Big Buddy: A Simulated Embodied Moderating System to Mitigate Children's Reaction to Provocative Situations within 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

The use of social Virtual Reality (VR) among children is increasing, but with it comes new forms of harassment that can be difficult for parents to monitor. To address this issue, we have developed "Big Buddy", a prototype AI-moderator that aims to safeguard children from potential harassment in social VR. We conducted a study in which 43 children (aged 8-16) participated in a simulated social VR classroom, with fictitious competitors disrupting their game. When Big Buddy intervened, the children reported feeling significantly less negative emotions and felt safer. This is the first study to empirically examine the use of an embodied AI-moderator in social VR from the perspective of children, and it provides important insights for designing AI-moderators in social VR.

CCS CONCEPTS

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

KEYWORDS

social virtual reality, children, online harassment, embodied moderating system, artificial moderator

ACM Reference Format:
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), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585840

1 INTRODUCTION

Social Virtual Reality (VR), which was originally intended for adults and older teenagers, has seen an increase of younger children under the age of 13 using the simulated social environment [27–29]. While social VR can create an innovative way of engaging and interacting with others due to unique embodiment and presence in VR with the illusion of "being there" [30], users can inflict virtual harm on others, leading to an increase of new forms of harassment and bullying [8, 27, 29, 40]. For example, children and adults have reported harassment ranging from name calling to virtual sexual harassment [27, 29]. Unfortunately existing mitigation features, such as blocking, personal space bubbles, muting, reporting players [1, 3] or trust systems to keep users safe from nuisance users [4], suffer from significant limitations. They place the responsibility of their application on potentially ill-equipped users including children or guardians (e.g., unfamiliar with the technology or not knowing the best approaches) [9, 18, 20, 22]. Moreover, there is a growing need for understanding the effectiveness of safety-enhancing technologies for children in social VR. While research has shown that human-based moderators can help establish norms for appropriate behaviours [8], research also showed 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 [17].

In this study, we introduced Big Buddy, an artificial avatar as a Wizard of Oz (WOZ) AI-moderator prototype, to safeguard children from disruptive social VR. 43 children (8-16 years old) played a researcher-constructed VR interactive tower of blocks construction game. It involved fictitious competitors disrupting the participant's game in a virtual classroom mimicking a social VR environment. We evaluated children's emotions and perceptions towards Big Buddy and his interventions. 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 perceive the embodiment of the moderating system (Big Buddy) and the system of punishments? Will children feel safer and/or inhibited knowing there is an embodied moderator agent?

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

This paper provides new insights into HCI and child-computer interaction. We conduct the first study of a VR-based WOZ AI-moderator to improve children's safety and comfort in social VR environments. We also gather data on design features for an AI-moderator that

CHI EA '23, April 23–28, 2023, Hamburg, Germany

balances safety with children's agency and enjoyment. We propose future design directions for effective AI-moderators 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 reality (e.g., at school) or online [41]. As children begin to communicate with others they learn desirable and undesirable behaviours in social settings [24]. 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 [21, 23, 31, 41]. The design and methodologies of these 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 study evaluated children's aggressive behaviour using interactive VR scenarios [41]. Boys aged 8-13 were individually tested in a silent room. Results showed that VR scenarios involving peer provocation (participants were refused to join a game by two virtual peers and participants' game was ruined by the virtual peer) led to more aggressive responses than neutral and instrumental gain scenarios (participants could choose to steal a block or ball from the virtual peer to obtain additional points or could win the game by sabotaging the virtual peer's game). The study inspired our VR game as it was shown that the game was challenging enough but not too difficult for children aged 8-13, and demonstrated a provocative situation designed to provoke an emotional response.

Anger and frustration are negative emotions according to Russell's complex [38] that can be elicited in tasks such as toy removal and were shown to orient children towards desirable goals or objects [21]. The study, involving 40 children aged 5-6, used self-reported emotions with the established Self-Assessment Manikin (SAM) [25] and a game where participants would compete with an unfamiliar player and would always lose. Results showed that anger is associated with attention biases towards rewards rather than punishment.

