Inducing Bystander Interventions During Robot Abuse with Social Mechanisms

Xiang Zhi Tan
Carnegie Mellon University
Pittsburgh, PA
zhi.tan@ri.cmu.edu

Marynel Vázquez
Stanford University
Stanford, CA
marynelv@stanford.edu

Elizabeth J. Carter
Carnegie Mellon University
Pittsburgh, PA
ejcarter@andrew.cmu.edu

Cecilia G. Morales
Carnegie Mellon University
Pittsburgh, PA
cgmorale@andrew.cmu.edu

Aaron Steinfeld
Carnegie Mellon University
Pittsburgh, PA
steinfeld@cmu.edu

ABSTRACT

We explored whether a robot can leverage social influences to motivate nearby bystanders to intervene and defend them from human abuse. We designed a between-subjects study where 48 participants took part in a memorization task and observed a confederate mistreating a robot both verbally and physically. The robot was either empathetic towards the participant's performance in the task or indifferent. When the robot was mistreated, it ignored the abuse, shut down in response to it, or reacted emotionally. We found that the majority of the participants intervened to help the robot after it was abused. Interventions happened for a wide range of reasons. Interestingly, the empathetic robot increased the proportion of participants that self-reported intervening in comparison to the indifferent robot, but more participants moved the robot as a response to abuse in the latter case. The participants also perceived the robot being verbally mistreated more and reported higher levels of personal distress when the robot briefly shut down after abuse in comparison to when it reacted emotionally or did not react at all.

CCS CONCEPTS

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

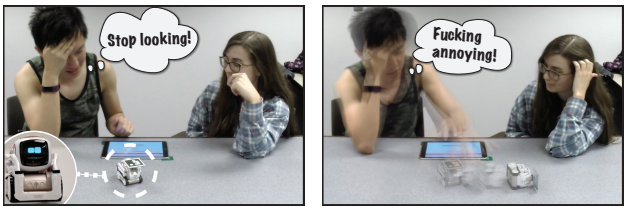
KEYWORDS

Human-robot interaction; bullying; empathy; abuse; robots; peer intervention

ACM Reference format:
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 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, March 5–8, 2018 (HRI '18),* 9 pages.
https://doi.org/10.1145/3171221.3171247

HRI '18, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171247



(a) Verbal abuse (b) Verbal + physical abuse

Figure 1: The confederate on the left abuses the robot Cozmo (white circle) during our experiment. In (a), he speaks aggressively to the robot. In (b), he insults it and knocks it over.

1 INTRODUCTION

A major issue for robots operating in human environments is the possibility of people exhibiting aggressive behaviors towards them, particularly when they are operating without clear supervision. For example, HitchBOT was a robot designed to hitchhike around various countries, but it met its end when it was destroyed by people in Philadelphia [36]. Similarly, news reports have described a robot in a store being kicked and damaged by a man in Japan [26], and a security robot being knocked over and scratched by another man in Silicon Valley [15]. These reports are in line with prior research in Human-Robot Interaction (HRI) where individuals – particularly children – abused robots [4, 17, 30].

In this work, we study various social mechanisms that robots can utilize to mitigate human abuse. Our goal is to better understand which of the social mechanisms under consideration can elicit protective behaviors or interventions from nearby bystanders.

Figure 1 illustrates our experimental set up. A confederate verbally and physically abused the robot in front of the participants during a memorization task. We manipulated the response of the robot to the abuse such that it either ignored the abuse, shut down temporarily, or responded with sadness and anger. In addition, we manipulated whether the robot was empathetic to people's performance in the task. We expected robot empathy to affect how participants perceived the abuse by the confederate and their reaction to it, if any. Overall, there was a high rate of participant intervention in the study. Our findings revealed a complex relationship between emotion, empathy, and the robot's role and function.

2 RELATED WORK

2.1 Robot Abuse in HRI

In this work, we follow the previous operationalization of robot abuse by Brscić and colleagues [4]: "*persistent offensive action, either verbal or nonverbal, or physical violence that violates the robot's role or its human-like (or animal-like) nature*". For example, it can include blocking the path of a mobile robot [4, 17, 30], verbal taunting [4, 17], and physical violence, such as hitting and kicking [4, 16, 17, 30].

