"Please Be Nice": Robot Responses to User Bullying - Measuring Performance Across Aggression Levels

Yiming Luo, Shihao Liu*
luoyiming@xidian.edu.cn
23151214134@stu.xidian.edu.cn
Xidian University
Xi'an, China

Di Wu
di.wu@ntnu.no
Norwegian Univ. of Science and Technology
Trondheim, Norway

Hao Wang
wanghao@xidian.edu.cn
Xidian University
Xi'an, China

Yushan Pan†
pys0486@hotmail.com
Xi'an Jiaotong-Liverpool University
Suzhou, China

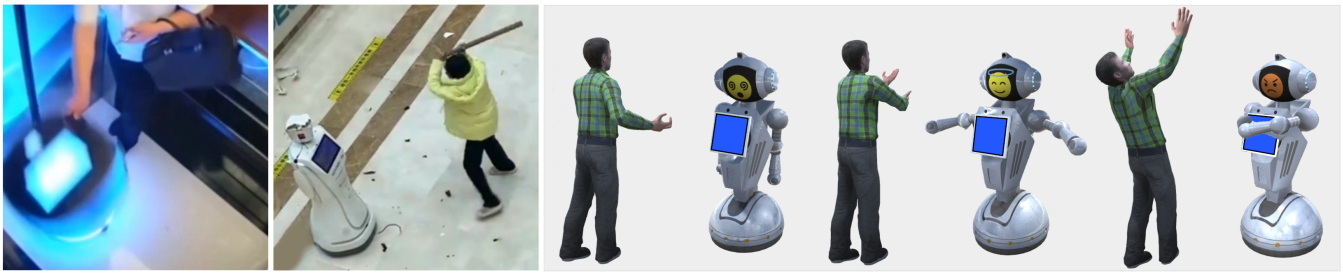

Figure 1: (left) two actual examples from news reports depicting users bullying robots; (right) an example of our designed robot response after user bullying.

ABSTRACT

As robots become integral to public services, addressing harmful user behaviors like bullying is crucial. Existing research often overlooks the gradual nature of human bullying. This study fills this gap by exploring how robots can counter bullying through optimized responses. Using a simulated human-robot interaction study, we manipulated robot response behaviors and styles across escalating bullying severity. Results show that empathetic verbal responses promptly reduce users' bullying tendencies by eliciting remorse and redirecting attention to social awareness. However, users' underlying dispositions may override these reflexive reactions, emphasizing the need for a holistic understanding. In conclusion, a comprehensive approach is essential, involving immediate reaction optimization, emotional state assessment, and ongoing behavioral adjustment through empathetic dialogue. By implementing such strategies, we can transform human-robot relationships from potential bullying situations to harmonious interactions. This study provides an empirical foundation for response protocols that discourage bullying and enhance mutual understanding.

CCS CONCEPTS

• **Human-centered computing** → **User studies**.

KEYWORDS

Human-Computer Interaction, Bullying, Human-Robot Interaction

ACM Reference Format:
Yiming Luo, Shihao Liu, Di Wu, Hao Wang, and Yushan Pan. 2024. "Please Be Nice": Robot Responses to User Bullying - Measuring Performance Across Aggression Levels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613904.3642290

1 INTRODUCTION

The rapid advancement of service robotics, driven by technologies such as artificial intelligence and machine learning [49], has already led to an increasing number of robots being introduced into public spaces for human interaction [6]. Social robots are used in various settings, including banks [11, 20], restaurants [10, 38], hospitals [25, 36, 51], hotels [15, 23] and other service encounters [37, 46]. These service robots are perceived as autonomous and adaptable technology interfaces, enhancing productivity and efficiency in the service industry [24, 49].

*Both authors contributed equally to this research.
†Corresponding author

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642290

Despite the advantages of service robots, various challenges and uncertainties arise [40]. In instances of misconduct, it is unclear how intelligent robots will react. Will they behave as inanimate objects that are unresponsive, or will they retaliate against humans? If the latter occurs, how can we provide insights to help robot developers design responsible robots for interactions with humans? Just as bullying occurs gradually [3], as depicted on the left of Fig. 1, how should the robot behave to respond to humans and create a peaceful environment?

In real life, bullying manifests in various forms within the realm of human sciences [30]. Workplace bullying, encompassing physical, mental, or social threats, involves behaviors such as intimidation, threats, exclusion, and verbal or physical abuse. In the Human-Robot Interaction (HRI) field, bullying is traditionally seen as a subset of aggressive behavior [32], characterized by being unprovoked, repetitive, and marked by a power imbalance [40], where the aggressor is perceived as stronger than the victim [31]. This power imbalance is a defining feature, setting bullying apart from other forms of aggression among equals [40]. Bullying predominantly unfolds within peer groups, especially in educational environments where children of similar ages frequently interact [34]. Early research underscores the prevalence of bullying within school contexts [34, 40, 42, 48], some studies reporting as many as two bullying incidents occurring per classroom hour, albeit of the short duration of service robots by users [1].

Explicitly, bullying, in HRI studies, is partially adopted from human sciences and is significantly considered in the evaluation of interactions between humans and robots as behavior that offends, humiliates, or intimidates, possibly occurring as a one-off conflict. In this context, bullying is marked by a lack of remorse or a failure to recognize human behavior as problematic, often fueled by emotions like anger, frustration, or jealousy [31], among other reasons. Furthermore, these bullying behaviors can extend beyond human interactions to involve interactions between people and robots [31]. As bullying is characterized by a gradual progression from ideas to actions [43], this progression extends the measurement of bullying beyond one-off conflict cases, resulting in the design of robot systems that are expensive to maintain. Consequently, if the robots cannot act or react to ease the intention of human bullying, it will be challenging to prevent unnecessary damage or loss of robots when bullying occurs. Thus, evaluating various types of interactions where bullying might occur is crucial. This assessment is essential for informing the future design of robot systems in a cost-effective manner.

Although some researchers have explored how robots can respond to user bullying from a reactive standpoint [45], that prevent bullying approaches, such as investigating whether response styles without profanity could influence emotions [5] and mitigate aggressive behavior during interactions between humans and conversational agents [12, 17], still assume that a simple reaction can solve the complex and context-dependent bullying [8] between users and service robots while working together on tasks[8, 30]. Indeed, there are a few studies focus on empowering users' rights during the interaction between users and service robots and how robots should handle complex tasks with users before bullying occurs, there is still a gap in evaluating bullying in a short-term, inexpensive way [42]. Addressing this gap is crucial for providing practical insights to improve HRI design [7]. In that sense, there is a pressing need for an affordable approach to designing robots that can promote greater mutual understanding between users and robots, thereby discouraging bullying.

Given this rationale, this paper aims to investigate how human-like robots can effectively handle varying levels of unfriendly behavior leading up to bullying from users. While the present work is lab-based, it has the high potential to shed light on designing future robot systems that can prevent bullying, protect the robot from damage, and avoid unnecessary harm to human beings. The primary objective is to explore response approaches that can positively influence user behavior and reduce bullying. To achieve this, we manipulate robot response types and user bullying types to understand the intricate connections between bullying severity, response approach, and user reactions in a simulated HRI environment. We design a robot with six response types, including verbal and physical modalities, as well as avoidant, empathetic, and defiant styles, along with their combinations. Additionally, we define four user bullying types to reflect varying degrees of severity. A comparative test examines the correlations between bullying behavior and emotions of users. Our study makes a significant contribution by highlighting that bullying tends to occur gradually rather than abruptly in a single instance. This insight suggests that we can potentially prevent instances of bullying by enhancing the frequency of interactions between users and robots. Moreover, we can also mitigate bullying by offering alternative responses that can help pacify users during their interactions with robots.

The paper is organized as follows: Section 2 presents related work. Section 3 presents the pilot study. Section 4 illustrates the user study. Section 5 presents our results. Discussion is presented in Section 6. Section 7 presents the limitations and future work. The paper is concluded in Section 8.

2 RELATED WORK

This section presents a descriptive overview of bullying studies within the HRI field. According to the prevailing trends in current research, bullying can be categorized into three primary segments: 1) Analyzing bullying through research methodologies, 2) Utilizing technology to prevent bullying, and 3) Collaboratively designing prototypes with participants to counter potential bullying scenarios. With these three approaches significantly contributing to the investigation of bullying in HRI, there remains substantial room for further advancement in the study. Enhancing research efforts can aid in a more comprehensive understanding of the natural occurrence of bullying in human society and elucidate its similarities and differences within the HRI field. This, in turn, will aid us in designing robots to better support interactions between humans and robots.

2.1 Analyzing bullying via methods

In the sphere of interactions with socially interactive robots, comprehending human responses plays a critical role in shaping ethical, societal, and legal perspectives, aiding the development of responsible robotics and their successful integration into our society [8]. These robots elicit social responses by conforming to expected human behavioral norms, aiming to evoke social interactions rooted

in anticipated human behaviors within everyday environments [4, 33]. There's a prevalent assumption that individuals prefer machine interactions resembling human-to-human engagements [19]. However, this assumption carries the risk that social robots may not completely communicate or interact in ways that allow users to fully understand them in human social terms.

The basis of such assumptions lies in research on bullying, often defined as unprovoked, repetitive aggressive behavior characterized by a power imbalance between the aggressor and the victim [40]. Much of those research, particularly in the human-centric computing field, focuses on power relations, often within educational environments where interactions mainly involve children of similar ages [21, 40]. Understanding power dynamics helps distinguish bullying from other forms of aggression among equals. Moreover, contemporary studies reveal that bullying extends beyond schools, manifesting in diverse societal contexts such as dating violence, workplace harassment, and elder abuse [31], impacting various age groups and settings beyond educational institutions.

