodied AI-moderator with children’s and parents’ perspectives, and propose design directions for embodied AI-moderators in Social VR.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

social virtual reality, children, parents, online harassment, embodied moderating system, automated moderator

ACM Reference Format:
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 *Interaction Design and Children (IDC ’23), June 19–23, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3585088.3589374

IDC ’23, June 19–23, 2023, Chicago, IL, USA

1 INTRODUCTION

Social Virtual Reality (VR) is a simulated social environment, initially designed for adults and older teenagers, that has attracted a large amount of minors under 13 years old [43–45]. Social VR has the potential to mimic true face-to-face interactions with social presence, allowing users to interact via 3D avatars in virtual environments through head-mounted devices [44]. While social VR can create an innovative way of engaging with others due to unique embodiment and the illusion of "being there" [47], this also opens the door to negative traumatic experiences that mirror harassment and bullying in reality [11, 45, 59]. For example, children and adults have reported harassment ranging from name calling to physical stalking and sexual harassment [43, 45].

Unfortunately existing mitigation features, such as such as blocking, personal space bubbles, muting, reporting players [2, 4] or trust systems to keep users safe from nuisance users [5], suffer from significant limitations. First, they place the responsibility of their use on potentially ill-equipped users including children or guardians (e.g., unfamiliar with the technology or not knowing the best approaches) [12, 31, 34, 36]. Second, the consequences are unclear to the abuser and bystander (e.g., parent), creating the perception of environments without consequences that might dissuade negative behaviours. Third, they do not allow remote parental oversight nor inform parents of their children’s negative experiences as bullies or as victims of bullying. There is a growing need to understand the effectiveness of safety-enhancing technologies for children in social VR as well as child and parental perceptions towards mitigation tools. Research has shown that human-based moderators can help establish norms for appropriate behaviours [11]. However, it was also shown that users would lack trust due to possible personal and subjective biases of the moderators and the worry that they may be effective only in small-scale social VR environments [28]. As research has shown the effectiveness of automated moderation approaches in forums or games (e.g., AutoModerator Bot [1]) [13, 20, 39], automated moderation approaches in social VR could therefore be a solution that could address the challenges described above. A recent study [56], drawing on 39 interviews with adult social VR users, investigated opportunities and limitations of AI-based moderation and provided insight into its potential in creating a safer and more inclusive social VR environment. However, we do not have an understanding of how they should be integrated in the social VR environment from children's and parents' perspectives.

This work addresses the gap in understanding how embodied AI-moderators should be integrated into a social VR environment from children's and parents' perspectives. To explore the effectiveness and perception of automated embodied moderators by children and parents, we introduced an AI-moderator named "*Big Buddy*" and employed the Wizard-of-Oz (WOZ) prototyping approach in a simulated social VR environment. The insights gained from this study can inform the design of future AI-moderators.

We explored the impact of the presence of the embodied AI-moderator, and the visibility of its interventions on child and parental perceptions and sense of safety of the moderation experience. 43 children (8-16 years old) played a researcher-developed VR tower-block-construction game. It involved fictitious competitors disrupting the participant's game in a virtual classroom mimicking a social VR environment. Parents watched a video of what the child was seeing under the VR headset before semi-structured interviews. We evaluated children's emotions and perceptions towards Big Buddy and his interventions as well as parents' perceptions. In this paper, we aimed to answer the following questions:

RQ1 How is children's emotional valence impacted in provocative situations by the presence of Big Buddy?

RQ2 How do children and parents perceive the embodiment of the moderating system (Big Buddy) and moderation actions (punishments)? Will children feel safer and/or inhibited given its presence?

RQ3 How do children and parents envision an AI-moderator in social VR?

RQ4 How involved would parents want to be with the embodied moderating system?

This paper makes a number of contributions to child-computer interaction. Firstly, we present the first experimental research of a VR-constructed WOZ embodied moderator that aims to mitigate the saboteur's actions, with the anticipated benefit of making children feel safer and more comfortable in the simulated social virtual environment through visible safeguarding interventions. We get first impressions and perceptions from both children and parents. Secondly, we gathered the first data about design features an AI-moderator would need to not only safeguard children but also be perceived positively and reassuringly to children and parents without completely removing children's sense of agency and enjoyment or overly restricting their freedom. Based on our findings, we propose future design directions for an effective AI-moderator to enhance safety and allow parental oversight in social VR.

2 RELATED WORK

2.1 Child Perception Towards Sanctions and Unfairness in Social Disruptive Situations

Children can encounter social disruptive situations in physical reality (e.g., at school) or online [60]. As children begin to communicate with others in school, they learn desirable and undesirable behaviours in social settings [38]. Children want to be listened to and supported by someone who would notice their pain [38]. Recent studies evaluated children's perception towards rewards, sanctions and unfairness for different psychological and educational applications such as reinforcement learning, effective parenting, discipline and social norms for appropriate behaviour [35, 37, 48, 60]. The design and methodologies of the latter studies are of interest to our experimental design and interventions.

2.1.1 Children's Emotional Reaction and Self-Reflection in Simulated Provocative Situations. A recent study evaluated children's aggressive behaviour, in particular aggressive social information processing (SIP) [60]. Researchers evaluated how emotionally engaged children were in disruptive situations via interactive and realistic Virtual Reality (VR) scenarios. Boys of age 8 to 13 years old (N=32) (from regular and special education) were individually tested in a silent room. The VR games included building a tower of blocks with six scenarios: starting with a practice, one neutral scenario where no engaging event occurred, two instrumental gain scenarios (participants could choose to steal a block or ball from the virtual peer to obtain additional points and participant could win the game by sabotaging the virtual peer's game) and two provocative situations (participants were refused to join a game by two virtual peers and participants' game was ruined by the virtual peer). Results showed that peer provocation led to more anger, hostile intent attributions and revenge goals than the instrumental gain and neutral contexts, and more aggressive responses than the neutral context. Our VR game and provocative situation are inspired by this study as it was shown that the game was challenging enough but not too difficult for children aged 8-13 and the disruptive situation was designed to provoke an emotional response.

Anger and frustration are negative emotions according to Russell's complex [55] that can be elicited in tasks such as toy removal and were shown to orient children towards desirable goals or objects [35]. The latter study evaluated the influence of children's (N = 40, 5-6 years old) anger under reward and punishment conditions. Emotions were self-reported by children using the established Self-Assessment Manikin (SAM) [40]. The first experiment manipulated a situation where children met obstacles (unfair treatment from the competitor) in the pursuit of a desirable goal. Each participant would compete with an unfamiliar player. The game was made so that the participant would always lose. Participants were then asked to rate how they were feeling using SAM scale. The experiment showed that anger is associated with attention biases towards rewards rather than punishment.

2.1.2 Perceptions and Influence of Sanctions on Children and Parents. Reinforcement learning has been widely used by teachers in education to maintain discipline in the classroom [48]. Perceptions of punishments and rewards for pupils' behaviour have been investigated looking at the relative effectiveness of school-initiated rewards and punishments as perceived by children in primary school and parents via a survey (N_children = 49, N_parents = 64). Regarding punishments, children rated '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 [48]. We base our interventions on this rank.

Prosocial behaviour of children is largely influenced by adult figures, authority and media. Several psychological studies looked at the effects of being watched, monitored and agency in parent-child and pupil-teacher relationships [32]. Children make a distinction in their perception of agency depending on the relationship context. A study showed children perceived the least agency with teachers and the most agency with peers [32]. Another study showed Disney characters can inspire children to help others. Children (N = 113, 8-10 years old) were divided into two groups, one was exposed to a Disney clip where the character was helping and the other group was exposed to a Disney clip without helping behaviour. It was shown that children were more likely to help their friends after watching the animated clip with the helping character [21]. Another study investigated children's conception of authority from an individual, showing it would depend on their status, the context and the domain of act depending on children's age [58]. Younger children's awareness of mothers and teachers authority was shown to be greater than that of police for instance [14, 58]. Therefore, there is the potential need of having a visible embodied moderator as an authority figure for children in SVR.