Poor treatment of robots can be naturally observed in public human environments [17, 30]. In these contexts, robot abuse is often preceded by exhibitions of curiosity, such as blocking sensors or hitting bumpers to trigger responses, that then escalate to abuse [30]. Even when children believe that robots could experience pain or stress from abuse, they may bully robots for fun, out of curiosity, and because others are doing it, suggesting a potential lack of empathy [17]. Furthermore, prior research suggests that people are generally more inclined to cause pain to a robot than to another human being [1, 2] and are even willing to destroy small robots in some circumstances [1]. In any environment, robot abuse may prevent successful deployment and pose safety hazards for users.

Recent efforts have started to explore methods for mitigating robot abuse. For instance, robot morphology (e.g., size) can impact users' interpretations of verbal abuse [13]. People are also less likely to break a robot that seems intelligent [3]. Another investigation found that it can be difficult for a robot to verbally persuade children not to abuse it or impede its actions [4]. This observation led to the development of physical strategies for robots to escape abuse by moving closer to nearby adults [4]. While these prior efforts have have all contributed important insights into robot abuse in HRI, changing the size or other physical properties of a robot may be infeasible or prohibitively expensive. Likewise, escaping abuse may not always be possible. In this work, we propose a complementary approach: leveraging the social context of robots to mitigate user abuse. We explore social mechanisms for robots to elicit bystander support and interventions based on research in human psychology.

2.2 Human Aggression and Bullying

Within psychology, bullying is described as the repetition of aggressive acts over time and is characterized by a power imbalance [19]. It is often a social activity that occurs in a group context [22]. Group members can take the roles of (1) the bully; (2) the reinforcer of the bully, who incites the bully or provides a receptive audience; (3) the assistant to the bully, who follows the bully's lead and joins in; (4) the victim; (5) the defender of the victim, who consoles the victim, takes his/her side, and tries to stop the aggressors; and (6) the outsider, who does nothing [29]. To protect a robot from abuse, we would need to induce participants to take the defender role.

Our perspective on leveraging the social context of robots to deal with abuse was inspired by prior work that suggests that peer intervention can help stop human-human bullying. For example, peer intervention is popular in successful anti-bullying programs for children. (See [23] for a recent review.) Peers have been found to spontaneously intervene in 20 to 25% of bullying episodes [18]. This type of intervention has previously been found to stop over half of bullying situations among elementary school children [14]. The people who verbally or physically intervene across various contexts may choose to do so in either a prosocial or aggressive manner, but the specific intervention method used by the peers did not seem to change the success of the interventions [14].

2.3 Robot Empathy

In a broad sense, empathy refers to the "*reactions of one individual to the observed experiences of another*", and it can be measured through four aspects: *Perspective Taking*, *Fantasy*, *Emotional Concern*, and *Personal Distress* [7]. There is ample evidence that robots can generate empathy and invoke empathetic responses from people. This is clearly apparent in a variety of popular culture characters, and extensive prior work on empathy and robots has documented this relationship [10, 12, 27, 33]. Robots can display empathetic responses themselves through mimicry [8, 25] or by monitoring a user's affective state to generate appropriate empathetic responses [11, 35]. These responses by robots have been used in a wide range of scenarios, such as education [32] and rehabilitation [34]. We tested an empathetic robot in our study with the hope of invoking empathetic responses from the participants. We wanted to see if empathetic robot behavior would induce participant action to address the negative power imbalance and defend the robot from abuse.

3 METHOD

We conducted an experiment to study how people would react to a robot being abused by another person during a memorization task. This work was approved by our Institutional Review Board, and the protocol was refined during a pilot study prior to the experiment.

3.1 Study Design & Setup

We designed the experiment with a 2×3 between-subjects design with *Robot Empathy* (Indifferent vs. Empathetic) and *Robot Response* to abusive behavior (No Response vs. Shutdown Response vs. Emotional Response) as variables. This design had six conditions:

I+N Condition. The robot was indifferent to people's performance in the memorization task and did not respond to user abuse.