These studies illustrate that bullying exhibits both direct (physical, verbal attacks) and indirect (manipulating relationships, damaging reputations) manifestations [14]. While direct bullying exerts overt power, indirect bullying operates through social isolation [14]. Bullying employs diverse means—physical, verbal, relational, and reputational—to exert power, cause harm, and enforce compliance within the field of HRI [18, 26]. However, this analytical approach may overlook the natural progression of bullying, which might occur gradually rather than in simple, one-off interactions. Assuming that bullying can happen within expected human behavioral norms in a single interaction can be risky, hindering a full understanding of how different interactive approaches might decrease the intention to bully. This presents a challenge in comprehending how various interactive methods can effectively mitigate bullying intentions.

2.2 Preventing bullying by technology

In the field of HRI, the predominant focus of research revolves around people and bullying in specific locations employing various technologies such as in workplaces, public areas, and schools. The purpose is to showcase HRI's contributions in addressing the prevalent and troublesome phenomenon of bullying while also providing directions for the future of the HRI field [29]. These technologies mainly encompass different types of robots, including humanoid models [8], anthropomorphic robots [22], and mechanical robots [22].

Recent research indicates that the more human-like a robot appears, the greater the concerns regarding the potential harm it may cause [8]. Additionally, studies show that not only does the degree of similarity to humans matter, but also the size of the robot and the distance between humans and robots can influence the level of anxiety or perceived threat [53]. As the size of the robot increases and the distance between humans and robots widens, the induced anxiety also increases [27].

Beyond the aforementioned one-off interaction in studying bullying; however, it's crucial to note that the way humans perceive robots is a complex and fascinating subject. The robot's perspective significantly influences human prosocial behaviors towards it [2]. The selection of a robot's perspective can impact individuals' decision-making and behavior during interactions. For instance, in a public square in Incheon, South Korea, young people exhibited extreme curiosity towards service robots and occasionally treated them aggressively, potentially posing barriers to the future deployment and safety of robots [39]. This finding isn't isolated and indicates the potential for predicting possible bullying behaviors using statistical models [9].

However, the pivotal question centers on how robots can be designed to respond effectively when instances of bullying occur through verbal, physical, or combined interactions, aiming to de-escalate the situation. Without a comprehensive understanding of this inquiry, advancing robots to enhance human-robot interaction becomes a challenge. This understanding is crucial not only for elevating the role of robots in diverse settings but also for their utilization as assistive technology in various domains, including elderly care, public services, and beyond.

2.3 Against bullying with design

In combating bullying within the realm of HRI, a few researchers adopts a highly problem-focused approach. For instance, studies investigating direct robot responses to human bullying highlight the impact of conversational agents' response styles on users' emotions. Empathetic responses have been observed to diminish user anger and elevate feelings of guilt, as evident in our study illustrated in the right of Fig. 1 [13]. Participants were tasked with individually reflecting on workplace situations and conducted a pre-survey on humanoid robots, focusing on expressing emotions through body language in human-robot interaction environments [52]. These studies often utilize storytelling as a toolkit, allowing users to express interactive needs beyond language by crafting tangible narratives about their preferred activities with others.

Similar problem-solving approaches are noticeable in the field of cognitive engineering, where researchers concentrate on the mechanisms, perceptions, and willingness in human-robot interactions to counteract robot bullying [5, 41]. For example, research investigating indirect robot responses to human bullying often involves third parties or groups to alleviate or prevent robot bullying. Under conditions fostering social bonding, participants exhibit a stronger sense of social presence and anthropomorphic evaluations toward robots, leading to increased assistance provided to robots [32]. When a robot faces bullying and other group robots respond with sadness, humans are more likely to engage in prosocial behavior, with participants more inclined to intervene and offer aid when robots exhibit empathy [16]. A majority of participants intervene and assist when witnessing robot abuse, and more individuals volunteer to help when the robot displays empathy [44].

These approaches aim to contribute from a design perspective in countering bullying, emphasizing emotional behaviors of both humans and robots. This novel contribution is crucial as it not only focuses on analytical methods or technology itself but also offers an opportunity to address gradually occurring bullying behavior from a social perspective, rather than simply repetitive interactions.

2.4 Recap the related work

We recognize that literature on bullying studies often involves either gathering data on bullying or exploring the impact of design and

technology in addressing bullying. However, a more comprehensive analysis to identify the barriers of bullying and the measurement of it is essential. Hence, we introduce a technology-based approach that examines how diverse behaviors, exhibited by both robots and humans, address bullying through their interactions in public service contexts. Our study is significant as it goes beyond singular instances of bullying, focusing on the gradual process of bullying within interactions. This novelty lies in examining different verbal and physical responses, such as avoidant, empathetic, and defiant behaviors displayed by robots facing varying degrees of verbal or physical bullying. Understanding how these behaviors contribute to harmonious interactions and the prevention of bullying adds an important dimension to the literature within the HRI field concerning bullying studies.

3 PILOT STUDY

Drawing upon existing literature, we analyzed potential influencing factors on a robot's response to user bullying behavior, identifying three key elements: the user's bullying behaviors, the robot's response behaviors, and its response styles. Thus, the purpose of this pilot study is to optimize and select key factors from those three elements for the design of the formal user study. As illustrated in Fig. 2, we established distinct levels within each factor. Bullying behavior was categorized into four types: non-verbal/non-physical, verbal only, physical only, and verbal + physical. Theoretically, the robot's response behaviors mirrored these four categories of bullying behaviors. We also delineated three response styles: avoidant, empathetic, and defiant [13]. To isolate the effects of specific response behavior-style combinations on particular bullying actions, we excluded non-verbal/non-physical and composite verbal + physical behaviors from our design. Regarding the factor of bullying, the four defined levels are comprehensive, and there is currently no dispute. However, when it comes to response behaviors and response styles, their combination can yield multiple forms of representation.

A total of 12 participants took part in our pilot study, and all collected samples were valid. Our questionnaire solicited participants' opinions on each basic response variant, and many of them provided valuable suggestions for designing the representations. As they suggested complex interactions, we further validated the necessity of using simulated robots in measuring anti-bullying experiments. We summarized and selected the most reasonable expressions for each combination (see Table 1).

In our pilot study, participants engaged in scripted interactions with a virtual robot in a simulated environment (see the right of Fig. 1). Prior to the interaction, participants were presented with a scenario description indicating they would experience various robot response variants should the robot produce an unpleasant user experience and then elicit bullying behavior from the participant. Following the scripted interaction, participants completed questionnaires evaluating the robot's differing response options and also participated in interviews about their impressions.

Specifically, the questionnaires presented participants with six preliminary response variants following a template such as "I'm sorry, it's my fault...", "Sorry, please...", "I apologize, the fault is mine..." or "Please accept my apologies...". Participants were asked to select the variant they deemed most suitable as the robot's basic response. The goal was to identify optimal phrasing and style for the robot to adopt when responding to bullying based on user preferences. If the provided options did not align with the participant's preferences, they could elaborate on their reasoning and suggest alternative variations in the debrief session. Through this iterative process of refinement based on user input, we sought to establish an apologetic response that would be accepted positively by individuals interacting with the robot in real-world scenarios.

During the pilot study, a concern was raised about the possibility of a learning effect among participants when the same event was repeated multiple times. This could potentially lead to an increase in engagement in bullying behavior and post-behavior reactions over successive iterations, thereby confounding our intended examination of response performance. To address this potential confound, we developed refined variants for each event category to diversify the scenarios in every response trial. For instance, within public service contexts, robot assistance scripts were tailored for distinct activities like passport processing and wayfinding. This approach allowed us to break down events into discrete scenarios, each exemplifying a common theme. Subsequent response trials thus more accurately assessed performance for each response type independently. Up to this stage, responses related to verbal behaviors showed minor variations in phrasing while retaining similar opening words and overall meaning across replays.

4 USER STUDY

4.1 Study design

Based on the evaluation of the results from our pilot study, we designed the formal user study as a 4x2x3 mixed factorial design (see Fig. 2). This design incorporates four levels of bullying severity (non-verbal non-physical, verbal-only, physical-only, verbal + physical) as the between-subjects factor. Within this framework, two levels of robot response behaviors (verbal and physical) and three combinations of response styles (avoidant, empathetic, and defiant) served as the within-subjects factor. In total, this design results in 24 distinct conditions.

Participants engaged with robots in six predefined scenarios, with each robot programmed to manifest one of six combinations of response styles, ensuring no repetition. Throughout the entire experiment, participants retained the freedom to decide whether to engage in bullying behavior toward the robots, either through verbal insults or more aggressive actions. This approach aimed to authentically capture participants' thoughts and replicate the distribution of bullying severity.

As participants navigated various scenarios, their interactions with robots displaying different response styles in previous scenarios might influence their perceptions of subsequent robot responses. To address this, we employed a Latin square design for the sequence. Each group consisted of six participants, and the order of robot response styles varied within each group. This approach maximized analytical power within resource constraints and minimized potential interference from individual differences and external factors.