2.2 Bullying and Cyberbullying Prevention and Intervention

The most effective interventions to tackle bullying and victimisation appeared to be disciplinary sanctions by communicating to the bully that the behaviour was unacceptable and reporting the event to other adults [17]. As a second most effective intervention, they showed that teacher-facilitated group discussions had beneficial effect in becoming a defender. An increase in non-intervention showed unfavourable effects increasing the likelihood of being a victim and decreasing the likelihood of being a defender [17]. Teachers therefore have a major influence towards student behaviours. Moreover, non-interventions can reinforce the message that victimising others has no negative consequences and bullies would not recognise their behaviours as unacceptable. However, supporting the victim had no direct beneficial effects and would not allow to communicate clear boundaries [17]. While most anti-bullying interventions involve teacher supervision and consequences, a study [25] suggests that youth mentoring interventions (e.g., Big Brothers/Big Sisters of America) and support groups have beneficial effects on peer relationships. In a recent study, bullied children in elementary school (N=12) were paired with mentors who visited the school twice a week. It was shown that children paired with a mentor experienced less peer victimisation. Within classrooms, having classmates who showed more peer support and who were older in general than children in other classrooms, reduced the risk of being a victim of bullying [23]. Indeed, it was shown that peers act as a group to influence the development of individual children [22].

Cyberbullying (i.e., bullying enacted using technology) [9] can lead to anger and loneliness, negative emotions that are often undisclosed by victims and intensified by passive bystanders' presence. Yet, observers' reactions is critical as they may influence those involved and the occurrence of the events [9]. They can change the course of situation and reduce the negative effects on the victim by confronting bullies and reporting to adults [10]. Research proposes different strategies to reduce adolescents cyberbullying on social media including 1) active parental involvement and monitoring of social media use, 2) training bystanders to intervene and, 3) educating about online safety [30]. Another solution involved reflective messages to dissuade someone from posting content detected as cyberbullying. Royen and her colleagues [54] found that such reflective messages can reduce the intention to engage in cyberbullying. Automated methods to detect cyberbullying have been used in social media, including via text classification to detect harassment keywords and Natural Language Processing (NLP) and have been shown to identify hate speech with high accuracy [8, 42, 53]. Automated detection could feasibly allow to detect harassment events in social VR in the future. We need to consider how we would integrate it in social VR, how it would intervene and how we would make the detection and intervention capability visible to improve the user experience.

2.3 Bullying and Harassment in Social VR

Experiences of harassment and bullying in social VR have increased and are shown to be more intense than bullying on social media sites due to the embodiment and presence in VR [11]. Children and adults have reported harassment, from name calling to physical stalking [43, 45]. While these issues raise concerns, current mediation tools (e.g., reporting and blocking) are shown to have flaws in the process and lack trust and feelings of unfair treatment discouraging users to use them [43]. Interventions to protect children from bullying are lacking and mitigation options that exist in digital media (e.g., Microsoft Family Safety, Apple Families, Google Family Group) now have (largely) yet to be transposed to social VR [7, 18, 19, 62]. It is difficult for parents to act as a bystander and intervene as social VR requires head-worn devices that completely occlude reality, and do not support bystander awareness that could allow effective supervision [50]. Research has shown that human moderators can help establish norms for appropriate behaviours [11], with nuanced views towards human-based moderators [28] as users could lack trust due to possible personal and subjective biases of the moderators and the worry that they may be effective only in small-scale social VR environments [28]. In our study, we introduce a simulated embodied AI-moderating system (WOZ prototype) that would potentially alleviate some of the concerns around human moderators in social VR environments.

This paper builds upon our previously published late-breaking work paper [26] broadening the scope to include not only children's perspectives but also parents' viewpoints.

2.4 Summary

Our study design is based on multiple prior studies to ensure the realism of the VR game and create a socially disruptive situation that is ethical but still evokes an emotional response. Our VR game and provocative scenario are based on [60], and measures of emotional reactions in studies with children were taken using SAM scales [37]. Furthermore, there is a class of interventions proven to work in schools and to reduce cyberbullying in social media [17, 23, 48, 54], from which we base our interventions. Additionally, the presence of visible authority figures and parental involvement can affect the effectiveness of interventions [9, 14, 21, 32, 58]. Given the feasibility of automatically detecting harassment in social media [8, 42, 53], and as new forms of harassment in social VR have become a growing concern [11, 28, 43, 45], it is important to investigate suitable interventions. Our proposed approach is to investigate child and parental perceptions of an embodied AI-moderator system to mitigate harassment towards children, with notable interventions in a simulated social VR environment and disruptive situations.

3 METHOD

Our study focuses on the design and perceptions of an embodied AI-moderation for social VR, evaluated in a social VR gaming experience with children and parents. We simulated a social VR environment game with disruptive situations based on prior research [60] and implemented a WOZ embodied AI-moderator, Big Buddy, who put in place interventions when a disruption occurred based on the class of punishments shown to be effective in prior work [17, 23, 48, 54]. We measured child reactions using SAM scales [37], and designed Likert scales and interview questions to measure child and parental perceptions towards Big Buddy. The protocol was approved by our ethics committee.

3.1 Procedure

Children (one or two participants per session) were accompanied by their parent or by their teacher and were welcomed in an empty room. While the children played the VR game, the parent waited at the other side of the room separated with a screen where they watched a video recording of what the children would be experiencing in the VR game before interviews. Before playing the game, the children were presented slides showing what they would be doing. They were told they would be in a virtual classroom competing with real players from all around the world. To verify their understanding of the three SAM scales, they were asked how they were feeling in that moment and to give a score: from sad (1) to happy (5) (M = 4.4, SD = 0.7), calm (1) to angry (5) (M = 1.4, SD = 0.5) and scared/intimidated (1) to safe (5) (M = 4.5, SD = 0.7). Children used an Oculus Quest 2 headset with an adjustable strap for more comfort. They were standing, they did not need to move around the room. They were immersed in a virtual classroom with non-player virtual characters (e.g., peers sitting or talking in the background) to mimic a social VR environment. The game consisted of a practice session and 4 rounds where they competed with each of the 4 fictitious players. The goal was to build a tower of 5 blocks as quickly as possible and ensure it remains stable until the time limit ends. Points were awarded to players who successfully met this goal. If they had N blocks stacked together they would earn N points. They started with a practice session where they could build the tower, and could practise until they felt confident. The researcher checked they had managed to get at least 4 points (4 blocks stacked) before starting the competition as the sabotage would mainly affect the tower-building. Overall, based on on 4-point scales (0 'a little' to 3 'extremely'), children found the VR game a little easy (M = 1.6, SD = 0.80), requiring a little effort (M = 1.1, SD = 0.65) and the headset was very comfortable (M = 2.3, SD = 0.9). After each round, they had to fill in a form on a laptop, with the questions described in the section 3.4. At the end of the game, each child and parent were individually interviewed.

3.2 Experimental Design

The study follows a Within-Participant design with Big Buddy's interventions as the independent variable of 4 levels:

(1) ***C1 [No BB]***- Big Buddy is absent, no intervention is taken.

(2) ***C2 [BB, Reset Points]***- Big Buddy is present and intervenes: resets saboteur's points back to 0.

(3) ***C3 [BB, Reset+Notify Parents]***- Big Buddy is present and intervenes: resets saboteur's points back to 0 and notifies saboteur's parents.

(4) ***C4 [BB, Reset+Notify Parents+Exclusion]***- Big Buddy is present and intervenes: reset saboteur's points back to 0, notifies saboteur's parents and saboteur is excluded from the game.

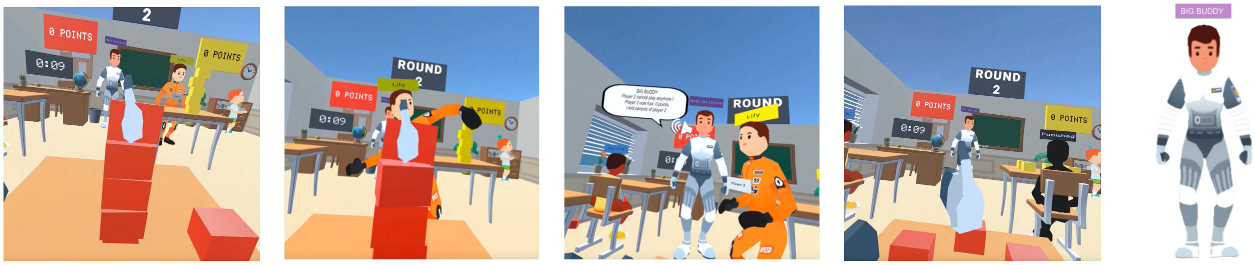
Each level represents a mix of punishments that increases in intensity, from no punishment to highly restricting actions. Big buddy announced verbally these punishments (via generated audios) and punishments were written in a bubble (see Figure 1 c), ensuring that the child participant is aware of the actions taken. The punishments were based on the top three of the perceived effective sanctions in a school according to children and parents [48]. The four conditions were counter-balanced with Latin Square. However, we have the following randomisation: [C2,C3,C4,C1; N = 10], [C3,C2,C1,C4; N = 10], [C4,C1,C2,C3; N = 13], [C1,C4,C3,C2; N = 10] (as 5 participants were removed, see section 3.6). The steps of the experiment and conditions can be seen in Appendix B. The provocative situation was the same throughout the game (i.e., destroying the participant's tower). Therefore, we randomised the number of games (2 to 4) for each round and the moment when the competitor would sabotage the game, randomising the fairness within each of the four conditions (fair win/loss, unfair loss) to reduce the effect of repetitiveness of sabotaging.