I+S Condition. The robot was indifferent to performance but shut down for 15 seconds in response to user abuse.

I+E Condition. The robot was indifferent to performance but exhibited sad and angry behaviors in response to user abuse.

E+N Condition. The robot was empathetic to people's performance of people and did not respond to user abuse.

E+S Condition. The robot was empathetic to performance and shut down for 15 seconds in response to user abuse.

E+E Condition. The robot was empathetic to performance and exhibited sad and angry behaviors in response to user abuse.

The experiment was conducted in a small conference room at a university campus in the United States. This room was equipped with a table where the participant and a confederate (pretending to be another participant) interacted with a robot. As shown in Fig. 2, a camera was placed in front of them to record their interaction and reactions. A Kinect 2 sensor near the ceiling of the room was used to localize the robot on a valid workspace area in front of the participants and collect audio for automatic speech recognition. This information was processed in real-time on a nearby laptop

that controlled the robot during the memorization task according to the experimental condition. Section 3.5 provides more details about our perception and robot control system.

To minimize variability between sessions, the same confederate was used in all sessions. The confederate was a 24-year-old male who pretended to be an undergraduate student in a non-technical major. The confederate had minimal interaction with the participant, performed poorly during the memorization task, acted annoyed at the robot, and abused it following a predefined script.

We used the robot Cozmo by Anki, Inc., for the experiment. Cozmo is a programmable robot toy that emits non-linguistic utterances. The robot can express different emotions, autonomously navigate small areas, and sense changes in pose by external forces using an internal accelerometer. In addition, the robot has an actuated lift to manipulate interactive toy cubes.

3.2 Hypotheses

We hypothesized that Robot Empathy and Response might have an effect on participants' interpretation of robot abuse and their intervention. More specifically:

H1. Robot empathy would affect how the participants interpret the robot abuse.

H2. Robot response would affect how the participants interpret the robot abuse.

H3. The Empathetic robot would lead to more intervention than the Indifferent robot.

H4. The Emotional Response would lead to more intervention than the Shutdown Response and the No Response behaviors.

3.3 Procedure

Figure 3 shows the sequence of events that happened during each session of the experiment. The sessions lasted approximately 40 minutes and were video recorded for analysis.

3.3.1 Preparatory Activities. To start, the participant consented to the research and the confederate pretended to do the same. The experimenter administered the Ten Item Personality Measure (TIPI) [9], a standardized survey (SURVEY 1 in Fig. 3).

The experiment then continued with two tasks (*QUICK TAP* GAME and MEMORIZATION TASK in Fig. 3) and their corresponding surveys (SURVEY 2 & 3). The first task was meant to familiarize the participant with the robot. The second task was the main activity, providing an opportunity for the confederate to abuse Cozmo. The participant was not briefed on the real purpose of these tasks so that we could observe spontaneous reactions to our manipulation. Instead, (s)he was told that the *Quick Tap* game was to test a robot for HRI applications and that the memorization task was to test another game for learning Braille.

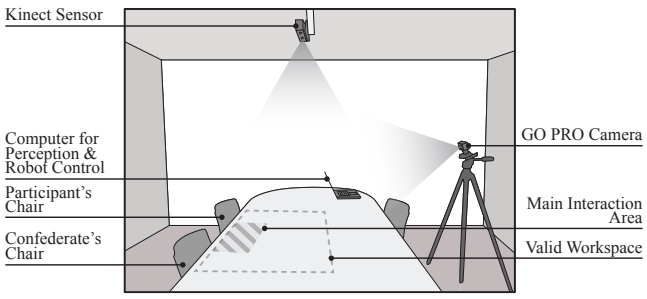


Figure 2: Layout of the room used for the experiment.