2.1.2 Perceptions Towards Sanctions and Authority. 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 [31]. 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 [19]. 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 [19]. Another study investigated children's conception of authority from an individual and showed that it would depend on their status, the context and the domain of act depending on children's age [39]. Younger children's awareness of mothers and teachers authority was shown to be greater than that of police for instance [10, 39]. Therefore, there is the potential need of having a visible embodied moderator as an authority figure for children in social VR.

2.2 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 [8]. Children and adults have reported harassment, from name calling to physical stalking [27, 29]. Current mediation tools (e.g., reporting and blocking) have been shown to be insufficient, have yet to establish user trust, and have led to feelings of unfair treatment - discouraging users from using them [27]. Interventions to protect children effectively 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 [6, 13, 14, 43]. Indeed, 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 [33]. Research has shown that real moderators can help establish norms for appropriate behaviours [8], with nuanced views towards human-based moderators [17] 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 [17]. Moreover, 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 [7, 26, 36]. 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. Therefore, we introduce a simulated embodied AI-moderating system and evaluate child perceptions towards it.

3 METHODS

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. We simulated a social VR environment game with disruptive situations based on prior research [41] and implemented a WOZ prototype of an 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 [12, 15, 31, 37]. We measured child reactions using SAM scales [23], and designed Likert scales and interview questions to measure child perceptions towards Big Buddy. The protocol was approved by our ethics committee.

3.1 Procedure

Children, accompanied by a parent or teacher, were given a presentation and were told they will be playing a virtual classroom game with real players from 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 a Meta Quest 2 headset with an adjustable strap for more comfort. Participants were standing and 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 players. The goal was to build a tower of 5 blocks as quickly as possible and ensure it remains stable until the time limit. Points were awarded to players who successfully met this goal. After each round, they had to fill in a form on a laptop, with the questions described in section 3.4. At the end of the game, each child was 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 (i.e., 4 rounds, each round with one level): (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 to 0; (3) *C3: [BB, Reset+Notify Parents]* - Big Buddy is present and intervenes: resets saboteur's points to 0 and notifies saboteur's parents ; (4) *C4: [BB, Reset+Notify Parents+Exclusion]* - Big Buddy is present and intervenes: reset saboteur's points to 0, notifies saboteur's parents and saboteur is excluded from the game.

Big buddy announced these punishments and punishments were written in a bubble, ensuring that the child participant is aware of the actions taken. The punishments were based on the top three punishments of the perceived effective punishments in a school according to children and parents [31]. 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] (due to omitting 5 participants from the analysis, see section 3.1). The provocative situation was the same throughout the game. 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 that could lead to fed up emotions or lack of game credibility.

3.3 Social VR Game and Avatars Implementation

The simulated social VR game, used to conduct the experiment, was developed by the experimenter using Unity 3D. Its design was based on the game used to evaluate the aggressive behaviour in boys [41]. We simulated similar provocative situations to the latter study, 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 [2] with animations and a voice generator. To stand out from the other avatars, Big Buddy was designed to be bigger and with a noticeable appearance. 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.) and available differentiable visual look from the Ubiq library (See Figure 1 e)). The game was first 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