3.3 Social VR Game and Avatars Implementation

The simulated social VR game used to conduct the experiment was developed using Unity 3D. Its design was based on the game used to evaluate the aggressive behaviour in boys in prior work [60]. Based on the latter study, we simulated provocative situations where a virtual character sabotages the participant's game, by destroying their built tower. We constructed avatars for the Big Buddy and the 4 competitors using the Ubiq library [3] with animations and voice generator. To stand out from the other avatars, Big Buddy was designed to be bigger and with a noticeable appearance (See Figure 1 e). We designed it such that it leaned towards a robotic artificial moderator using a monotone agent voice to avoid adding variables (e.g., different voice tones and intonations etc.) using available visual looks from the Ubiq library. The game was pilot tested with 8 adults before starting the experiment with children. A sample scenario in VR (from the participants' perspective) that occurs in one of the rounds for C3 is shown in Figure 1.

3.4 Measures

Quantitative: *After the practice session*, children were asked to give a score on 4-point scales (0 'a little' to 3 'extremely') for ease of use (standard SUS questionnaire item [16]), effort (standard NASA-TLX item [33]) and comfort and their performance by giving the maximum points they got. *At the end of each round of the game*,


2
0 POINTS
0:09
0 POINTS
ROUND 2
Lily
BIG BUDDY: Player 2 cannot play anymore ! Player 2 now has: 0 points. I told parents of player 2.
BIG BUDDY
Punished

(a) (b) (c) (d) (e)

Figure 1: Example of the VR game scenario occurring in round 2 with C4 [BB, Reset+Notify Parents+Exclusion] from the user's eyes. (a) Tower Building Game. (b) Provocative situation. (c) Big Buddy intervenes: points reset to 0, parents notified and exclusion. (d) Competitor appears punished. (e) Big Buddy Close-Up.

children were asked to self-rate the emotions felt when the other player destroyed their tower, using the three 5-point SAM Likert scales [40], re-adapting the third one (from scared/intimidated to safe) (see Appendix A). If Big Buddy was present in the round, they had to give a score on three 4-point scales (from 'a little' to 'extremely') focusing on the sense of agency: 1) Seeing Big Buddy when playing, 2) Feeling seen by Big Buddy, 3) Big Buddy helped with a fair punishment) and select among choices, the corresponding punishments that were put in place. *At the end of the game, during interviews,* children and parents were separately asked to give scores for preferred physical and social attributes, using 5-point scales from -2 to 2 (e.g., authoritarian to liberal, non-humanised to humanised, visible to invisible, amateur to expert). The latter scales were designed based on items used to design a tutor social robot [51], to gain a better understanding of what children and parents would want as a preferred embodied moderating system in social VR (See Appendix C). **Qualitative:** *At the end of the game, during interviews* children and parents were separately asked questions around their perceptions of the bully and Big Buddy, their sense of safety, and potential customisation of Big Buddy (see full questions in Appendix D). For example, children were asked to describe Big Buddy's role, how they felt when Big Buddy was present compared to when absent, and their perceptions of the different punishments. Parents were also asked about how they felt when Big Buddy was present, their perceptions on the punishments and how involved they would want to be given such a system.

3.5 Analysis

Quantitative: We analysed our quantitative data using R statistical tools. In particular using ARTool for Aligned ranks transformation (ART) ANOVA, a non-parametric approach for multiple independent variables, interactions, and repeated measures [24, 61]. We conducted one-way ART ANOVAs (significant level 0.05) and post-hoc comparisons with Tukey adjustment with 1) SAM score as the dependent variable and conditions as the independent variable and 2) scores for Big Buddy scales (seeing Big Buddy/feeling seen/fairness of punishment) as the dependent variable and conditions as the independent variable. We also used K-means clustering analysis to analyse preferred physical and social attributes of Big Buddy [6]. The analysis and codes can be found: (link to be added; see supplementary material for review). **Qualitative:** We used inductive thematic analysis techniques [15, 50] to analyse parents' and children's perceptions from audio recorded interviews. We did not seek inter-rater reliability because researchers may interpret the meaning of codes differently [41]. After generating transcripts, the experimenter listened to all recordings (~5min each) and corrected any mistakes in the automatically generated transcripts. A pair of researchers read and familiarised themselves with the data. The two researchers then created individual coding schemes independently using NVivo, line by line. The codes generated are words or short phrases that describe an idea. Then we collaborated to consolidate the two coding schemes into one combine scheme, by collating or distinguishing between codes. We created a set of higher-level codes by bringing together related codes. All authors collaborated in an iterative process to discuss, combine, and refine themes and features to generate a rich description.

3.6 Participants

3.6.1 *Recruitment.*

The children (8-16 years old) were recruited via their parents on a voluntary basis and required the parent or legal guardian's permission to participate. The consent form and demographics questions were completed by the children's parent/legal guardian before booking a slot. The first batch (N_children = 20, N_parents = 14) was found via the university forums and mailing lists. The second batch (N_children = 28, N_parents = 3) was from a school where the lead researcher spent two days. This required permission from the headteacher and the parents filling in the same form with the information sheet, consent form and demographics question but were not required to be present. The researcher asked for a verbal agreement from children to participate and informed them that they could take breaks or stop at any point. Due to technical issues and as two children did not want to play anymore, we had to omit in total 5 child participants from the data analysis. The total number of participants is therefore: **43 children and 17 parents.** An £8 Amazon-voucher and commute costs were compensated to adults for their time and a token of appreciation (sticker/keychain) was given to the children.

3.6.2 *Demographics.*

The participants' age, awareness and frequency of use of general VR, social VR and online games (6-point

Likert Scale; 0=never; 5=daily) are summarised in Table 1. **Children:** Among the 43 child participants, 20 are female, 21 are male, 1 is non-binary/third gender and 2 preferred not to say. Regarding ethnicity, there are 29 White Caucasian, 2 Asian, 4 Arab, 1 Hispanic/Latino, 1 Mixed (Scottish-North African), 4 Other and 2 prefer not to say. 40 children were interviewed at the end of the game. Children were from 31 families. **Parents:** 18 mothers, 10 fathers and 2 parents that did not disclose their gender responded to the form. There were 22 White Caucasian, 3 Arabs, 2 Asian, 1 Black/African, 1 Hispanic/Latino and 2 prefer not to say. Among the 31 parents, there were 17 who accepted to participate in the interview (12 female, 4 male and 1 undisclosed gender; 3 Arab, 2 Asian, 1 Black/African, 1 Hispanic/Latino and 10 White Caucasian).

Groups	**All Parents**	**Parents Interviewed**	**Children**
N	31	17	43
Age (in years)	42.37 [5.20]	42.35 [5.13]	11.49 [2.09]
Social VR Awareness (0-5 Likert scale ’not at all’-’extremely’)	3.63 [1.10]	3.41 [1.12]	NA
VR Use Frequency (0-5 Likert scale ’never’-’daily’)	0.33 [0.55]	0.35 [0.49]	0.68 [1.04]
Social VR Use Frequency (0-5 Likert scale ’never’-’daily’)	0.27 [0.78]	0.41 [1.00]	0.40 [1.08]
Online Games Use Frequency (0-5 Likert scale ’never’-’daily’)	NA	NA	3.54 [1.87]

Table 1: Table summarising demographics of parents, parents who were interviewed and children. (N_parents = 31, N_parents(interviewed) = 17, N_children = 43). MEAN [SD].