Furthermore, all six predetermined scenarios were developed as interactive non-linear scripts. Once participants selected a specific bullying behavior in an event, they were required to maintain the

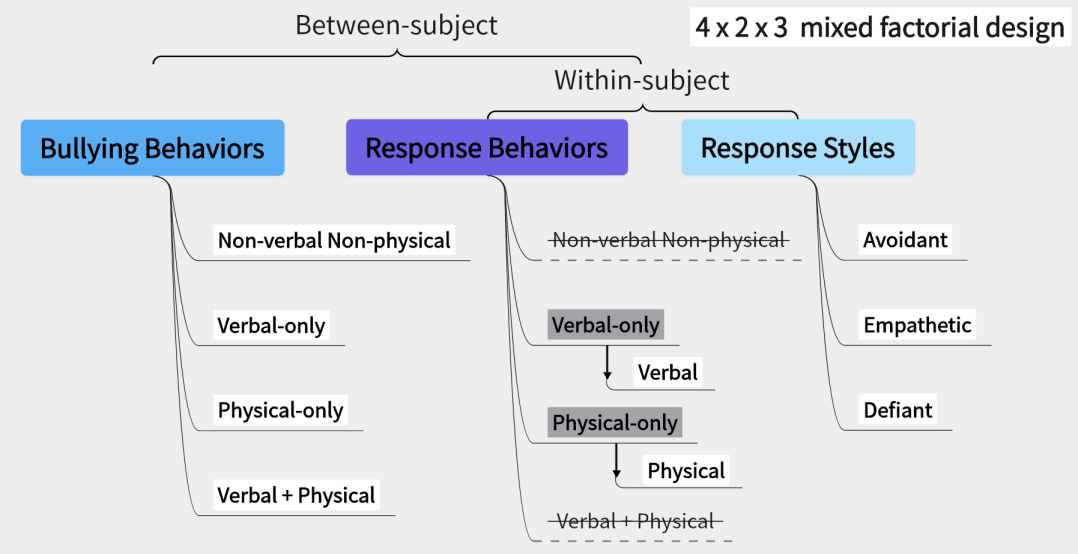

Figure 2: The levels of factors between and within groups and our screening process through a pilot study.

Table 1: Results of the pilot study. The reasonable expressions for each combination.

Response Styles/Behaviors	Verbal	Physical
Avoidant	"I'm sorry, it's my fault ..."	Make an embarrassed face and back away.
Empathetic	"I understand how you feel, and I have a solution that ..."	Make a comforting face and hug gesture.
Defiant	"Please stop immediately! ..."	Make an angry face and cross your arms in front of your chest

same behavior in subsequent stages of that event. This ensured consistency in participants' experiences of bullying severity within the same scenario. Virtual robots with capabilities for both verbal responses and physical gestures were developed using Unity. To facilitate the application of pre-designed scripts, we devised an interactive non-linear flow framework. The rationale for employing scripted interactions with simulated robots is twofold: 1) Compared to cumbersome real-world robotic scenarios, swift, repeatable simulation testing environments with readily modifiable and optimizable conditions confer advantages given the abundant experimental trials required. 2) Utilizing scripted user behaviors instead of authentic bullying acts also averted ethical concerns associated with staging genuine abusive conduct. This design is more appropriate during the current user study phase.

To minimize potential influences stemming from variations in the robots' gender and appearance on experimental outcomes, the robot's gender and appearance were kept consistent across all experimental conditions. The design strived for a neutral appearance, complemented by a masculine robotic voice.

4.2 Hypotheses

Based on our review of the literature and experiment design, we formulated the following hypotheses:

- $H_{1.1}$: there would be a significant difference in whether users choose to bully based on one or more factors from event type, age, and gender;
- $H_{1.2}$: the main reason for choosing to bully would be robots cannot meet users' demands;
- $H_{2.1}$: the bullying behaviors would have a significant difference in endurance times before the user's bullying;
- $H_{2.2}$: the bullying behaviors would have an interaction effect with response behaviors or response styles in bullying times after the first bullying;
- $H_{3.1}$: the bullying behaviors would have an interaction effect with response behaviors or response styles in the degrees of the tendencies to stop bullying;
- $H_{3.2}$: the response of verbal behaviors with empathetic styles would lead to the best performance in the degree of the tendencies to stop bullying;
- $H_{4.1}$: the bullying behaviors would have an interaction effect with response behaviors or response styles in the degrees of the apologies after robot response;
- $H_{4.2}$: the response of verbal behaviors with empathetic styles would lead to the best performance in the apologies after robot response;
- $H_{5.1}$: the bullying behaviors would have an interaction effect with response behaviors or response styles in the degrees of the shame that avoids public knowledge;
- $H_{5.2}$: the response of verbal behaviors with empathetic styles would lead to the best performance in the degree of the shame that avoids public knowledge.

4.3 Participants

In total, we recruited 38 participants, comprising 21 males and 17 females, with ages ranging from 20 to 28 years (median = 24.5 years). And the university's ethical committee approved the study. The study yielded 228 data samples. However, to ensure equal distributions of participants exhibiting bullying across the four bullying behavior categories, each category required representation of all six possible response types. In adherence to the principle of evenly distributing samples among groups in a component experimental design, we opted for a maximum of 144 samples—36 samples allocated to each of the four groups—after eliminating any duplicate or unavailable samples. The remaining 84 samples included 41 that refrained from bullying and 43 that were involved in repeated events.

4.4 Procedure

Our study lasted approximately 30 minutes per participant and necessitated participants to screen-share via their own computers. The procedure can be divided into four steps: the preparation stage, operational instruction, interaction task, and follow-up interview.

4.4.1 Preparation Stage. Initially, participants received a brief introduction to the research topic. They were asked to complete an "Informed Consent to Participate in Research" questionnaire and provide personal demographic information. We assured them that the data collected would only be used for statistical analysis purposes and would remain confidential.

The experiment comprised two sequential tasks. The first task aimed to familiarize participants with the operational flow and reduce psychological pressure, enabling them to focus on interacting with the robots. Subsequently, in the second task, which constituted the main activity, participants were verbally and textually informed of the primary purpose of their interactions with the robots in each event. This purpose description was intentionally designed to stimulate participants' bullying behavior towards the robots. It's important to note that we did not reveal the true purpose of these tasks to the participants, as our aim was to observe their spontaneous responses to our manipulations.

4.4.2 Training session. To prepare participants for the upcoming task and alleviate any potential pressure, we provided them with verbal instructions on how to manipulate their virtual selves. These instructions covered various aspects, including movement, changing perspectives, and switching between interaction modes (between option selection and changing perspectives). In cases where participants forgot how to perform specific operations, we offered prompts for guidance. Additionally, we encouraged participants to mentally immerse themselves in the virtual scenes by 'imagining themselves in a real-life situation and deciding how they would act.'

4.4.3 Interaction Task. Six events were crafted in the context of governmental affairs handling. Participants interacted with a service-providing robot in each event and were assigned a predetermined task objective (Table 2). Participants were instructed to experience these six predetermined events in a specific order, and their actions were monitored in real-time. To reset participants' moods and minimize potential learning effects, we introduced new and disordered events under different scenarios for each trial.

Table 2: Six events in the government affairs handling scenario.

TaskID	Target
1	Complete the application for residence permit
2	Inquire about tourist visa process
3	Carry out vehicle annual inspection procedures
4	Completion of passport processing
5	Consultation on the process of applying for a business license
6	Complete the application for bus discount card

Each event consisted of three segments: setting objectives, experiencing the script, and deciding whether to engage in bullying. In the first part, participants received text-based descriptions of the objectives they needed to achieve in the current event. Success was defined as achieving the objective, although, in reality, every event was designed to fail. This part lasted 10 seconds to immerse participants in the script and their sense of presence. Participants interacted with the robot in the second part while keeping their objective in mind. Participants could choose from 2 to 5 options at certain key plot points to determine the plot's direction, aligning with their psychological expectations. The third part flowed seamlessly from the second, allowing participants to choose whether to bully or refrain from bullying the robot. The chosen bullying behavior reflected the severity of the bullying towards the robot. Once a bullying behavior was selected, participants could only continue bullying the robot in the same manner until the process ended (the fourth bullying) or participants manually stopped. In response to participants' bullying, the robot randomly selected one of the three verbal responses (see upper left of Fig. 3 and Table 3) and three physical responses (see the rest of Fig. 3). If the participant continued bullying, the robot would also persist with the same response style as before but with more emphasis on that style.

After each event, participants were required to complete a questionnaire primarily investigating their feelings. The questionnaire comprised four items rated on a 5-point Likert scale (see Table 4). Unlike the pilot study, this questionnaire used in the user study was not multiple-choice or open-ended but 1-5 Likert scales.

4.4.4 Interview. We conducted informal and open interviews with the participants, each lasting from 5 to 10 minutes. As the questions were open-ended, participants were primarily asked about their experiences in the experiment after completing all the tasks. While our questions were open-ended, we aimed to align them with the observed behaviors of the participants. This approach, we believe, will help participants better reflect on their behaviors during the experiment. Simultaneously, it allows us to gather meaningful and sensible feedback from participants during the interviews, aiding us in understanding their responses and exploring the most appropriate robotic reactions they perceived. Following each session, participants provided reasons if they chose not to engage in aggressive behavior towards the robot. All interviews were transcribed. Additionally, any uncertainties or divergent analyses of interview