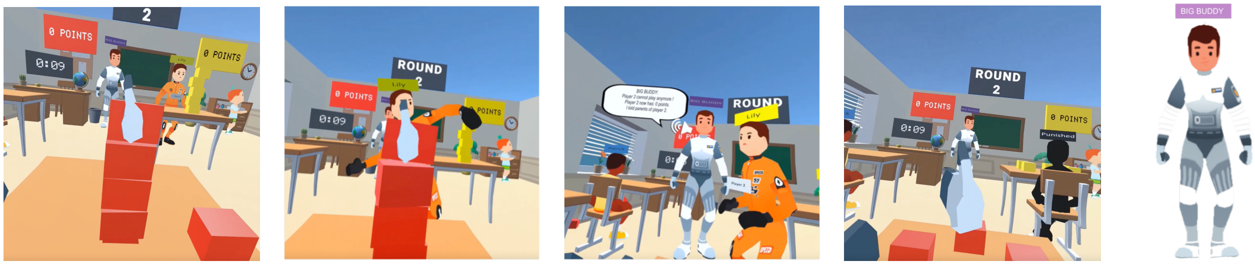
Quantitative: At the end of each round of the game, children were asked to self-rate the emotions felt when the other player destroyed their tower, using the three 5-point SAM Likert scales [25], readapting the third one (from scared/intimidated to safe). If Big Buddy was present in the round, they had to give a score on three 4-point-Likert scales (from 'a little' to 'extremely': 1) Seeing Big Buddy when playing, 2) Feeling seen by Big Buddy, 3) Big Buddy helped with a fair punishment). At the end of the game, during interviews, children were separately asked to give scores for preferred physical and social attributes, using 5-point Likert 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 [34], to gain a better understanding on what children would want as a preferred embodied moderating system in social VR (See supplementary material). *Qualitative:* At the end of the game, children were separately interviewed with questions around their perceptions of the bully and Big Buddy, their sense of safety, and potential customisation of Big Buddy (see interview questions in supplementary material).

3.5 Analysis

Quantitative: We analysed our quantitative data using R statistical tools. We conducted one-way ART ANOVAs (significant level at 0.05 and effect sizes: medium (between 0.06 and 0.14) and large (0.14 or higher) partial eta squared η_p^2 [32, 35]) using ARTool and post-hoc comparisons with Tukey adjusment [16, 42]. We also used K-means clustering analysis to analyse preferred physical and social attributes of Big Buddy [5]. The analysis and codes can be found: (link to be added, see supplementary material). *Qualitative:* We used inductive thematic analyses techniques [11, 33] to analyse children's perceptions from audio recorded interviews. After generating 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 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

The children (8-16 years old M = 11.49 [SD = 2.09]) 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)



(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.

was found via the university forums and mailing lists. The second batch (N_children = 28) was from a school where the lead researcher spent two days. This required permission from the headteacher and the parents filling in the form. 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 5 child participants from the data analysis. The total number of participants is therefore: **43 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 their 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. We note that 40 children were interviewed at the end of the game. Children were from 31 families. An £8 Amazon-voucher and commute costs were compensated to adults for their time and a token of appreciation was given to the children.

3.7 Limitations

Participants were mostly White Caucasian. The study was done in English in the UK which may introduce cultural bias. The experiment was conducted in two locations, with one group accompanied by parents and the other in a school, both with children separated from the adult to avoid their influence. Some sessions were done in pairs, but the researcher ensured they could not talk to each other and they answered their questions quietly on a separate device. Omitting 5 participants resulted in an imbalanced Latin Square randomisation of 4 conditions, which may limit the study due to possible valence effects, but countermeasures such as balancing wins and losses and number of games per round have been taken to reduce this impact.

4 RESULTS

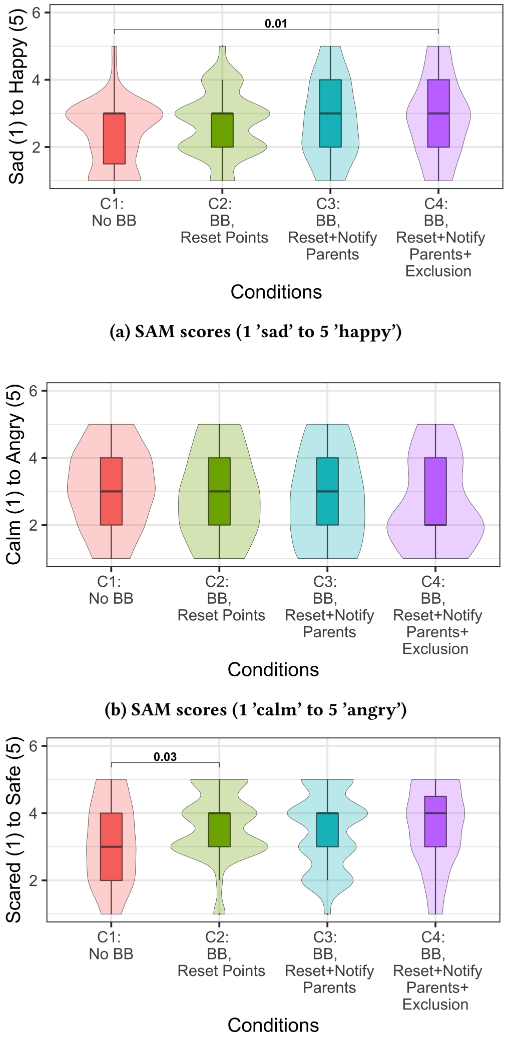
4.1 Emotional Reaction and Perceptions of Big Buddy after Disruption