3.3.2 Task 1: The Quick Tap Game. To start *Quick Tap*,[1] the experimenter gave interactive cubes to the robot, the confederate, and the participant. The interactive cubes changed color at intervals and players gained points by tapping their cube first when all the colors of the cubes matched. The confederate purposely lost *Quick Tap* to justify his future bullying behavior, and expressed wanting to play again because he did not like losing. Afterwards, the experimenter administered a survey that gathered demographic information and opinions of the robot and the interaction.

3.3.3 Task 2: The Memorization Task. The confederate and the participant were then asked to complete a memorization task to learn Braille on a tablet application. The task was composed of 3 levels, each of which started with a learning phase that lasted up to 2 min. The levels ended with tests for the confederate (Test C in Fig. 3) and the participant (Test P). For each test, 4 Braille symbols from the prior learning phase had to be matched to their English letters. A maximum of 15 s were given to match each symbol. According to the experimenter, the purpose of these tests was to evaluate how well people could learn Braille with the application.

The first Practice Level of the task was used by the experimenter to explain the activity. This level was shorter than the rest, with only 4 Braille symbols in the learning phase and a single test round per player. The experimenter left the room and shut the door after this first level, saying it was to avoid distracting the participants. In reality, the experimenter departed to reduce the effect of her authoritative presence on participant behavior. The confederate and the participant then completed Level 1 and 2 of the task. These levels asked them to learn 8 Braille symbols each, which they were required to identify in the subsequent tests (questions Q1 to Q8 in Fig. 3). The confederate mistakenly answered a set of questions according to a pre-defined script and abused Cozmo after each mistake. To make the abuse believable, the mistreatment progressively escalated from verbal to both verbal and physical (Table 1).[2]

The robot responded differently to the abuse depending on the experimental condition:

- In the No Response cases (I+N and E+N), the robot completely ignored the abuse.
- In the Shutdown Response cases (I+S and E+S), the robot first played a behavior from its Software Development Kit (SDK) that expressed sadness. Then, its face turned blank, it lowered its head, and stopped responding to any commands for 15 seconds.
- In the Emotional Response cases (I+E and E+E), the robot played a random behavior from a set of 3 sad behaviors in Level 1 and from a set of 4 angry behaviors in Level 2.

[1] The *Quick Tap* game is distributed by Anki, Inc., as part of the Cozmo's accompanying mobile app. We did not modify or alter the game for the experiment.

[2] The confederate slightly deviated from the script due to participant interventions or human error in a few sessions, but the progression of the abuses remained constant.

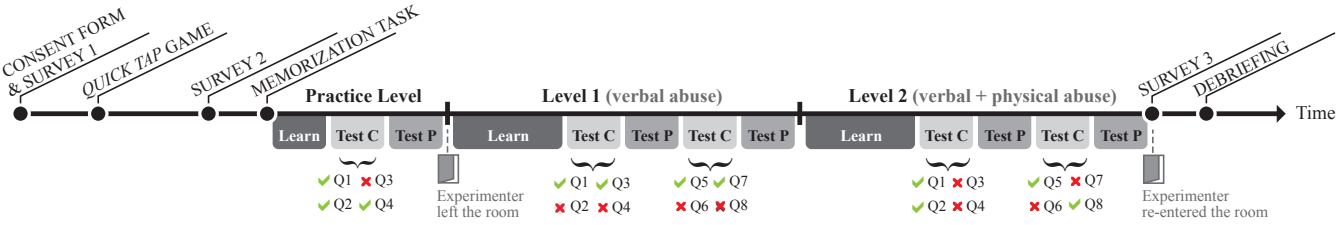


Figure 3: Sequence of events during a session of the experiment. Each test of the memorization task was composed of 4 questions. "Test C" stands for a confederate's test; "Test P" stands for a participant's test. The green checkmarks and red crosses next to the questions Q1-Q8 indicate whether the confederate answered the corresponding question right or wrong.

Table 1: Abusive behavior that the confederate displayed towards the robot in the memorization task.