3.7 Limitations

Demographics: Participants were mostly White Caucasian and among the parents, those who were interviewed were mostly female. The study was done in English in the United Kingdom which may introduce cultural bias. **Locations:** The experiment was carried out in two different locations, the first batch of children was accompanied by parents whereas the second one was not and was in a school environment. However, in both cases, children were separated from the teacher or parent to reduce possible influence. In some sessions, the participants were in pairs playing the VR game at the same time (parent with two children or in the school to avoid taking some of their valuable time), which may have influenced their responses. Nevertheless, the researcher ensured they could not talk to each other and they answered their questions quietly on a separate device. **Counter-balancing:** Due to omitting 5 participants, the Latin Square randomisation of the four conditions is imbalanced. As the study is within-subjects, there are some limitations as they might carry on valence effects (e.g., participants angry from C1 when indicating their anger levels after C2) but this is more ecologically valid as such a situation could occur in social VR and we counter-balanced the win-loss of the games (fair win or loss and unfair loss) as well as the number of games per round to reduce this effect. **Surprise Effect:** While participants may have expressed that they were very surprised in the first round but felt less surprised from the third round when the other player destroyed their tower, it is not the surprise or expectation of the attack that changed their perception of lack of fairness and need of interventions.

4 RESULTS

4.1 Children’s Emotional Reaction after the Disruptive Situation

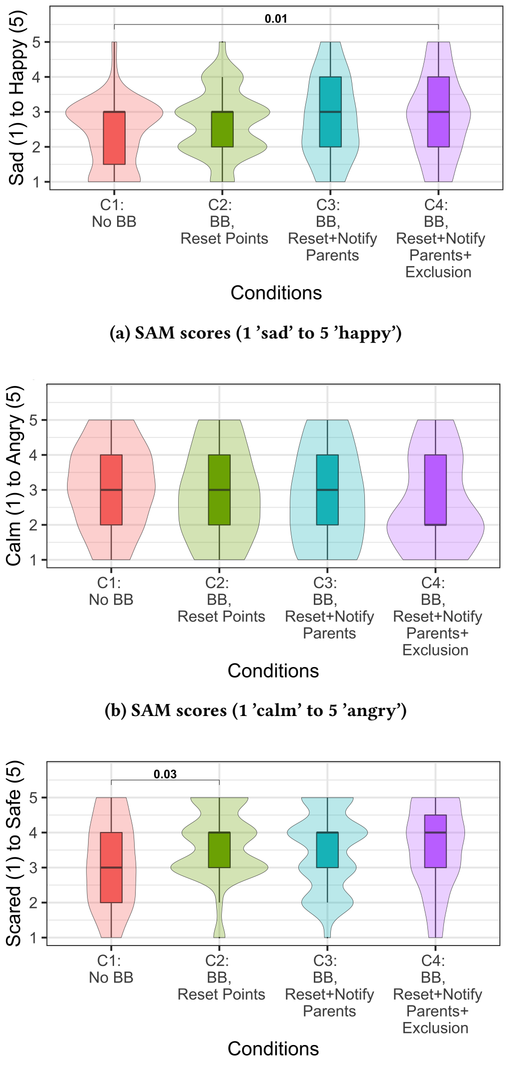
Overall, children felt more negative valence emotions (sadness) and felt less safe when the other player destroyed their tower and when Big Buddy was absent compared to when Big Buddy was present (see Figure 2). However, in terms of arousal, children’s anger was relatively the same throughout all the rounds which shows that even if the disruption was repeated, their level of anger did not particularly change. Significant results (significant p-values set at 0.05) with medium (between 0.06 and 0.14) and large effect sizes (0.14 or higher) partial eta squared η_p^2 [49, 52], were obtained for the different conditions for sadness (F(3,117)=3.55, p=0.02, η_p^2=.08, medium effect size, ART ANOVA) and for safety (F(3,117)=3.22, p=0.03, η_p^2=.08, medium effect size, ART ANOVA). In particular, children felt significantly sadder (p=0.01) in the round with C1 [No BB] compared to the round with C4 [BB, Reset+Notify Parents+Exclusion]. Moreover, children felt significantly safer (p=0.03) in C2 [BB, Reset Points] than in C1 [No BB].

4.2 Perceptions towards Big Buddy (Embodiment of Moderating System)

4.2.1 Children’s Perceptions towards Big Buddy. From self-ratings (sense of agency), children felt they were ’seeing Big Buddy’ while playing similarly in all conditions where Big Buddy was present. Nevertheless, there was a significant difference of scores across conditions when answering if they ’felt seen by Big Buddy’ when playing (F(2,84)=5.58, p=0.005, η_p^2=.12, medium effect size, ART ANOVA). More specifically, they felt that Big Buddy could see them significantly more in C4 [BB, Reset+Notify Parents+Exclusion] than C2 [BB, Reset Points](p=0.0037) (see Figure 3).

Big Buddy’s Role. Seven described him as being like a teacher, seven described his role as keeping the game fair, seven mentioned he is there to keep them safe, six said he is there to punish and 14 described him as someone to help regulate other users’ actions: *“His presence helped to regulate what the people did”* [SchoolChild17]. *“I think it was to help if any of the other children were mean then it meant that they would get the punishment that they deserved”* [Child22a].

Impact of Big Buddy’s Presence VS Absence. Children felt that Big Buddy’s presence was reassuring and they felt safer (n=10): *“I felt much safer when he was there and a lot more protected”* [Child32a]. However, two children had mixed feelings, one of them mentioned it was comforting but strange at the same time [SchoolChild14]. Children also linked Big Buddy’s presence with the interventions and moderation, they knew interventions would only occur if Big Buddy is shown in the scene (n=5): *“Well when he was present he made sure that if they did something wrong they would get in trouble. When he wasn’t there nothing happened”* [Child28a]; *“I didn’t like it when he wasn’t there because they’d knocked down my tower and nothing happened”* [Child25b]. Eight children were annoyed by the absence of Big Buddy due to no punishments. Twenty-five children responded that Big Buddy’s presence would make them feel safer in other bullying situations (e.g., name calling or someone



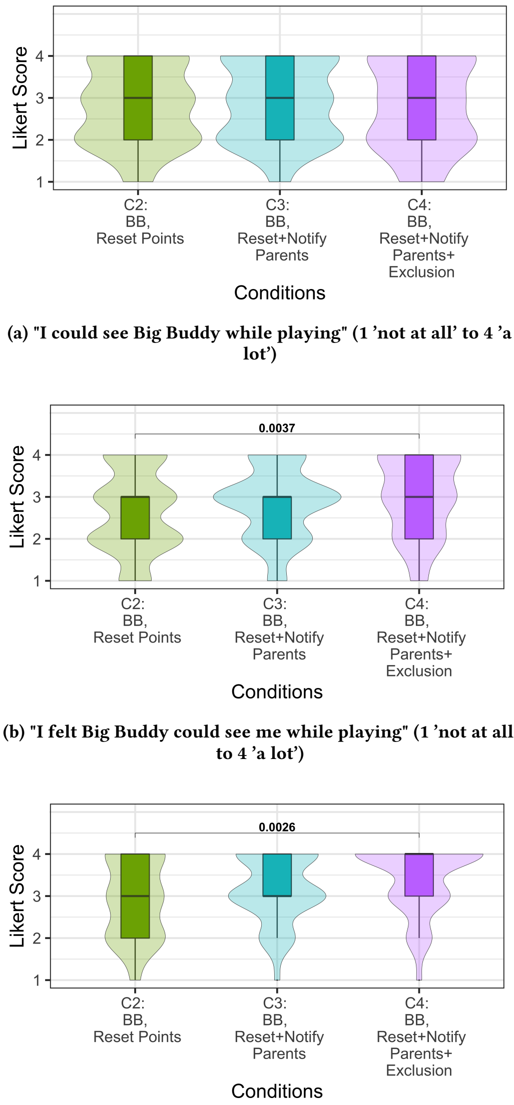
(a) SAM scores (1 'sad' to 5 'happy')

(b) SAM scores (1 'calm' to 5 'angry')

(c) SAM scores (1 'scared' to 5 'safe')

Figure 2: Violin-boxplots of three SAM Likert scales' scores for each condition. (a) Children felt significantly sadder in the round with C1 [No BB] compared to the round with C4 [BB, Reset+Notify Parents+Exclusion]. (b) No significant effect. (c) Children felt significantly safer in C2 [BB, Reset Points] than in C1 [No BB]. Significant pairwise comparisons are labelled.

making fun of them). Children noted he would be able to help them (n=9) Big Buddy's absence mostly led to making the child feel negative feelings (n=13) including: being annoyed (n=3), nervous (n=2), afraid (n=1), not safe (n=1), sad (n=1), tormented (n=1), having less fun when he was absent (n=1). *"I felt like I had more fun when he was there because he was making it fair"* [SchoolChild01b]. However, one child mentioned feeling better without Big Buddy *"when he was present it was just weird because all I felt was him staring at me not looking at the other student but when he was not there I felt better."* [Child01a] and found Big Buddy's presence inhibiting fun.