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 with medium and large partial eta squared η_p^2, were obtained for the different conditions for sadness (F(3,117)=3.55, p=0.02, η_p^2=.08, medium effect size) and for safety (F(3,117)=3.22, p=0.03, η_p^2=.08, medium effect size). 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, Tukey adjustment) in C2: [BB, Reset Points] than in C1: [No BB].

From self-ratings, 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). They felt Big Buddy could see them significantly more in C4: [BB, Reset+Notify Parents+Exclusion] than C2: [BB, Reset Points] (p=0.0037, Tukey adjusment).

Big Buddy's Role: Children were asked to describe 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.

Impact of Big Buddy's Presence and Absence: Regarding children's perception towards Big Buddy presence compared to his absence, on the one hand, children felt that his presence was reassuring and they felt safer (n=10) *"I felt a different way when he was there because it felt a lot safer."* [SchoolChild07b]. 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). Eight children were annoyed by the absence of Big Buddy was there were 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 making fun of them). Big Buddy's absence mostly led to making the child feel negative feelings including (n=13): 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). However, Child01a mentioned feeling better without Big Buddy and SchoolChild01b found Big Buddy's presence inhibiting fun. 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).



(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.

4.2 Perceptions towards Punishments (Fairness of Moderating System)

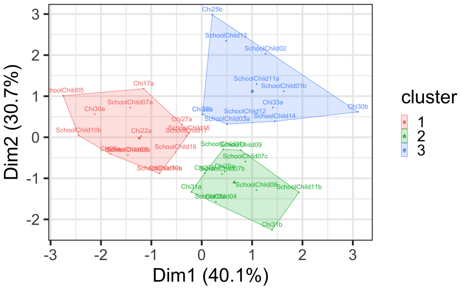
As part of self-rating, children evaluated if Big Buddy put in place a fair punishment using a 4-point Likert scale. Across conditions, results' scores were statistically significantly different (F(2,84)=5.94, p=0.004, η_p^2=.12, medium effect size). Children felt that the punishment in C4: [BB, Reset+Notify Parents+Exclusion] was significantly fairer than in C2: [BB, Reset Points] (p=0.0026). 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 but the latter was 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 be the same to all saboteurs as the disruptive situation was the same (n=5).

4.3 Envisioning an 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, giving thumbs up or wearing a t-shirt with a motivating note "You will be fine". Children's visions of an embodied AI-moderator based on the items of the scales were nuanced. Using K-means clustering analysis, we were able to observe three main clusters Figure 3. Cluster 2 leans towards a more realistic authority figure: authoritarian, visible, humanised, teacher whereas cluster 3 preferred a more friendly indulgent supervisor and cluster 1 would want it non-embodied.