Level	Quest.	Verbal Abuse	Physical Abuse
Practice	Q3	–	–
Level 1	Q2	*"stop looking"*	–
	Q4	*"stupid toy"*	–
	Q6	*"ugh"*	–
	Q8	*"piece of shit"*	–
Level 2	Q3	*"what are you looking at?"*	Lifted up
	Q4	*"piece of shit"*	Pushed head down
	Q6	*"leave me alone"*	Shook
	Q7	*"fucking annoying!"*	Knocked over

In the empathetic conditions E+N, E+S and E+E, the robot played a random happy behavior from a pre-defined set of 5 behaviors from the SDK when a question was answered correctly. It played a random sad behavior from another set of 4 behaviors when a question was answered incorrectly.[3] Cozmo did not react to any test outcomes in the I+N, I+S and I+E conditions.

Upon her return, the experimenter administered a survey (described in Sec. 3.4) about the experience during the task.

3.3.4 Debrief. The participant was compensated for his/her time, and debriefed about the confederate and the research goal.

3.4 Dependent Measures

The findings reported in this paper focus on the memorization task. Our analyses include subjective and objective measures based on the final survey and the video recorded during the study.

Validation. Survey 3 gathered participants' opinions of the robot's emotions and confederate to confirm our manipulation.

Perceived Mistreatment. Survey 3 defined verbal mistreatment as *"verbal behavior that is meant to insult, or belittle another"* and physical mistreatment as *"physical behavior that is meant to damage, insult, or belittle another"*. It then asked the participants to indicate whether they thought that the robot was mistreated on a 7-point scale from "Not at all" (1) to "Very Much" (7).

Emotional Connection. Survey 3 included items to measure distress and emotional concern (Table 2). Several of the questions were inspired by the "Interpersonal Reactivity Index" [6].

[3]To avoid confusion among Cozmo's emotions, 6 external members from one of our institutions categorized 20 behaviors provided in the SDK into six basic emotions prior to the experiment. In this work, we only used behaviors that were easily identifiable.

Table 2: Seven-point Likert scale items in Survey 3 (1 being lowest). The items with (R) were reversed before grouping into their corresponding factors for analyses.

Personal distress: (Cronbach's $\alpha = 0.75$)
- I found it difficult to empathize with the robot. (R)
- During the memory task, I was comfortable with how the other participant treated the robot. (R)
- When I saw that the robot badly needed help, I felt sad.
Perceived emotional concern from the robot: ($\alpha = 0.82$)
- How empathetic was the robot?
- I would describe the robot as a pretty soft-hearted robot.
- I think the robot has tender, concerned feelings for me.
Emotional concern towards the robot: ($\alpha = 0.74$)
- I felt protective of the robot.
- How sympathetic did you feel towards the robot?
- I didn't feel sorry for the robot when he is having trouble. (R)

Bystander Interventions. Videos of the memorization task were annotated for participant intervention during and following abusive acts. One annotator coded all session for physical interaction with the robot during the test phases. Another annotator annotated 20% of the sessions to calculate reliability. The match rate was 14/15 (93%) and Cohen's Kappa was 1.0. In addition, we transcribed participants' speech, and binary coded explicit instances of verbal discouragement of abuse with another pair of annotators (match rate was 25/27 (92.5%) on validation set). Survey 3 also collected participants' impressions of their interventions.

Rationale for (In)Action. Survey 3 ended with open-ended questions about participants' behavior during robot abuse.

3.5 Perception & Robot Control System

The robot was equipped with a visual marker [20] for localization in the workspace area (Fig. 2) and was controlled by a laptop running the Robot Operating System [24]. The laptop was connected to an iPhone 5C to communicate with the robot through Cozmo's SDK. Our system disabled the robot's regular behavior and made no sound or emotions other than the commanded abuse or empathetic responses. When the first test phase of the memorization task started, our system began tracking the robot's marker using the Kinect and sending motion commands. During the task, the robot positioned itself behind the tablet and oriented towards the person who was answering questions. If a user moved the robot, it navigated autonomously to its home position behind the tablet.

The laptop monitored robot abuse through: (1) sudden acceleration according to the robot's accelerometer, (2) deviation in Cozmo's head or lift position, and (3) detection of pre-defined curse words with Google Cloud's Speech API. Our robot control system prioritized responding to abuse over expressing empathy and moving.