(a) "I could see Big Buddy while playing" (1 'not at all' to 4 'a lot')

(b) "I felt Big Buddy could see me while playing" (1 'not at all' to 4 'a lot')

(c) "Big Buddy helped with a fair punishment" (1 'not at all' to 4 'a lot')

Figure 3: Violin and boxplots of the three 4-point scales' scores for Big Buddy conditions. (a) No significant effect. (b) Children felt that Big Buddy could see them significantly more in C4 [BB, Reset+Notify Parents+Exclusion] than C2 [BB, Reset Points]. (c) Children felt that the punishment in C4 [BB, Reset+Notify Parents+Exclusion] was significantly fairer than in C2 [BB, Reset Points]. Significant pairwise comparisons are labelled.

Some children did not really notice him or pay attention to him as they were focused on their game until the disruptive situation and punishments were put in place (n=9), other children did not realise when he was not there (n=3), and some mentioned they were not looking at him but they knew he was there (n=3).

4.2.2 Parents' Perceptions towards Big Buddy. As parents watched the recorded video of the VR game, they were asked about their first impressions of the disruptive situations and how they felt knowing their child went through these events. Parents felt negatively (n=17) including the feeling of anger (n=2) *"Maybe a bit of anger [...] there was nothing I could do [...] I wasn't sure he's been in such situations before [...]"* [Parent33], annoyance (n=2), confusion (n=2), frustration (n=3), worry (n=2), feeling sorry for the children (n=1), feeling surprised or shocked (n=5), discomfort (n=1), unfairness (n=1), feeling strange because it is not real (n=1), feeling passive as it is just a game (n=1), it reminded of their own childhood experience (n=1) or hoping that their parenting allowed their child to be able to deal with these situations (n=1). For example: *"I hope that we've helped our child to be able to deal with that sort of thing. If they want to play then they need to be prepared for these things happening [...]"* [Parent30]. Parents also shared how they would imagine their child feeling: children would feel angry or upset (n=2), not a pleasant experience (n=1).

Impact of Big Buddy's Presence. Parents noted the usefulness of an embodied safeguarding system such as Big Buddy to increase their children's safety and well-being in virtual social environments. Parents mentioned Big Buddy's presence was reassuring (n=9) as someone was watching them (n=4) with quite a reassuring appearance (n=1), they would have more fear if he was not there (n=1) and felt it would make the place safer for their child (n=3), where they can feel more confident (n=1). Parents also thought Big Buddy's presence is good and important (n=8) and would want him to be there rather than having no supervision (n=8). One parent noted: *"I think it depends how much of a presence there is and how much it feels like you're being supervised. I suppose in some instances it would feel a bit weird to have someone looking over your shoulder all the time but in the same sense because that other player had already destroyed the tower [...]"* [SchoolParent01]. Some parents found Big Buddy helpful and useful (n=7) *"I think it's a helpful system not just not just suitable purely in terms of how they will be punished but to have that as a safety measure I think might be nice for everyone so that if something goes wrong there will be consequences [...]"* [Parent30]. Parents, however, indicated a few dislikes about Big Buddy: too intimidating (n=1), too passive (n=1), potentially not effective, its effectiveness may depend on children's age and if they take him seriously (n=2).

4.3 Perceptions towards Punishments (Fairness of Moderating System)

4.3.1 Children's Perceptions towards Punishments. As part of self-rating, children evaluated if Big Buddy put in place a fair punishment using a 4-point Likert scale. Across conditions, results' scores were statistically significantly different ($F(2,84)=5.94$, $p=0.004$, $\eta_p^2=.12$, medium effect size, ART ANOVA). Children felt that the punishment in C4 [BB, Reset+Notify Parents+Exclusion] was significantly fairer than in C2 [BB, Reset Points]($p=0.0026$) (see Figure 3 c).

Preferences have been raised for punishments during interviews. Children considered that the punishments were a way to increase safety (n=3). Among the three actions taken, children had different opinions regarding which punishment or combination of punishments were the fairest. Six children noted that putting in place the three punishments (C4 [BB, Reset+Notify Parents+Exclusion]) was better than putting just one (C2 [BB, Reset Points]). Seven highlighted that the fairest punishment is having the player out of the game. Some children thought that the combination of punishments was a bit extreme and severe (n=7), and that points reset to 0 was fair (n=13). In particular, Child01a and SchoolChild15 both suggested to have at least a first warning before being banned. However, resetting the competitor's points felt useless to some children (n=2). Two children preferred when the saboteur's parents were contacted. The latter was particularly disliked by others: feeling weird and uncomfortable (n=1), seemed as an unfeasible punishment (n=1), or not necessary (n=1). Children pointed out that punishments should have been the same as the disruption was the same (n=5).

4.3.2 Parents' Perceptions towards Punishments. Similar to children, preferences and pros and cons have been raised for each punishment. Overall, parents found the punishments adequate and reasonable (n=8) and one noted that all is covered *"I thought that's about as much as you could do in that situation that's probably the most you could do"* [Parent03]. Parents particularly liked the immediate consequence of having the saboteur's points reset to 0 (n=2). Some parents were positive about the time-out punishment (n=6) and liked the label showing that the player is punished (n=1). Others, similar to some children found the latter punishment severe (n=4) and a parent suggested to have a warning before exclusion. Regarding the punishment of parents being notified, parents were sceptical as to whether it is effective (n=4) in terms of other parents' parenting and if they would notice, it might worry parents, and for practicalities, you would need parents' contact details. Parents noted that punishments would depend on children's age as there might be an age limit for punishments' efficiency (n=2) *"I think that's probably going to reach a point where your kids are old enough they don't really mind about that [...]"* [Parent17]. Finally, parents suggested a progression of punishments based on repeated bad behaviour or level of bad behaviour (n=3) and making sure children understand Big Buddy's intentions (n=1).

4.4 Customisation of Big Buddy

4.4.1 How children envision the embodied moderating system. While eleven children mentioned they liked Big Buddy the way it was presented in the experiment's game, eleven other indicated wanting Big Buddy more like a realistic human and less like a robot and AI-looking, wearing normal clothes (n=10) and having a normal name (n=1). They found Big Buddy a bit intimidating and would prefer it to be more friendly (n=3). They would also prefer visual and audio characteristics that tend towards more real-life authority figures (teacher/parents) (n=2) or familiar/positive-related figures (Game Characters) (n=2), or their friends (n=1). In particular, two children did not like the robotic voice and found it uncomfortable. Children suggested having the possibility of personalising it (n=3), *"I think different people would want to see themselves or like some of their interests or something"* [SchoolChild14]. SchoolChild08 suggested to have Big Buddy with positive-related clothes or gestures, for example, wearing a t-shirt with a motivating and helpful note "You will be fine" and having him give thumbs up. Children's visions of a virtual supervisor based on the items of the scales were nuanced. Using K-means clustering analysis, we were able to observe three main clusters Figure 4. Among the three clusters, we observed preferred attributes for cluster 2 leaning towards a more realistic authority figure: authoritarian, visible, humanised, teacher whereas cluster 3 preferred a more friendly indulgent supervisor but still visible and humanised and cluster 1 would want it non-embodied.