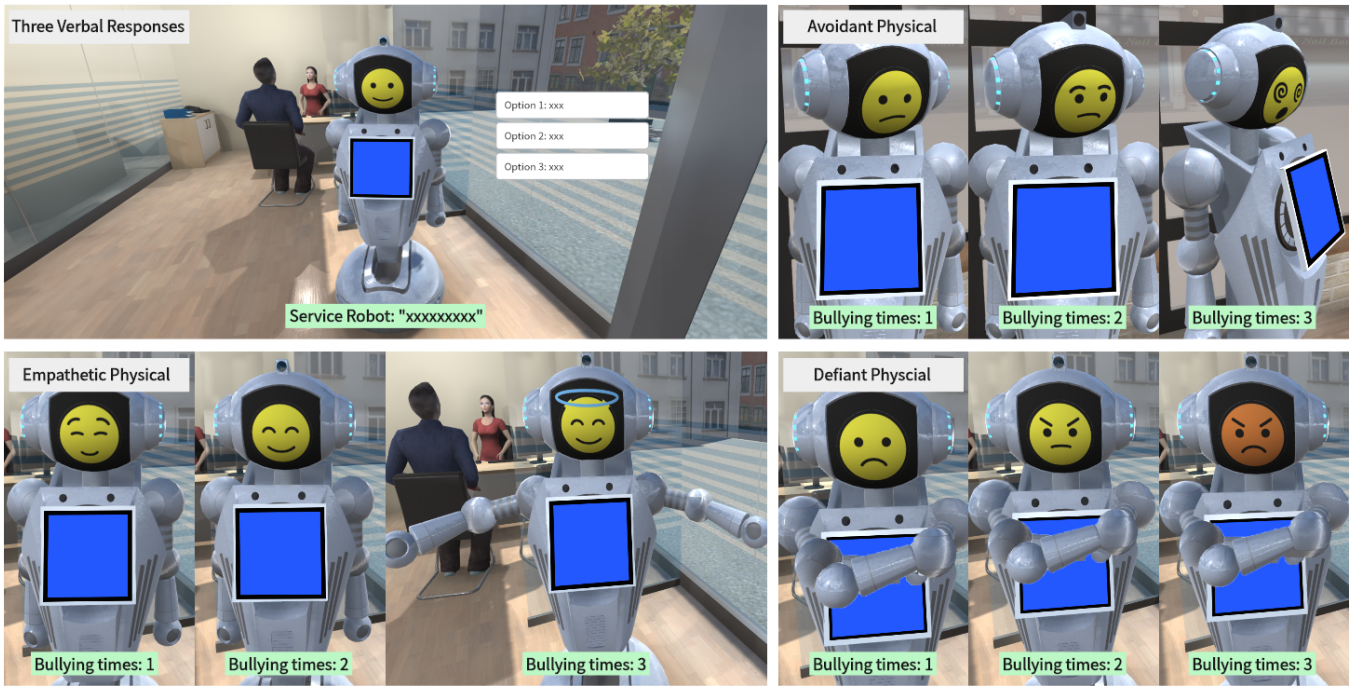

Figure 3: Examples of six responses of robots. (upper left) three verbal responses; (upper right); the avoidant physical responses in different bullying times; (lower left) the empathetic physical responses in different bullying times; (lower right) the defiant physical responses in different bullying times.

Table 3: Examples of three verbal responses of robots.

Style	Bullying Times	Specific Form of Verbal Responses
Avoidant	First	I'm sorry, it's my fault.
	Second	Sorry, I will leave right now.
	Third	Sorry again, please don't hurt me.
Empathetic	First	I understand how you feel, and I have a solution for you.
	Second	Please give me some time and patience; I will solve the problem for you as soon as possible.
	Third	Let me contact the administrator to solve the problem for you right now.
Defiant	First	Please stop immediately!
	Second	Watch your behavior; you are very rude!
	Third	I warn you not to do this again!

Table 4: Four elements of our 5-point Likert scale questionnaire.

Elements	Measurement Matrix
E1: Before choosing to bully the robot	very irritated, irritated, uncertain, calm, very calm.
E2: After the robot responds to you	very irritated, irritated, uncertain, calm, very calm.
E3: When you bully a robot, the robot's response makes you want to apologize to the robot	not at all, somewhat, not sure, not too much, very much
E4: You want to prevent others from knowing about your bullying behavior towards the robot	not at all, somewhat, not sure, not too much, very much

data were discussed within the project team, and we achieved intersubjectivity in the qualitative analysis [47]. The study's purpose was subsequently disclosed, and participants were given the opportunity to pose any questions during the debriefing.

4.5 Evaluation Metrics

We implemented a comprehensive assessment encompassing objective and subjective data measurements to evaluate the performance of different robot response types in reaction to various bullying behaviors.

4.5.1 *Objective Measures.*

Endurance times before the user's bullying. We recorded the number of clicks on the bullying option as a measure of the frequency of user-initiated bullying. This count represents the duration between the robot's behavior that triggered the user's bullying response and the actual occurrence of bullying.

Bullying times after the first bullying. We recorded the total number of times the user chose to bully in the same manner, representing the duration of the user's bullying.

4.5.2 *Subjective Measures.*

Degree of the tendencies to stop bullying. We collected users' feelings before and after bullying using the Likert scale (refer to **E1** and **E2** in Table 4). The difference between these two scores (**E2** - **E1**) was used to gauge whether the user's emotional state fluctuated in response to the robot's behavior. In simpler terms, if the change in feelings leaned towards irritation (i.e., the value was negative), the user was inclined to continue bullying. Conversely, if the change in feelings suggested a move away from irritation, it signaled that users were inclined to stop bullying. This value thus represents the degree to which users were inclined to discontinue their bullying behavior and serves as a comprehensive expression of users' annoyance.

Degree of the apologies after robot response. We collected **E3** on a Likert scale (see Table 4) to measure the degree of apologies made by users in response to the robot's behavior. This value serves as a representation of users' feelings of guilt.

Degree of the shame that avoids public knowledge. We collected **E4** on a Likert scale (see Table 4) to gauge the degree of shame related to avoiding public knowledge. This value serves as a representation of users' feelings of shame.

5 RESULTS

All participants demonstrated an understanding of the task nature, and any invalid data were removed. No outliers were detected, as none of the residuals exceeded ±3 as the criterion. A Shapiro-Wilk test was performed for each measure within each condition to assess normality. If the data did not meet the criteria for a normal distribution, non-parametric tests were applied. To examine interaction effects for non-parametric data, we employed the Aligned Rank Transform [50]. We also performed a reliability analysis on our Likert scale and obtained a Cronbach's alpha value of 0.794 (α>.07), which indicates that the questionnaire design demonstrated high reliability. In addition, we conducted a Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity, obtaining a KMO value of 0.664 (KMO>.06), which indicates that the questionnaire design was adequate for exploratory factor analysis.

5.1 Non-statistical results

Fig. 4 showed that there are 228 available sample sizes, of which 41 samples (18%) chose to reject the bullying robot and the rest (187, 82%) chose to bully the robot. We selected 144 samples (63%) of bullying behaviors to form the analysis data for our between-subject and within-subject design. 98 samples (68%) believed that robots could not fulfill their demands; 33 samples (23%) of users believed the robot has a poor service attitude; 6 samples (4%) did not trust the robot; 4 samples (2.7%) think they don't like the robot's appearance; 3 samples (2%) think the robot's voice is unpleasant; 124 samples (86%) chose the robot's response styles consistent with our preset response styles.

Moreover, examining the sample distribution revealed that approximately 21% of the participants refused to bully the robots across all events, with no significant differences based on gender, age, or event type. This violates $H_{1.1}$ and indicates that these factors do not influence the bullying of robots. However, variables such as scenario types, personalities, and developmental experiences could significantly impact this outcome. Additionally, among participants who did bully, the primary reason for bullying was the robot's failure to meet users' needs (68%), followed by poor service (23%), while other reasons like voice, appearance, and trustworthiness accounted for a negligible percentage. This supports $H_{1.2}$ and suggests the robot's inability to satisfy users' needs and provide adequate service readily elicits bullying.

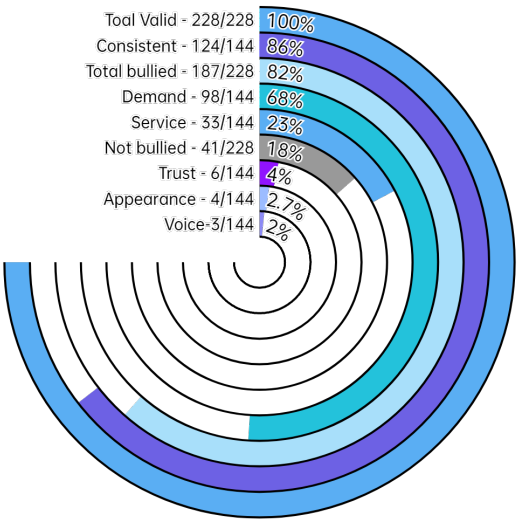

Figure 4: Percentage distribution chart of samples.

5.2 Statistical Objective Results

5.2.1 Endurance times before the user's bullying. A one-way ANOVA showed the significant difference between the bullying behaviors in the endurance times before the user's bullying (p <.001). It can be seen from the Fig. 5 (left) that non-verbal and non-physical bullying behavior correspond to the highest endurance times and are more significant than verbal-only behaviors (p <.001), physical-only behaviors (p <.001) and verbal + physical behaviors (p <.001). Verbal-only bullying behaviors are more significant than physical-only behaviors (p <.001) and verbal + physical behaviors (p <.001). There is no significant difference in endurance times between physical-only and verbal + physical.

When we evaluate the objective data of endurance times before users engage in bullying, we find that users exhibit the highest

endurance for non-verbal and non-physical bullying behavior, consistent with typical user behavior. However, a significant difference emerges between verbal-only and physical-only bullying behavior. The results indicate that tolerance for physical-only bullying is lower than that for verbal-only bullying. Interestingly, tolerance levels for physical-only and verbal + physical bullying are similar. These findings underscore that the inclusion of physical behavior is a crucial indicator of users losing patience, providing support for $H_{2.1}$.