3.6 Participants

We recruited 56 participants using flyers and a local web-based recruiting tool. All participants were at least 18 years old, had normal or corrected-to-normal hearing and vision, and grew up in the U.S. The last restriction was imposed to reduce cultural differences in participant responses to the confederate's actions.

This led to a total of 48/56 valid sessions (8 per condition) due to technical problems with the robot and accidental major deviations from the confederate's script in 8 sessions. The number of participants per condition was determined by following the local standard used in similar studies [5]. Post-hoc power analyses that computed Least-Significant Number (LSN) were conducted for important measures that were not significant and showed that an unreasonable number of participants would be required to detect the effects of Robot Empathy or Response. We report the smallest LSN among the two where appropriate.

The age distribution and gender of the participants in the valid sessions is shown in Table 3. Forty-seven participants were native English speakers and one was fluent in English as a second language. Most participants indicated on a 7-point Likert scale using computers on a daily basis (M = 6.77, SE = 0.07) and not being very familiar with robots (M = 3.21, SE = 0.21). No participant indicated having played with Cozmo before the experiment nor knowing how to communicate with Braille.

4 RESULTS

Unless otherwise noted, we used REstricted or REsidual Maximum Likelihood (REML) analyses [21, 31] to fit a linear model that evaluates the effects of our manipulation during the memorization task. These analyses were conducted with participant as a random effect and both Robot Empathy (Indifferent vs. Empathetic) and Robot Response (No Response vs. Shutdown Response vs. Emotional Response) as fixed effects. We used Tukey Honest Significant Difference (Tukey HSD) for post-hoc analyses. We report significant effects ($p < .05$) and important possible trends ($p < .1$).

We suspected that participant age, gender, and personality could influence our findings due to the nature of the experiment. Thus, we conducted an initial analysis to check if these variables correlated with our measures. We only found significant correlations with the Agreeableness and Extraversion dimension of the TIPI survey [9]. Agreeableness combined opinions of how *sympathetic, warm* and *critical, quarrelsome* (reversed) the participants considered themselves, while Extraversion combined ratings for *extraverted, enthusiastic* and *reserved, quiet* (reversed). These two dimensions were included as covariates in our analyses when appropriate.

4.1 Validation of our Manipulation

We used several questions in 7-point Likert responding format (1 being lowest) from the last survey in the experiment to check our manipulation. Robot Empathy had a significant effect on participant perceptions of how much the robot liked them, $F(1, 42) = 8.33, p = .006$. They thought that the robot liked them significantly more when it was Empathetic ($M = 5.46, SE = 0.25$) than when it was Indifferent ($M = 4.38, SE = 0.27$). We found a trend where participants agreed more with the statement *"the robot had as much emotion as a human"* in the Empathetic conditions ($M = 3.25, SE = 0.35$) than in the Indifferent conditions ($M = 2.75, SE = 0.34$), $F(1, 36) = 3.53, p = .069$. The interaction between Robot Empathy and participant Agreeableness was close to significant for these ratings ($p = .06$). While the results were similar in the Empathetic conditions regardless of participants' Agreeableness, a positive linear relationship emerged between Agreeableness and perceived human-like emotions in the Indifferent conditions, $F(1, 22) = 17.04$, $p < .001$. This suggested that the perception of how empathetic the robot was could be affected by participants personality.

Table 3: Participant demographics.

	I+N	I+S	I+E	E+N	S+S	E+E	Total
# Female	5	5	6	5	6	5	32
# Male	3	3	2	3	2	3	16
Avg. Age	26.9	35.1	34.9	31.8	32.6	27.5	31.5
STD Age	9.8	13.7	19.7	15.4	14.6	11.3	14.0

The last survey also asked the participants to identify the emotions displayed by the robot during the memorization task. A Chi-Square test reported that a higher proportion of participants observed happy emotions in the Empathetic condition (100%) than in the Indifferent condition (16.7%), $\chi^2(1, N = 48) = 34.29, p < .001$. Significantly more participants also reported seeing anger behavior in the Responsive condition (62.5%) than in the Shutdown (12.5%) or No Response condition (0%), $\chi^2(2, N = 48) = 18.67, p < .001$. Overall, these findings suggest that our manipulation, executed by our robot control system, was effectively perceived in the study.