4.4.2 *How parents envision the embodied moderating system.*

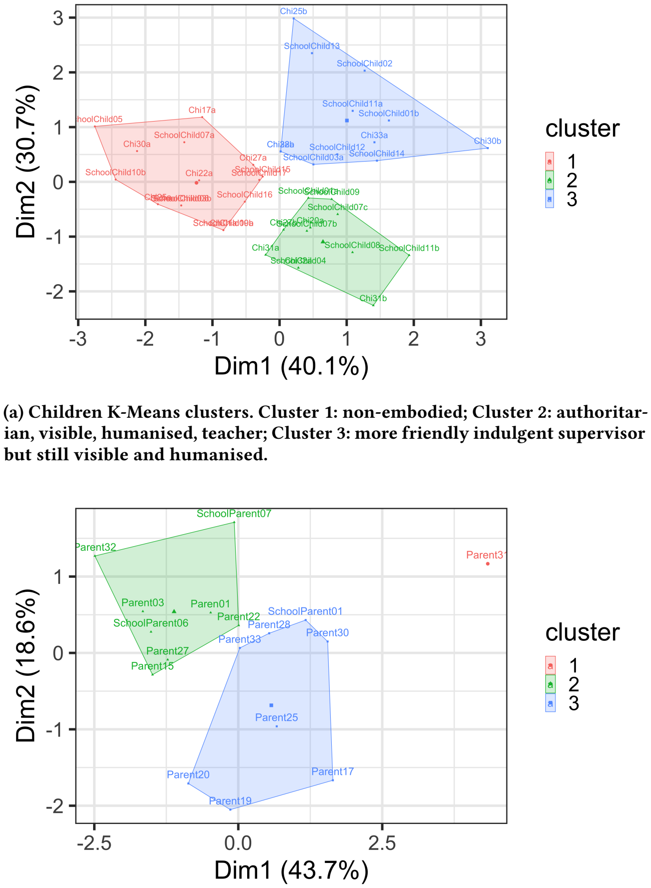
In terms of parents' preferred social and physical characteristics for an embodied moderating system, preferences among parents were shared again. Some said it should be context-dependent, based on the game children play (n=1), children's age and their preferences (n=1), and the situation (n=2). Four parents found Big Buddy too bulky, like a "big law enforcing robot" [Parent22] and would prefer someone more friendly (n=1), like an older kid as opposed to an adult (n=1), smaller and more discrete (n=2), with the interventions occurring in the background rather than in front of everyone (n=1). In contrast, it was suggested that Big Buddy names the player who did something wrong *"because they deserve to be like at least having something against the reputation"* [Parent32]. Parents also wanted something more familiar like a parent (n=1), a teacher (n=2), with a more natural voice (n=1), and more inclusive (e.g., gender, skin colour, personalised voice) (n=2) *"I would make them more um inclusive like showing white person and black person being supervisor and maybe personalise as well their tone of voice"* [Parent01]. Parent30 suggested it could be dressed differently for different/special roles but highlighted that it would not be the parents' role to choose the appearance etc. but it should be personalised by the player. Using K-means clustering analysis, we were also able to observe three main clusters among the sample of parents Figure 4. Preferred attributes for cluster 3 leaned towards an non-humanised and indulgent figure but formal and assertive to some extent whereas cluster 2 preferred a more realistic humanised expert and authoritarian supervisor and cluster 1, an outlier (Parent31) would want it as a complete opposite of the latter (indulgent, non-humanised, more friendly being an amateur, informal and not too assertive). Parent31 mentioned still wanting some regulation to some extent but would want to prioritise being a relaxing and fun experience for the children. Overall, preferred attributes from children and parents' perspectives can be visualised Appendix F. The reasons of their given scores are summarised in Appendix G.

4.4.3 *Parents' involvement in Safeguarding.*

While parents noted the usefulness of Big Buddy and how reassuring it might be, they would nonetheless remain in the supervision loop with different levels of involvement:

Trusting The System and Staying Involved to Safeguard Their Children Parents mentioned they would want to be more involved specifically in the beginning (n=8), to see how it works and if it is efficient before trusting the system. *"If I could see it operating okay and knew that it was effective for a certain period of time then I could probably get to the point where I say okay I know that this does work [...] I wouldn't trust it like I wouldn't trust anybody that I don't know"* [Parent22]. *"I'd like to know what the system is and how it works but once I've done that, I just think it's good for kids to be able to work things out between them so long as it's a fairly genuine system where all the kids will actually be genuine and they actually exist and so on"* [Parent30]. Parents would not trust the system or lack trust to some extent (n=7) until it has proven its effectiveness (n=2).

Real-Time Notification of Disruptive Events in Social VR Eight parents mentioned they would want to be notified real-time about the disruptive situations and the outcomes of the behaviours if there was a notification system. Parent15 and Parent19 pointed out they would want to be involved without real-time notifications as it may be difficult to manage and take care of real-time, and the latter system might pressure and inhibit children's fun and expression according to Parent15. Parent19 suggested to receive one notification or a report at the end of the day of what happened. Furthermore, parents suggested it would depend on different factors: SchoolParent01 highlighted that the use of the notification system (e.g., content and recurrence of notifications) may also depend on the age of the child and Parent25 pointed out it would depend on the level of the situation, if it is repeated for example.



(a) Children K-Means clusters. Cluster 1: non-embodied; Cluster 2: authoritarian, visible, humanised, teacher; Cluster 3: more friendly indulgent supervisor but still visible and humanised.

(b) Parents K-Means clusters. Cluster 1 (outlier): indulgent, non-humanised, amateur, informal, not too assertive; Cluster 2: humanised, expert, formal, authoritarian, visible, assertive; Cluster 3: non-humanised, visible, indulgent but formal and assertive.

Figure 4: Clusters for an AI-moderator attributes scores given. The optimal number of clusters was chosen based on the elbow method. Labels inside clusters correspond to participants' IDs.

5 SUMMARY AND DISCUSSION

5.1 The Children's Perspective

RQ1: child reaction and experience of Big Buddy - Big Buddy's presence and interventions have significantly impacted children's reactions to the disruption. Its absence induced negative feelings (e.g., sadness, frustration), whilst applying the full set of punishments (C4) significantly reduced sadness compared to no intervention (C1). **RQ2: perceptions towards Big Buddy and punishments** - Interestingly, C2 has also been perceived as substantially fairer than C1 which led us to believe that perceived fairness is a potential determinant of children's post-harassment reactions. Big Buddy's perceived role is congruent with these results, as children compared him with a referee to keep people safe, punish bad behaviours and ensure game fairness. Its presence is associated with active punishments and increased perceived safety. Regarding punishments, some children underlined the unfairness of applying different punishments for the same action. They also criticised the ban and notifying parents as too harsh. However, others thought it was appropriate and preferred the most severe interventions. **RQ3: customisation of moderator** - Children nonetheless raised concerns about Big Buddy's appearance and shared their design preferences. They can be clustered in three trends: 1) a more realistic authority figure, 2) a more friendly indulgent moderator, and 3) non-embodied moderator.

5.2 The Parents' Perspective

RQ2: perceptions towards Big Buddy and punishments - Parents showed empathy to their children experiencing the attacks and said they felt reassured by the presence of an embodied moderator such as Big Buddy. Parents provided similar feedback to children's and added that notifying parents may not be ideal. Its effectiveness relies on others' parents, and bullies may not be sensitive to it. Some proposed graduated interventions and an initial warning before intervening. **RQ3: customisation of moderator** - They expressed similar design preferences as their children but mentioned that its appearance and type of intervention would depend on various factors (e.g., gaming context, children's preferences, bad behaviour characteristics). **RQ4: parents' involvement** - They were not completely convinced by letting an AI moderating system ensure the safety of their children in Social VR and expressed their desire to better understand the process and remain in the loop.

5.3 Implications for Designing Embodied Moderation to Safeguard Children in Social VR

5.3.1 *Child and Parental Preferences are both Important.*

Grounded in our findings, children's and parents' answers were quite similar. Parents valued what their children would want, their needs and preferences and noted that enjoyment is important. Most children and parents felt reassured having a system such as Big Buddy to safeguard social VR. This means the design of embodied AI-moderation should involve the perspectives of both, parents and children.

5.3.2 *Leveraging the Opportunities brought forth by AI-driven and Embodied Moderation Systems.*

Immediate Consequences: One of the many opportunities enabled by systems like Big Buddy is allowing "*Immediate Consequences*". Indeed, some parents highlighted they particularly appreciated how the sabotage was detected and dealt with in real time. This is in contrast to typical moderation systems where decisions are delayed until they are reviewed by human moderators. Future systems can use social signal processing to detect when children are in distress and intervene accordingly [27, 46].