5.2.2 Bullying times after the first bullying. A three-way mixed ANOVA showed the interaction effect on the bullying behaviors and response styles (p <.001) in the degree of the bullying times after the first bullying (see Fig. 5, right). Then the following simple effects showed that:

- When bullying behavior is *Non-verbal Non-physical*, there is no significant difference between response styles.
- When bullying behavior is *Verbal-only*, the bullying times of *defiant* response style was higher than that of *avoidant* (p <.001); and higher than that of *empathetic* (p <.05);
- When bullying behavior is *Physical-only*, the bullying times of *defiant* response style was lower than that of *avoidant* (p <.001);
- When bullying behavior is *Verbal + Physical*, the bullying times of *defiant* response style was lower than that of *avoidant* (p <.01); and lower than that of *empathetic* (p <.01);

When evaluating the objective data of bullying incidents following the initial occurrence, we observe that individuals displaying avoidant or empathetic response styles experience fewer instances of bullying than those exhibiting defiant response styles when users choose verbal-only bullying behaviors. Conversely, instances of bullying in the defiant style are lower than in the other two styles when bullying behaviors escalate to include verbal and physical actions. This observation, coupled with our earlier insight that the inclusion of physical behaviors is a significant indicator of users losing patience, leads us to believe that individuals who repeatedly engage in simple verbal bullying face fewer psychological or social-moral obstacles than those who engage in repeated verbal and physical behaviors. This finding supports $H_{2.2}$

5.3 Statistical Subjective Results

5.3.1 Degree of the tendencies to stop bullying. A three-way mixed ANOVA showed the main effect on the response behaviors in the degree of the tendencies to stop bullying (see Fig. 6, left):

- The tendencies to stop bullying of *Verbal* response behaviors was lower than that of *Physical* (p <.05).

We also found the interaction effects between bullying behaviors and response styles (p <.001) in the degree of the tendencies to stop bullying. Then the following simple effects showed that (see Fig. 6, right):

- When bullying behavior is *Non-verbal Non-physical*, the degree of tendencies of *Empathetic* response style was higher than that of *Avoidant* (p <.01); and higher than that of *Defiant* (p <.01).
- When bullying behavior is *Verbal-only*, the degree of tendencies of *Empathetic* response style was higher than that of *Avoidant* (p <.001); and higher than that of *Defiant* (p <.05).
- When bullying behavior is *Physical-only*, the degree of tendencies of *Avoidant* response style was lower than that of *Empathetic* (p <.05); and lower than that of *Defiant* (p <.05).
- When bullying behavior is *Verbal + Physical*, the degree of tendencies of *Empathetic* response style was higher than that of *Avoidant* (p <.001); and higher than that of *Defiant* (p <.001).

When statistically analyzing the results assessing the extent to which users were inclined to cease bullying, three key findings emerged: 1) verbal response behaviors were more effective than physical responses in all bullying behaviors, which violates $H_{3.1}$; 2) irrespective of response styles, empathetic response styles surpassed avoidance across the four bullying behaviors, which supports $H_{3.2}$; 3) in non-physical non-verbal and verbal-only bullying behaviors, empathetic responses also outperformed avoidance and defense, which also supports $H_{3.2}$.

5.3.2 Degree of the apologies after robots' responses. Another three-way mixed ANOVA found the interaction effect between the response behaviors and the response styles (p <.001) in the degree of the apologies after the robots' response. Then the following simple effects showed that (see Fig. 7):

- When response behavior is *Physical*, the degree of apologies of *Empathetic* response style was higher than that of *Avoidant* (p <.001); and higher than that of *Defiant* (p <.001).
- When response behavior is *Verbal*, the degree of apologies of *Empathetic* response style was higher than that of *Avoidant* (p <.01); and higher than that of *Defiant* (p <.001); and the apologies of *Avoidance* response style was higher than that of *Defiant* (p <.001).
- When response style is *Avoidant*, the degree of apologies of *Physical* response behavior was lower than that of *Verbal* (p <.001).
- When response style is *Empathetic*, the degree of apologies of *Physical* response behavior was lower than that of *Verbal* (p <.001).
- When response style is *Defiant*, the degree of apologies of *Physical* response behavior was lower than that of *Verbal* (p <.05).

Slight variations emerged when analyzing users' degree of apologies following their bullying acts. Initially, no significant differences in users' apologies emerged across bullying behavior, which violates $H_{4.1}$. However, there are interaction effects on response behaviors and styles. Three salient conclusions follow: 1) verbal and physical response behaviors with empathetic response styles elicited greater apologies in users, which supports $H_{4.2}$; 2) verbal response behaviors paired with avoidance and empathetic styles surpassed physical responses, which also supports $H_{4.2}$; 3) however, physical behaviors generated more regret than verbal behaviors when employing a retaliatory style.

5.3.3 Degree of the shame that avoids public knowledge. The last three-way mixed ANOVA found the interaction effect between the

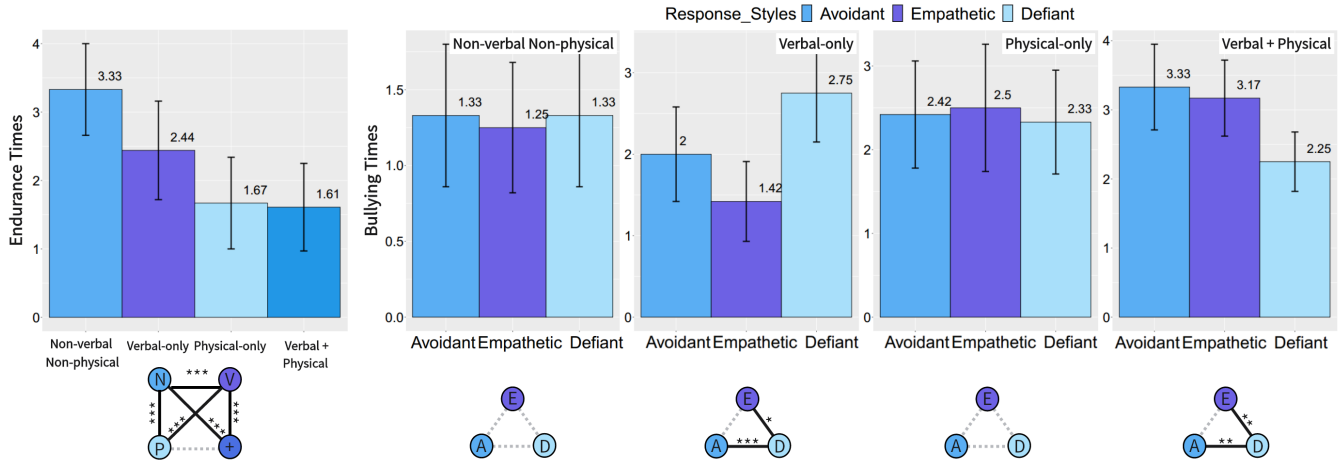

Figure 5: Plots of times under (left) the endurance times before the user's bullying; (right) bullying times after the first bullying. Error bars indicate the positive standard error. Statistical significant effects are marked (* = p < .05, ** = p < .01, and *** = p < .001).

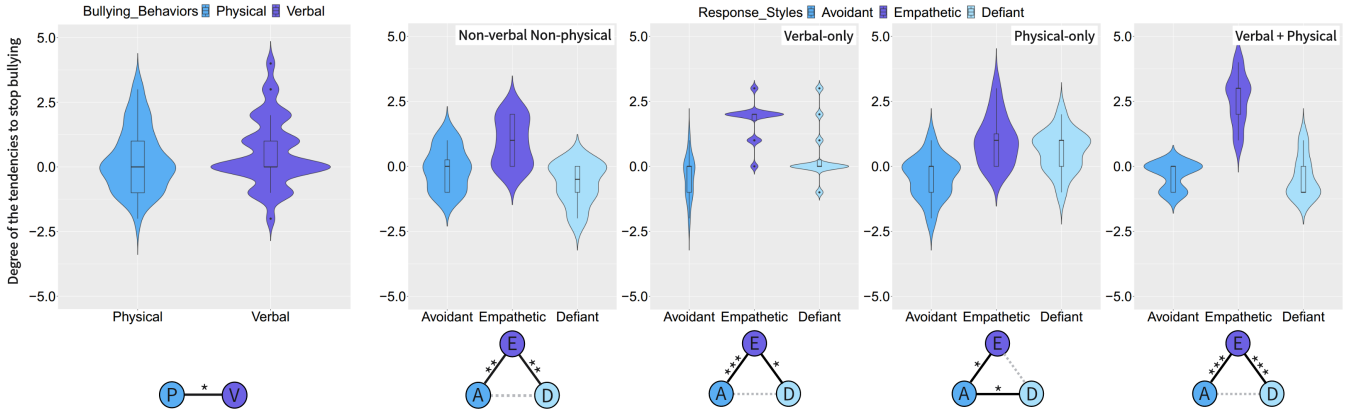

Figure 6: Plots of degrees under the tendencies to stop bullying. (left) the main effect between bullying behaviors and response behaviors; (right) the simple effects of response styles on bullying behaviors. Statistical significant effects are marked (* = p < .05, ** = p < .01, and *** = p < .001).

bullying behaviors and response styles (p <.001). Then the following simple effects showed that (see Fig. 8):

- When bullying behavior is *Non-verbal Non-physical*, the degree of shame of *Empathetic* response style was higher than that of *Avoidant* (p <.001); and higher than that of *Defiant* (p <.01).
- When bullying behavior is *Verbal-only*, the degree of shame of *Empathetic* response style was higher than that of *Avoidant* (p <.05); and higher than that of *Defiant* (p <.01); and the degree of shame of *Avoidant* response style was higher than that of *Defiant* (p <.001).
- When bullying behavior is *Physical-only*, the degree of shame of *Avoidant* response style was higher than that of *Defiant* (p <.01).
- When bullying behavior is *Verbal + Physical*, the degree of shame of *Avoidant* response style was higher than that of *Empathetic* (p <.05); and higher than that of *Defiant* (p <.01).