5 SUMMARY AND DISCUSSION

RQ1: child reaction of Big Buddy - Children felt significantly less sad in C4: [BB, Reset+Notify Parents+Exclusion] and significantly safer in C2: [BB, Reset Points] compared to C1: [No BB]. Children also found C4: [BB, Reset+Notify Parents+Exclusion] fairer compared to C2: [BB, Reset Points]. It seems the selection of the best punishment and resulting feelings are driven by perceived fairness. The absence of Big Buddy and thus, the absence of punishments induced negative feelings (e.g., frustration). **RQ2: perception towards Big Buddy and punishments** - Big Buddy's role is perceived as a referee to keep people safe, punish bad behaviours and ensure fairness of the game. It has been compared to a teacher, perhaps as it was in a virtual classroom. Its presence is associated with active punishments and increases perceived safety. Yet, being watched raised discomfort of being watched, they would prefer a more discrete moderator that appears only when necessary and its



(a) K-Means clusters. Labels inside clusters are participants’ IDs.

	Authoritarian (-) Indulgent (+)	Non-humanised (-) Humanised (+)	Visible (-) Non-visible (+)	Teacher (-) Friend (+)
Red	Around 0	- or 0	+	Mixed
Green	- or 0	+	- or 0	- or 0
Blue	+ or 0	+	-	+

(b) Preferred attributes for each cluster based on scores given.

Figure 3: Children clusters for preferred AI-moderator attributes. Optimal clusters’ number based on elbow method. Cluster 1 (red): non-embodied; Cluster 2 (green): authoritarian, visible, humanised, teacher; Cluster 3 (blue): more friendly indulgent supervisor but still visible and humanised.

efficiency to prevent attacks has been questioned. **RQ3: customisation of moderator** - Social and physical preferred characteristics of an embodied AI-moderator tend towards more real-life authority figures (teacher/parents) or familiar/positive-related figures (Game Characters). Robotic features (i.e., voice, space suit) have not been well received and described as intimidating. Most children prefer an embodied moderator that is visible, looks and talks more human-like, with a balance between being authoritative and indulgent, friendly yet respected.

Our study reveals a number of particularly novel and useful features of AI-driven embodied moderating systems, that should be capitalised on when designing such systems. However, further research is needed to determine which attributes should be personalised by the child for comfort and safety, which settings and rules can be personalised by the parents, and which features should be fixed for all players in social VR. While our study shows potential usefulness of an embodied AI-moderating system, its 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. For younger children, someone visible and familiar would most likely be more effective than for older teenagers. Therefore, there is a need of personalised embodied AI-moderators but in terms of practicalities in social VR, a set of pre-designed moderators may be needed for different age groups, games and situations. Research should also 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. We also need to consider if moderation alone is sufficient for addressing the needs of children. Furthermore, ethical considerations must be taken into account when implementing AI-moderation, including privacy, transparency, bias, accountability, and human oversight to ensure fair and responsible operation of the AI-moderator.

6 CONCLUSION

The increased use of social VR platforms by children and the emergence of new forms of harassment in embodied social VR experiences has led to a growing need for moderation and effective safety-enhancing tools. Our experiment examined the perception of an embodied WOZ AI-moderator, Big Buddy, among children and identified design features that could help increase feelings of safety and comfort. The results showed that children felt reassured and safer with the presence and intervention of Big Buddy, and perceived its role as a referee. Additionally, children preferred more realistic and humanised authority figures, and familiar and positive-related figures. We hope that our findings contribute to a better understanding of children’s perceptions towards embodied AI-moderators in social VR, and lead to future research on safer, inclusive, and personalised AI-moderators for different groups based on children’s age, social VR environments, and parental oversight needs.

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.

REFERENCES

[1] 2022. Comfort and Safety — Rec Room. https://recroom.com/safety Last Accessed: 30-08-2022.
[2] 2022. UCL-VR/ubiq. https://github.com/UCL-VR/ubiq Last Accessed: 29-11-2022.
[3] 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.
[4] 2022. VRChat Safety and Trust System. https://docs.vrchat.com/docs/vrchat-safety-and-trust-system Last Accessed: 30-08-2022.
[5] n.d.. K-means Cluster Analysis · UC Business Analytics R Programming Guide. https://uc-r.github.io/kmeans_clustering Last Accessed: 08-01-2023.
[6] 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
[7] 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
[8] 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
[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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.

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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.

[24] 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

[25] 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.

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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/

[43] 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