Customisation: Another advantage is that embodied moderation systems are "*customisable*". This allows for personalisation in multiple ways as it can be programmed to be perceived differently by different users [29]. This way, each user would see two versions of Big Buddy: a version that is on their side and a version that is effective in stopping said user from performing further offensive actions. The question remains as to which attributes should be personalised by the child to feel more comfortable and safe, which settings and rules that can be personalised by the parent, whether parents would make good arbiters of how Big Buddy should act and be presented, and which features would be fixed for all the players in social VR. Customisation raises concerns about the ability of parents and children to make informed decisions and the extent to which parents should be able to influence punishments for their child as a bully and/or a victim, particularly if they may be promoting harmful behaviour. Future research may be required to determine which aspects of moderation should be applied universally, tailored to specific demographics of children, or customised for individual families.

Parents can stay involved: Alongside our results, the experimental design showed that having parents view the recording of the VR game in real-time could allow parents to offer sympathy and empathy to their children. Warm and positive parent-child relationships are important for the growth and development of children and would need to be maintained through moderation and monitoring. In particular, some parents valued giving space to their children, letting them have a relaxing experience and being able to express themselves and still have fun. They may trust the AI-moderation system only to an extent, until it has proven its effectiveness. However, they would also want be aware and know what is happening and were positive about receiving real-time notifications if something bad occurs. There are also tensions around feasible parental involvement. Despite the benefits of real-time observation (that is often impractical), we need more research into asynchronous forms of parental involvement that can still give them the same opportunity for insight, oversight, and support.

Perceived Effectiveness of AI-Moderator Based On Multiple Factors: Our study shows the *potential* usefulness of an embodied AI-moderating system. However, its real-world usefulness and effectiveness may vary based on the children's age, their maturity, background, education and their preferences. It may also vary on the game played and the situation in the game as parents pointed out. Parents suggested that for younger children, someone visible and familiar would most likely be more effective than for older teenagers. Therefore, there is a need for personalised, embodied AI-moderators. For Social VR, a promising approach is the creation of pre-designed, validated moderators using proven intervention strategies as a necessary first step to adapt for different contexts.

5.3.3 *Importance of Safe Virtual Spaces for Children: Considering Context and Real-World Applications.*

This study on AI-moderating systems in social VR settings is crucial, given the increasing number of children in these environments and the potential for new forms of harassment. However, the effectiveness of AI moderation systems depends on various contextual factors such as the type of social VR experience, age range of users and type of threat (e.g., verbal, physical or environmental). To design safe virtual spaces for children, we need to consider these factors and explore potential redesigns that prevent harassment and foster positive interactions.

This study may also have implications for real-world bullying interventions. The principles used to create safe and positive virtual spaces may be applied to real-world situations to prevent bullying and harassment. By understanding the contextual factors that contribute to harassment and bullying in virtual environments, we can identify similar patterns in real-world settings and apply similar intervention strategies. This highlights the potential of virtual environments as a testing ground for interventions that can be used in both virtual and real-world settings.

6 CONCLUSION AND FUTURE WORK

The increased use of social VR platforms by children and the emergence of harassment enabled by embodied social VR experiences, underline the need for moderation, parental awareness and effective, accepted safety-enhancing technologies for children and parents. Through our experiment, we explored the extent to which an embodied AI-moderator prototype, Big Buddy, is perceived by children and parents as helping increase feelings of safety, comfort and reassurance and ensuring fairness. Children and parents prefer more realistic humanised authority figures, and familiar and positive-related figures. Parents would still like to remain in the supervision loop at different levels of engagement, through notifications, settings or direct parental watching, mainly due to a lack of trust in the system. Parents proposed progressive punishments based on repeated bad behaviours and including a warning phase. Preferences for the moderator's attributes were ambivalent. We concluded with implications for designing Embodied AI-moderators This work is crucial if we are to appropriately and effectively leverage and act upon advances in AI detection of harassment to create safer social VR environments.

Research needs to strongly consider challenges in embodied AI-moderator design, the risks posed by use of embodied and non-embodied AI moderators with children, and how to reliably and effectively detect provocative events. AI-moderation systems, while providing a level of constant monitoring, may undermine children's ability to respond to moderation, parenting, and supervision in real-life situations. Future research should focus on understanding how these AI systems affect children's social and emotional development, and whether or not they have a negative impact on their relationships with their parents. These systems may also affect children's privacy, autonomy, and well-being. We also need to consider if moderation alone is sufficient for addressing the needs of children. Future research should focus on understanding how AI-based moderation can be integrated with the pastoral role, such as providing emotional support and guidance to children. AI-based moderation systems may have gaps in their ability to detect problematic events and produce false positives. Dealing with uncertainty and missing contextual factors (e.g., body language and social cues), is a challenge that must be addressed to ensure the effectiveness of these systems. If trust in the AI-based moderation system is undermined by parents or children, it could also have a negative impact on its effectiveness. Additionally, there are ethical considerations for AI-moderation such as transparency, bias, accountability, and human oversight to ensure that the AI-moderator operates in a fair and responsible manner.

7 SELECTION AND PARTICIPATION OF CHILDREN

We followed a similar selection process as the study published in IDC 2021 [57]. Our study and experimental design were approved by the ethics committee of our university. Before participation, the details of the study were explained to the children's legal guardians/parents in the information sheet sent after they expressed interest and gave their consent in which it was noted that the child's participation is voluntary. At the start of each session, the researcher explained to the child what they would be doing, with slides and a video clip of the VR game. The researcher asked for a verbal agreement to participate and informed them that they could take breaks or stop at any point. The first batch (N = 20) was found via university forums and mailing lists (researchers or lecturers with children). The second batch (N = 28) was invited through the school's teacher who also filled in a form giving consent and contacted colleagues who had children in the school. All participants' personal data were stored securely, and all personally identifiable data were removed.

ACKNOWLEDGMENTS

We want to thank the families and Giggleswick School for participating. This work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1 and partly sponsored by a 2020 Meta Research Award on Responsible Innovation. We also want to thank Scotland's anti-bullying service RespectMe for sharing our study with their network to find participants.

REFERENCES

[1] 2022. AutoModerator. https://automoderator.app/bot/automoderator Last Accessed: 16-01-2023.

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

[3] 2022. UCL-VR/ubiq. https://github.com/UCL-VR/ubiq Last Accessed: 29-11-2022.

[4] 2022. User safety and moderation - AltspaceVR | Microsoft Docs. https://docs.microsoft.com/en-us/windows/mixed-reality/altspace-vr/user-safety Last Accessed: 30-08-2022.

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

[6] n.d.. K-means Cluster Analysis · UC Business Analytics R Programming Guide. https://uc-r.github.io/kmeans_clustering Last Accessed: 08-01-2023.

[7] Suzan Ali, Mounir Elgharabawy, Quentin Duchaussoy, Mohammad Mannan, and Amr Youssef. 2021. Parental Controls: Safer Internet Solutions or New Pitfalls? *IEEE Security and Privacy* (2021). https://doi.org/10.1109/MSEC.2021.3076150

[8] Nasiru I Aliyu, Musbau D Abdulrahaman, Fatimah O Ajibade, and Tosho Abdurauf. 2020. Analysis of Cyber Bullying on Facebook Using Text Mining. 1 (2020), 1–12. https://doi.org/10.48185/jaai.v1i1.30

[9] Kimberley R. Allison and Kay Bussey. 2017. Individual and collective moral influences on intervention in cyberbullying. *Computers in Human Behavior* 74 (9 2017), 7–15. https://doi.org/10.1016/J.CHB.2017.04.019

[10] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann Desmet, and Ilse De Bourdeaudhuij. 2014. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *http://dx.doi.org/10.1080/0144929X.2014.983979* 34 (4 2014), 425–435. Issue 4. https://doi.org/10.1080/0144929X.2014.983979

[11] 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

[12] 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

[13] Hannah Bloch-Wehba. 2020. Automation in Moderation. *Cornell International Law Journal* 53 (3 2020). Issue 1. https://scholarship.law.tamu.edu/facscholar/1448

[14] Lila Ghent Braine, Eva Pomerantz, Debra Lorber, and David H. Krantz. 1991. Conflicts With Authority: Children's Feelings, Actions, and Justifications. *Developmental Psychology* 27 (1991), 829–840. Issue 5. https://doi.org/10.1037/0012-1649.27.5.829

[15] 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

[16] J Brooke. 1996. Usability evaluation in industry, chap. SUS: a "quick and dirty" usability scale.