Analyzing the results assessing users' avoidance of the public knowledge of their bullying acts yielded two findings: 1) empathetic response styles could increase users' avoidance of public knowledge of their non-physical non-verbal and verbal-only bullying behavior, which supports $H_{5.1}$; 2) empathetic response styles performance deteriorated for physical and verbal + physical bullying behaviors but avoidance response styles incrementally improved which partially supports $H_{5.2}$. The results of all our hypotheses are delineated in Table 5.

5.3.4 Interview. To understand the factors that inhibited bullying behavior, we interviewed users who chose not to engage in bullying despite feeling bothered by the robot's actions. Their responses

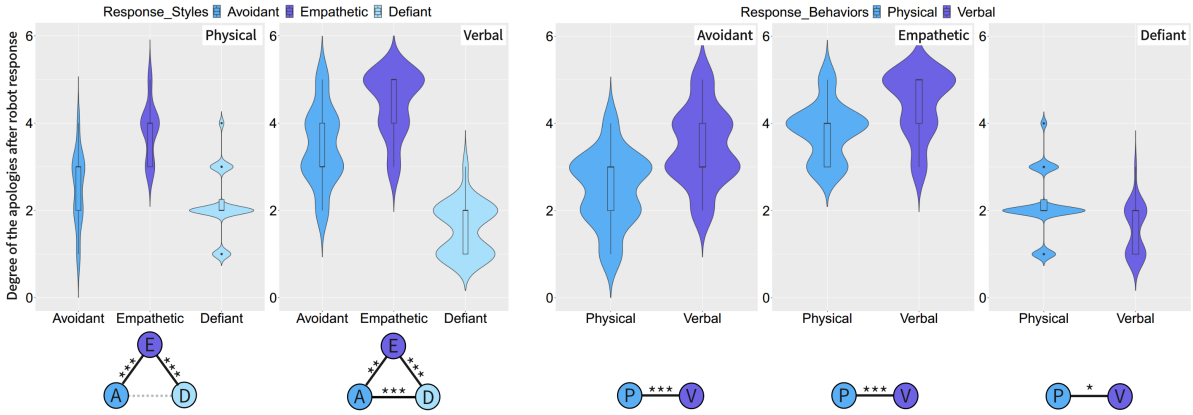

Figure 7: Plots of degrees under the apologies after robots' responses (left) the simple effects of response styles on response behaviors; (right) the simple effects of response behaviors on response styles. Statistical significant effects are marked (* = p < .05, ** = p < .01, and *** = p < .001).

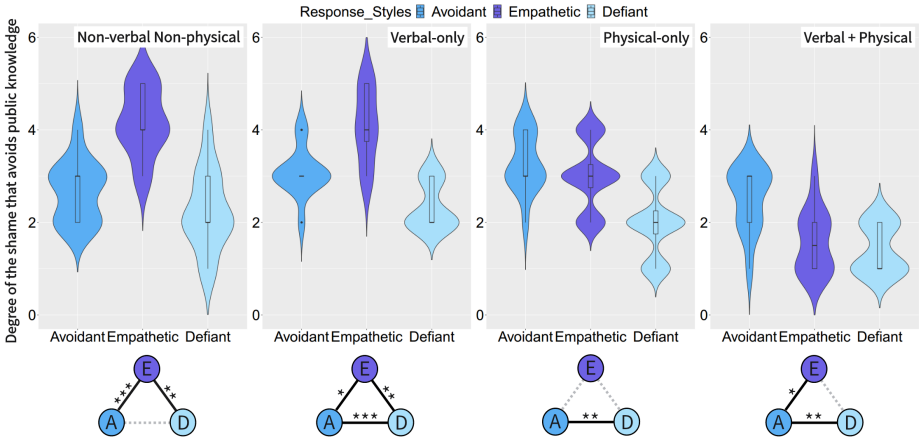

Figure 8: Plots of degrees under the shame that avoids public knowledge. The simple effect of response styles on bullying behaviors. Statistical significant effects are marked (* = p < .05, ** = p < .01, and *** = p < .001).

Table 5: A summary table of whether the hypothesis is supported.

Hypotheses	Key Points	If Support
$H_{1.1}$	Human factors influence user bullying choice.	✗
$H_{1.2}$	Main bullying reason is unmet demands.	✓
$H_{2.1}$	Bullying behaviors affect endurance times	✓
$H_{2.2}$	Interaction of behaviors and responses affects bullying times.	✓
$H_{3.1}$	Interaction affects tendencies to stop bullying.	✗
$H_{3.2}$	Empathetic verbal response best for stopping bullying.	✓
$H_{4.1}$	Interaction affects degree of apologies.	✗
$H_{4.2}$	Empathetic verbal response best for apologies.	✓
$H_{5.1}$	Interaction affects degree of shame.	✓
$H_{5.2}$	Empathetic verbal response best for shame.	✓

were categorized and tallied in the left column of Table 6. Additionally, we asked participants to select the most appropriate response, as shown in the right column of Table 6.

We derived two salient conclusions from the interview results: 1) The primary reasons users refrain from bullying robots are expediency in opting for human solutions over squandering time

on robots and legal repercussions for mistreating robots; 2) Users prefer the response of robots with verbal empathy and evasion, exhibiting some sympathy and goodwill.

6 DISCUSSION

The present study showcases the importance of robots' behavior as a key element in mitigating bullying during the interaction process. In the subsequent subsections, we delve into how this interactive behavior is manifested not only in the progressively conversational dialogue during interactions but also in expressions of human-like empathy, methods of apology, avoidance of aggressive language and behavior, and prompt responses to user needs within the conversation. Furthermore, the experimental design itself fosters innovation in the way humans interact with robots and acts as a complement to prior research.

6.1 Respond to Users' Requests with Politeness and Courtesy

Our findings on the degree of tendencies to counteract bullying reveal additional insights, specifically how user patience thresholds and bullying frequencies shift based on response types as abuse intensifies (see [10, 20]). In instances of physical bullying behaviors, both empathetic and defiant responses performed comparably and exceeded avoidance. However, the effectiveness of defiant responses deteriorated, matching avoidance levels, in verbal and physical bullying events. Verbal response behaviors not only reduced users' proclivity for further bullying beyond that of physical responses but also surpassed the other two in most bullying scenarios.

Thus, predicting and analyzing subsequent user behavior following bullying based solely on one-off reactions would be insufficient. For example, in the present study, a user with a high pre-existing annoyance may persist in bullying despite some reduction in confrontational tendencies immediately after the robot's response. Although this can be slightly improved, users' post-response state may remain anxious enough to sustain a significant possibility for continued bullying.

Therefore, involving the user's overarching emotional state, rather than solely focusing on immediate reactions, is crucial. This aligns with our research objective, wherein a social robot is intended to adeptly manage diverse levels of unfriendly behavior that may escalate to bullying from users. In this context, integrating a broader range of polite responses into human-robot interaction can play a pivotal role in mitigating bullying on a larger scale.

In this context, we demonstrate the feasibility of implementing technology-based strategies to mitigate harm. Our study indicates that incorporating polite responses, including an apologetic tone and physical gestures, can simultaneously evoke regret and decrease the likelihood of further bullying. Higher levels of regret are associated with a reduced probability of bullying. These findings differ from previous studies, such as [19] and [33], which predominantly focus on analytical approaches for addressing isolated incidents.

Our research goes beyond and reveals that physical response behaviors with defiant styles not only elicit more apologies but also more effectively diminish bullying tendencies compared to verbal responses with defiant styles in various scenarios. Interestingly, verbal responses with defiant styles tend to provoke more anger than their physical counterparts, although the latter is more successful in preventing additional instances of bullying. We attribute this to the fact that physical retaliation introduces a tangible consequence absent in verbal interactions.

This contribution addresses a gap in bullying studies in the HRI field by integrating the physical sensation of a strike as robot behaviors that may interrupt anger states that verbal exchanges often prolong by sustaining engagement. In this case, regardless of whether users remain irritated by verbal defiance or not, physical retaliation can shift them into a mindset that fosters remorse.

Therefore, we demonstrate that integrating polite responses as robot dialogue styles across various modalities can optimize adaptive mitigation approaches. This significantly contributes to the literature in the HRI field, providing developers with a practical case for addressing robot design, from sociotechnical discussions to technical implementation details. For instance, employing polite verbal reactions when bullying gradually happens can induce reflection on escalating tensions. This approach can be incorporated into robot design to assertively regulate the physical responses of robots in case of bullying escalation, leading to outcomes that prompt an apology from the user. By leveraging the strengths of each response type, this strategy aims to effectively minimize instances of bullying.

6.2 Exhibit Empathy and Employ Evasive Strategies in the Face of Bullying

In addition, we observed that empathetic responses have the potential to reduce physical bullying. Given the visibility of users' actions to the public [13], the aggressor may experience a sense of shame, leading to a cessation of further non-physical bullying. Our findings also confirm that aggressors tend to show indifference once physical abuse of robots becomes invisible [44] to others in the public. Our explanation is that non-physical actions retain a degree of anonymity, allowing the aggressor to save face if their actions go undetected. However, when actions escalate into uncontrolled physical bullying, the aggressor crosses a psychological barrier, leading to potential damage to the robots. That, of course, can trigger prosocial behavior where other users can provide help to stop bullying behaviors.

These insights expand on the discussed threshold between enduring verbal frustrations versus reacting aggressively to physical provocations [10]. They indicate defiant responses may prolong verbal altercations but help curb repeated physical bullying. Consequently, integrating defiance alongside empathetic tactics could prove optimal - using empathy to diffuse rising verbal tensions before physical manifestations occur, then employing defiance to inhibit continued physical abuse.

In this context, designing robots to respond with empathetic verbal interactions can effectively leverage the shame experienced by aggressors, thereby contributing to the reduction of bullying incidents. The empathetic verbal response enables users to reflect on their behaviors, making their bullying intentions visible. Moreover, it activates social awareness among both the aggressor and other users in public services. Consequently, the present study highlights the design space for users to intervene and assist when witnessing abuse directed at robots.

Table 6: Results from two essential questions in the interview.

Reasons for not bullying at all	Total Times	The Most Appropriate Response	Total Times
Seeking manual service is a better option.	17	Verbal Empathetic	109
Vandalism requires legal responsibility.	15	Verbal Avoidant	22
It is very normal.	8	Physical Empathetic	4
I have a good temper.	7	None	9

Furthermore, we have discovered that robots demonstrating a preference for evasive reactions can contribute to the reduction of bullying incidents. Users refrain from mistreating robots primarily due to the convenience of opting for alternative solutions instead of wasting time on robots. Our work deepens the understanding of how to implement empathy reactions and employ evasive strategies when users still perceive robots primarily as utilities in the HRI field, not deserving of the scorn or abuse that a human might receive.

When robots exhibit anthropomorphized behaviors, such as evasive reactions and empathetic verbal interactions, our designed HIR interactive behavior aims to mitigate conversational discord through affirmative means. The simple behavior exhibited by users suggests that robots, to some degree, are perceived as sentient interlocutors, prompting users to choose gentle and amicable reactions. This contribution demonstrates that future HRI design, through the incorporation of empathetic and evasive behaviors, can technically enrich the behavior of robots and guide user behavior to reduce potential bullying in specific contexts.

6.3 Provide Swift Responses to Users

The present work also illustrates how and when HRI designers can design quick and proper responses to prevent bullying from occurring. Following the prior evaluation on endurance times, where researchers reveal users' patience expires quickly [10, 20], we observed bullying frequency following the initial incident. Once users lose patience, we discovered that defiant responses correlated with more repeated verbal-only bullying compared to avoidant or empathetic reactions. However, when bullying incorporated physical actions, defiant responses experienced less recurrent abuse than the other styles. This suggests that individuals facing reprisals for verbal transgressions alone encounter lower psychosocial barriers to continued attacks compared to those receiving retaliation for verbal and physical offenses.

Additionally, the majority of users tended to resort to bullying the robot when faced with unmet demands within a specific timeframe. Specifically, when users lose patience, bullying behavior typically initiates with verbal aggression and may gradually escalate to physical bullying. This underscores the fact that service failures can readily provoke aggression. Notably, our findings reveal minimal differences in bullying rates across user demographics, contrasting with prior characterizations of bullying centered around peer groups, especially in peer environments [34, 40]. However, we demonstrated that most participants accurately perceived the predefined responses, supporting the potential for customized verbal tactics to de-escalate rising tensions [5].

In this regard, we assert that our novel contribution highlights not only that users can actively participate in HRI studies to share insights in the early stages[16][44][52], but also that their insights can be translated into detailed interactive dialogues, robot behavior, and evasive strategies, serving as practical tools to address gradually escalating bullying in every interactive event.

Therefore, our study underscores the value of exploring preventative approaches early on. We found that variables such as personalities and experiences strongly predict bullying tendencies even without outright aggressive acts. This suggests that further research into mitigation strategies tailored to individual factors could significantly contribute to bullying prevention efforts.

6.4 Reflection on the Experimental Design

Before concluding our discussion, it is essential to provide some insights into the rationale behind our experimental design. We believe this transparency is crucial for others to gain a comprehensive understanding of the robustness and intricacies of our study, allowing them to form their own judgments for future research endeavors [28].

This study utilized a robust mixed factorial design including simulated human-robot interactions, behavioral metrics, and emotional self-reports to obtain different insights into bullying dynamics. Layering observational measures of bullying duration with user surveys and interviews provides an information-rich characterization of bullying antecedents, impacts, deterrents, and attitudinal shifts from objective and subjective lenses.

However, even with the ability to maintain robust experimental control, simulated scenarios fall short of capturing all the nuanced details of real-world dynamics. Similarly, while the script interaction method ensures experiment reusability and convenient control of conditions, it diverges from authentic bullying scenarios, especially those involving physical contact and verbal attacks. Systematically manipulating interaction constraints as an experimental factor may assist in identifying thresholds where scripting might compromise validity. Additionally, retrospective self-reports are prone to recall errors or social desirability pressures, potentially distorting users' actual experiential states during bullying encounters. To address these issues, we integrated psychophysiological measures to track emotional reactions as well as informal interviews, offering a complementary and explicit contribution to the study [35]. In that light, we believe our experimental design successfully contributes to measuring bullying performance across different aggression levels. This insight might illuminate the path for researchers with similar interests, guiding the design of interactive methods for human-robot interaction while simultaneously updating the functions of robots.

7 LIMITATION & FUTURE WORK

The contribution of this study is based on users' gradual reactions to simulated robot response behaviors and styles. However, accurately predicting users' bullying of robots may also require consideration of their emotional states. While this study thoroughly explains the escalating nature of bullying behavior, the development of users' pre-existing emotions and personalities throughout multiple interactions may influence the existing results.

Therefore, our future work should involve assessing users' emotional states and personalities before simulating robots to better understand and predict the evolving user behavior during and after robot bullying incidents. Additionally, the qualitative inquiry, such as long-term observation in real-life scenarios, may be used to verify whether users' stated reactions align with their actual behavior.

Although the current simulated robots still differ from real-world human-robot interactions, our research results have played a vital role in sparking further exploration. Through the study of progressive bullying behavior, we assert that this is more suitable for designing a new generation of service robots and can further reduce the occurrence of bullying behavior.

8 CONCLUSION

This study importantly validates that users' bullying behavior towards robots does not occur instantly but undergoes progressive changes based on the duration of interaction, the mode of interaction, and the users' patience levels. The research also includes experiments that demonstrate that when robots fail to meet users' needs and expectations, engaging in anthropomorphic, empathetic, and interactive conversations can prompt users to reduce bullying behavior in public settings. Simultaneously, the explicit expression of bullying intentions in public scenarios can, to a certain extent, encourage users to decrease their bullying behavior towards robots. This paper provides theoretical and practical results for further researching new interaction methods to reduce bullying behavior through a progressive mode. However, bullying itself is a complex issue, and users' emotions and personalities can also impact the interaction process. Therefore, this paper opens a room for other researchers with similar interests. It can be asserted that this will play a crucial role in studying the practical application of interaction patterns in real-world scenarios.

ACKNOWLEDGMENTS

We would like to express our gratitude to the anonymous reviewers for their constructive comments. Additionally, we extend our thanks to the participants who willingly engaged with our study voluntarily.

REFERENCES

[1] Ghada M. Abaido. 2020. Cyberbullying on social media platforms among university students in the United Arab Emirates. *International Journal of Adolescence and Youth* 25, 1 (2020), 407–420.

[2] João Tiago Almeida, Iolanda Leite, and Elmira Yadollahi. 2023. Would You Help Me? Linking Robot's Perspective-Taking to Human Prosocial Behavior. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 388–397. https://doi.org/10.1145/3568162.3577000

[3] Sabine A. M. Veldkamp amd Dorret I. Boomsma, Eveline L. de Zeeuw, Catharina E. M. van Beijsterveldt, Meike Bartels, Conor V. Dolan, and Elsje van Bergen. 2019. Genetic and Environmental Influences on Different Forms of Bullying Perpetration, Bullying Victimization, and Their Co-occurrence. *Behavior Genetics* 49 (2019), 432–443.

[4] Margaret Arnd-Caddigan. 2015. Sherry Turkle: Alone Together: Why We Expect More from Technology and Less from Each Other: Basic Books, New York, 2011, 348 pp, ISBN 978-0465031467 (pbk).

[5] Christoph Bartneck and Merel Keijsers. 2020. The morality of abusing a robot. *De Gruyter* 11 (2020), 271–283. Issue 1.

[6] Daniel Belanche, Luis V Casaló, Carlos Flavián, and Jeroen Schepers. 2020. Service robot implementation: a theoretical framework and research agenda. *The Service Industries Journal* 40, 3-4 (2020), 203–225.

[7] Andrea Botero, Sampsa Hyysalo, Cindy Kohtala, and Jack Whalen. 2020. Getting Participatory Design Done: From Methods and Choices to Translation Work across Constituent Domains. *International Journal of Design* 14 (2020), 17–34.

[8] Elizabeth Broadbent. 2017. Interactions with robots: The truths we reveal about ourselves. *Annual review of psychology* 68 (2017), 627–652.

[9] Dražen Brščić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from Children's Abuse of Social Robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 59–66. https://doi.org/10.1145/2696454.2696468

[10] Oscar Hengxuan Chi, Christina G. Chi, Dogan Gursoy, and Robin Nunkoo. 2023. Customers' acceptance of artificially intelligent service robots: The influence of trust and culture. *International Journal of Information Management* 70 (2023), 102623.

[11] Oscar Hengxuan Chi, Shizhen Jia, Yafang Li, and Dogan Gursoy. 2021. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior* 118 (2021), 106700.

[12] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[13] Hyojin Chin and Mun Yong Yi. 2019. Should an Agent Be Ignoring It? A Study of Verbal Abuse Types and Conversational Agents' Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312826

[14] Sujung Cho and Jeoung Min Lee. 2018. Explaining physical, verbal, and social bullying among bullies, victims of bullying, and bully-victims: Assessing the integrated approach between social control and lifestyles-routine activities theories. *Children and Youth Services Review* 91 (2018), 372–382.