[17] Christoph Burger, Dagmar Strohmeier, and Lenka Kollerová. 2022. Teachers Can Make a Difference in Bullying: Effects of Teacher Interventions on Students' Adoption of Bully, Victim, Bully-Victim or Defender Roles across Time. *Journal of Youth and Adolescence 2022* (9 2022), 1–16. https://doi.org/10.1007/S10964-022-01674-6

[18] Heather M. Clarke and Lorne M. Sulsky. 2019. The Impact of Gender Stereotypes on the Appraisal of Civic Virtue Performance. *Journal of Research in Gender Studies* 9 (2019). https://heinonline.org/HOL/Page?handle=hein.journals/jogenst9&id=221&div=19&collection=journals

[19] Caitlin R. Costello and Danielle E. Ramo. 2017. Social Media and Substance Use: What Should We Be Recommending to Teens and Their Parents? *Journal of Adolescent Health* 60 (6 2017), 629–630. Issue 6. https://doi.org/10.1016/j.jadohealth.2017.03.017

[20] Maral Dadvar and Franciska de Jong. 2012. Cyberbullying Detection: A Step toward a Safer Internet Yard. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 121–126. https://doi.org/10.1145/2187980.2187995

[21] 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

[22] Giulio D'Urso and Ugo Pace. 2019. Homophobic bullying among adolescents: The role of insecure-dismissing attachment and peer support. *https://doi.org/10.1080/19361653.2018.1552225* 16 (4 2019), 173–191. Issue 2. https://doi.org/10.1080/19361653.2018.1552225

[23] Giulio D'Urso, Jennifer Symonds, Seaneen Sloan, and Dympna Devine. 2022. Bullies, victims, and meanies: the role of child and classmate social and emotional competencies. *Social Psychology of Education* 25 (2 2022), 293–312. Issue 1. https://doi.org/10.1007/S11218-021-09684-1/TABLES/3

[24] Lisa A Elkin, Paul G Allen, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests; An Aligned Rank Transform Procedure for Multifactor Contrast Tests. 15 (2021). Issue 21. https://doi.org/10.1145/3472749.3474784

[25] L. Christian Elledge, Timothy A. Cavell, Nick T. Ogle, and Rebecca A. Newgent. 2010. School-based mentoring as selective prevention for bullied children: A preliminary test. *Journal of Primary Prevention* 31 (6 2010), 171–187. Issue 3. https://doi.org/10.1007/S10935-010-0215-7

[26] 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

[27] 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

[28] 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

[29] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. 2020. My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382923

[30] Gary W. Giumetti and Robin M. Kowalski. 2022. Cyberbullying via social media and well-being. *Current Opinion in Psychology* 45 (6 2022), 101314. https://doi.org/10.1016/J.COPSYC.2022.101314

[31] 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.

[32] Sevtap Gurdal and Emma Sorbring. 2019. Children's agency in parent–child, teacher–pupil and peer relationship contexts. *https://doi.org/10.1080/17482631.2019.1565239* 13 (6 2019). Issue sup1. https://doi.org/10.1080/17482631.2019.1565239

[33] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[34] Heidi Hartikainen, Netta Iivari, and Marianne Kinnula. 2016. Should We design for control, trust or involvement? A discourses survey about children's online safety. *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children* (6 2016), 367–378. https://doi.org/10.1145/2930674.2930680

[35] Jie He, Xinyi Jin, Meng Zhang, Xiang Huang, Rende Shui, and Mowei Shen. 2013. Anger and selective attention to reward and punishment in children. *Journal of experimental child psychology* 115 (7 2013), 389–404. Issue 3. https://doi.org/10.1016/J.JECP.2013.03.004

[36] 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

[37] Julie A Hubbard. 2001. Emotion expression processes in children's peer interaction: The role of peer rejection, aggression, and gender. *Child development* 72, 5 (2001), 1426–1438.

[38] Catherine Knibbs. 2022. *Children, technology and healthy development : how to help kids be safe and thrive online.* 183 pages. https://www.routledge.com/Children-Technology-and-Healthy-Development-How-to-Help-Kids-be-Safe-and/Knibbs/p/book/9780367770150

[39] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 437, 12 pages. https://doi.org/10.1145/3411764.3445279

[40] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 2005. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* NIMH, Center for the Study of Emotion & Attention Gainesville, FL.

[41] Shaun Alexander MacDonald, Euan Freeman, Stephen Brewster, and Frank Pollick. 2021. User Preferences for Calming Affective Haptic Stimuli in Social Settings. *ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction* (10 2021), 387–396. https://doi.org/10.1145/3462244.3479903

[42] Thabo Mahlangu, Chunling Tu, and Pius Owolawi. 2019. A review of automated detection methods for cyberbullying. *2018 International Conference on Intelligent and Innovative Computing Applications, ICONIC 2018* (1 2019). https://doi.org/10.1109/ICONIC.2018.8601278

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

[44] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. *CHI PLAY 2020 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 472–483. https://doi.org/10.1145/3410404.3414268

[45] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. *Proceedings of Interaction Design and Children, IDC 2021*, 69–79. https://doi.org/10.1145/3459990.3460703

[46] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. "Talking without a Voice": Understanding Non-Verbal Communication in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 175 (oct 2020), 25 pages. https://doi.org/10.1145/3415246

[47] 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

[48] 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

[49] Peter E. Morris and Catherine O. Fritz. 2013. Effect sizes in memory research. *http://dx.doi.org/10.1080/09658211.2013.763984* 21 (10 2013), 832–842. Issue 7. https://doi.org/10.1080/09658211.2013.763984

[50] 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)*. 211–220. https://doi.org/10.1109/ISMAR52148.2021.00036

[51] Aaron Powers and Sara Kiesler. 2006. The Advisor Robot: Tracing People's Mental Model from a Robot's Physical Attributes. *HRI 2006: Proceedings of the 2006 ACM Conference on Human-Robot Interaction* 2006, 218–225. https://doi.org/10.1145/1121241.1121280

[52] John T.E. Richardson. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review* 6, 2 (2011), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

[53] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (4 2019), 333–345. https://doi.org/10.1016/J.CHB.2018.12.021

[54] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior* 66 (1 2017), 345–352. https://doi.org/10.1016/J.CHB.2016.09.040

[55] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (12 1980), 1161–1178. Issue 6. https://doi.org/10.1037/H0077714

[56] 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. (2023), 17. https://doi.org/10.1145/3544548.3581090

[57] Evropi Stefanidi, Maria Korozi, Asterios Leonidis, Dimitrios Arampatzis, Margherita Antona, and George Papagiannakis. 2021. When Children Program Intelligent Environments: Lessons Learned from a Serious AR Game. *Proceedings of Interaction Design and Children, IDC 2021* (6 2021), 375–386. https://doi.org/10.1145/3459990.3462463

[58] Marie S. Tisak, Dushka Crane-Ross, John Tisak, and Amanda M. Maynard. 2000. Mothers' and Teachers' Home and School Rules: Young Children's Conceptions of Authority in Context. *Merrill-Palmer Quarterly* 46, 1 (2000), 168–187. http://www.jstor.org/stable/23093347

[59] Wen-Jie Tseng, Elise Bonnail, Mark McGill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2022. The Dark Side of Perceptual Manipulations in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 612, 15 pages. https://doi.org/10.1145/3491102.3517728

[60] Rogier E.J. Verhoef, Anouk van Dijk, Esmée E. Verhulp, and Bram O. de Castro. 2021. Interactive virtual reality assessment of aggressive social information processing in boys with behaviour problems: A pilot study. *Clinical Psychology and Psychotherapy* 28 (5 2021), 489–499. Issue 3. https://doi.org/10.1002/cpp.2620

[61] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. (2011). http://faculty.washington.edu/wobbrock/art/

[62] Serkan Çankaya and Hatice Ferhan Odabaşi. 2009. Parental controls on children's computer and Internet use. *Procedia - Social and Behavioral Sciences* 1 (1 2009), 1105–1109. Issue 1. https://doi.org/10.1016/J.SBSPRO.2009.01.199