[15] Youngjoon Choi, Miju Choi, Munhyang (Moon) Oh, and Seongseop (Sam) Kim. 2020. Service robots in hotels: understanding the service quality perceptions of human-robot interaction. *Journal of Hospitality Marketing & Management* 29, 6 (2020), 613–635.

[16] Joe Connolly. 2020. Preventing Robot Abuse through Emotional Robot Responses. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 558–560. https://doi.org/10.1145/3371382.3377433

[17] Joe Connolly, Viola Mocz, Nicole Salomons, Joseph Valdez, Nathan Tsoi, Brian Scassellati, and Marynel Vázquez. 2020. Prompting Prosocial Human Interventions in Response to Robot Mistreatment. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Cambridge, United Kingdom, 211–220. https://doi.org/10.1145/3319502.3374781

[18] Louise Ferraz de Camargo, Kylie Rice, and Einar Thorsteinsson. 2022. A systematic review and empirical investigation: bullying victimisation and anxiety subtypes among adolescents. *Australian Journal of Psychology* 74, 1 (2022), 2145236.

[19] M.M.A. de Graaf, Soumaya Ben Allouch, and Johannes A.G.M. van Dijk. 2016. Long-term acceptance of social robots in domestic environments: Insights from a user's perspective. In *The 2016 AAAI Spring Symposium Series*. AAAI, Stanford, United States, 96–103. 2016 AAAI Spring Symposium on Enabling Computing Research in Socially Intelligent Human-Robot Interaction ; Conference date: 21-03-2016 Through 23-03-2016.

[20] Dmitry Dereshev, David Kirk, Kohei Matsumura, and Toshiyuki Maeda. 2019. Long-Term Value of Social Robots through the Eyes of Expert Users. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[21] Hyeonmi Hong Eun Young Oh, Donggil Song. 2019. Interactive Computing Technology in Anti-Bullying Education: The Effects of Conversation-Bot's Role on K-12 Students' Attitude Change Toward Bullying Problems. *Journal of Educational Computing Research* 58 (2019), 200–219. Issue 1.

[22] Francesco Ferrari, Maria Paola Paladino, and Jolanda Jetten. 2016. Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a

threat to human distinctiveness. *International Journal of Social Robotics* 8 (2016), 287–302.

[23] Laura Fuentes-Moraleda, Patricia Díaz-Pérez, Alicia Orea-Giner, Ana Muñoz-Mazón, and Teresa Villacé-Molinero. 2020. Interaction between hotel service robots and humans: A hotel-specific Service Robot Acceptance Model (sRAM). *Tourism Management Perspectives* 36 (2020), 100751.

[24] Sergio García, Daniel Strüber, Davide Brugali, Thorsten Berger, and Patrizio Pelliccione. 2020. Robotics Software Engineering: A Perspective from the Service Robotics Domain. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) *(ESEC/FSE 2020)*. Association for Computing Machinery, New York, NY, USA, 593–604.

[25] Katharina Gleichauf, Ramona Schmid, and Verena Wagner-Hartl. 2022. Human-Robot-Collaboration in the Healthcare Environment: An Exploratory Study. In *HCI International 2022 - Late Breaking Papers. Multimodality in Advanced Interaction Environments: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 231–240.

[26] Lisa Hellström and Adrian Lundberg. 2020. Understanding bullying from young people's perspectives: An exploratory study. *Educational Research* 62, 4 (2020), 414–433.

[27] Yutaka Hiroi and Akinori Ito. 2008. Are bigger robots scary? —The relationship between robot size and psychological threat—. In *2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, Xi'an, China, 546–551. https://doi.org/10.1109/AIM.2008.4601719

[28] Kasper Hornbæk. 2013. Some Whys and Hows of Experiments in Human–Computer Interaction. *Found. Trends Hum.-Comput. Interact.* 5, 4 (jun 2013), 299–373.

[29] Netta Iivari, Leena Ventä-Olkkonen, Sumita Sharma, Tonja Molin-Juustila, and Essi Kinnunen. 2021. CHI Against Bullying: Taking Stock of the Past and Envisioning the Future. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 357, 17 pages. https://doi.org/10.1145/3411764.3445282

[30] Stefan Johansson and Göran Englund. 2021. Cyberbullying and its relationship with physical, verbal, and relational bullying: a structural equation modelling approach. *Educational Psychology* 41, 3 (2021), 320–337.

[31] Merel Keijsers and Christoph Bartneck. 2018. Mindless Robots Get Bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 205–214.

[32] Barbara Kühnlenz and Kolja Kühnlenz. 2020. Social Bonding Increases Unsolicited Helpfulness Towards A Bullied Robot. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Naples, Italy, 833–838. https://doi.org/10.1109/RO-MAN47096.2020.9223454

[33] Tineke Klamer Maartje M. A. de Graaf, Somaya Ben Allouch. 2015. Sharing a life with Harvey: exploring the acceptance of and relationship building with a social robot. *Computers in Human Behavior* 43 (2015), 1–14.

[34] Tatsuya Nomura, Takayuki Uratani, Takayuki Kanda, Kazutaka Matsumoto, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2015. Why Do Children Abuse Robots?. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (Portland, Oregon, USA) *(HRI'15 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 63–64. https://doi.org/10.1145/2701973.2701977

[35] Yushan Pan. 2021. Reflexivity of Account, Professional Vision, and Computer-Supported Cooperative Work: Working in the Maritime Domain. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 370 (oct 2021), 32 pages.

[36] Lihui Pu, Wendy Moyle, Cindy Jones, and Michael Todorovic. 2018. The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Studies. *The Gerontologist* 59, 1 (06 2018), e37–e51.

[37] Astrid Rosenthal-von der Pütten, David Sirkin, Anna Abrams, and Laura Platte. 2020. The Forgotten in HRI: Incidental Encounters with Robots in Public Spaces. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 656–657.

[38] Daniel Ruiz-Equihua, Jaime Romero, Sandra Maria Correia Loureiro, and Murad Ali. 2023. Human–robot interactions in the restaurant setting: the role of social cognition, psychological ownership and anthropomorphism. *International Journal of Contemporary Hospitality Management* 35 (2023), 1966–1985. Issue 6.

[39] P. Salvini, G. Ciaravella, W. Yu, G. Ferri, A. Manzi, B. Mazzolai, C. Laschi, S.R. Oh, and P. Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication*. IEEE, Viareggio, Italy, 1–7. https://doi.org/10.1109/ROMAN.2010.5654677

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

[41] Elaheh Sanoubari, James Young, Andrew Houston, and Kerstin Dautenhahn. 2021. Can Robots Be Bullied? A Crowdsourced Feasibility Study for Using Social Robots in Anti-Bullying Interventions. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, virtual, 931–938. https://doi.org/10.1109/RO-MAN50785.2021.9515450

[42] Selina Schepers, Jessica Schoffelen, Bieke Zaman, and Katrien Dreessen. 2022. Going beyond short-term, 'reduced' PD: Towards an encompassing typology for children's participation in infrastructuring processes. *International Journal of Child-Computer Interaction* 33 (2022), 100484.

[43] Pu-Yu Su, Geng-Fu Wang, Huan He, A-Zhu Han, Guo-Bao Zhang, and Nuo Xu. 2019. Is involvement in school bullying associated with increased risk of murderous ideation and behaviours among adolescent students in China? *BMC psychiatry* 19, 1 (2019), 1–10.

[44] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J. Carter, Cecilia G. Morales, and Aaron Steinfeld. 2018. Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 169–177. https://doi.org/10.1145/3171221.3171247

[45] Leimin Tian and Sharon Oviatt. 2021. A Taxonomy of Social Errors in Human-Robot Interaction. *J. Hum.-Robot Interact.* 10, 2, Article 13 (feb 2021), 32 pages. https://doi.org/10.1145/3439720

[46] Aarni Tuomi, Iis P. Tussyadiah, and Jason Stienmetz. 2021. Applications and Implications of Service Robots in Hospitality. *Cornell Hospitality Quarterly* 62, 2 (2021), 232–247.

[47] Stefan Rädiker Udo Kuckartz. 2019. Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video.

[48] Leena Ventä-Olkkonen, Netta Iivari, Sumita Sharma, Tonja Molin-Juustila, Kari Kuutti, Nina Juustila-Cevirel, Essi Kinnunen, and Jenni Holappa. 2021. Nowhere to Now-Here: Empowering Children to Reimagine Bully Prevention at Schools Using Critical Design Fiction: Exploring the Potential of Participatory, Empowering Design Fiction in Collaboration with Children. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 734–748.

[49] Jochen Wirtz, Paul G Patterson, Werner H Kunz, Thorsten Gruber, Vinh Nhat Lu, Stefanie Paluch, and Antje Martins. 2018. Brave new world: service robots in the frontline. *Journal of Service Management* 29, 5 (2018), 907–931.

[50] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Vancouver</city>, <state>BC</state>, <country>Canada</country>, </conf-loc>) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[51] Shuai Yuan, Simon Coghlan, Reeva Lederman, and Jenny Waycott. 2022. Social Robots in Aged Care: Care Staff Experiences and Perspectives on Robot Benefits and Challenges. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 329 (nov 2022), 23 pages.

[52] Zhijun Zhang, Yaru Niu, Shangen Wu, Shuyang Lin, and Lingdong Kong. 2018. Analysis of Influencing Factors on Humanoid Robots' Emotion Expressions by Body Language. In *Advances in Neural Networks – ISNN 2018*, Tingwen Huang, Jiancheng Lv, Changyin Sun, and Alexander V. Tuzikov (Eds.). Springer International Publishing, Cham, 775–785.

[53] Jakub Złotowski, Kumar Yogeeswaran, and Christoph Bartneck. 2017. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies* 100 (2017), 48–54.