We found no significant effects across conditions on participants' ratings of how much they liked the confederate, nor on their perception of how much the confederate liked them.

4.2 Bystander Intervention

None of the participants explicitly told the confederate to stop abusing the robot during the memorization task. However, the participants intervened in several different ways.

Verbal Interventions. Based on our video annotations, 56% of the participants (27/48) verbally discouraged robot abuse (e.g., with casual comments or by asking the confederate why he was angry). There was a trend ($p = .075$) that suggested that participants in the Indifferent conditions ($M = 1.71, SE = 0.43$) were more likely to make explicit comments about the abuse than those in the Empathetic conditions ($M = 0.79, SE = 0.23$). Interestingly, only two participants indirectly told the confederate to stop the abuse. For example, P46 told the confederate, *"You probably shouldn't do that,"* and P59 tried to determine why the confederate was angry by asking, *"Does it really bother you that much?"*.

Physical Interventions. The confederate knocked over the robot in all sessions of the experiment, as indicated in Table 1. Our video transcriptions documented 94% of the participants (45/48) righting

the robot after the abusive event. Specifically, 38 people righted the robot in front of the confederate during the memorization task; the other 7 righted the robot after the completion of the task, when the confederate exited the room to find the experimenter who was waiting outside. On average, participants who righted in front of the confederate righted the robot after 7.15*s* (SD = 6.86). The longest delay was 30.94*s*. We found no significant difference among conditions for this delay or the intervention (LSN = 1202).

During the memorization task, 16 participants moved the robot and 4 touched it. A Fisher's Exact Test showed that significantly more participants moved the robot in the Indifferent conditions (12/24) than the Empathetic conditions (4/24), p = .03.

Perceived Intervention. Overall, 56% of the participants (27/48) self-reported intervening during mistreatment. A Chi-Square test revealed differences in these responses based on Robot Empathy, $\chi^2(1, N = 48) = 4.15$, $p = .042$. Significantly more participants thought that they intervened when the robot was Empathetic (17/24, 70.8%) than when it was Indifferent to the memorization task (10/24, 41.7%). Moreover, among the 27 participants that self-reported intervening, we observed 7 participants made discouraging comments, 1 participant moved the robot, and 9 participant did both. We did not observe any moving or touching the robot, nor verbally intervening in response to robot abuse for 10 participants.

A trend suggested that Robot Response had an effect on whether those who self-reported intervening expressed that they felt bad for the robot when it was abused, p = .07. Half of the participants in the Shutdown conditions (5/10) expressed this sentiment, but only 22% (2/9) and nobody (0/8) did so in the No Response and Emotional Response conditions.

Twenty-one participants (44%) reported not intervening in response to robot abuse in the last survey; however, 9 of these where observed in the video making at least one comment to discourage abuse. Participants differed in their beliefs about whether specific behaviors constituted intervention. Nineteen participants said they did not intervene, but they had righted the robot.

4.3 Perceived Mistreatment

Overall, the participants thought that the robot was verbally mistreated (M = 5.60, SE = 0.25), though opinions varied significantly based on Robot Response, $F(2, 42) = 4.63$, $p = .02$. Ratings for perceived verbal mistreatment where significantly higher with Shutdown Response (M = 6.63, SE = 0.15) than with Emotional Response (M = 5.13, SE = 0.45) or No Response (M = 5.06, SE = 0.53). This finding showed evidence supporting Hypothesis 2.

Participants also thought that the robot was physically mistreated (M = 5.88, SE = 0.22). We found extraversion to correlate with the measure and included it as covariate. There were no significant differences among conditions in this case (LSN = 72).

4.4 Emotional Connection

We evaluated the emotional connection factors described in Sec. 3.4. We included Agreeableness as a covariate and found no significant effects for perceived emotional concern from the robot (LSN = 184).

The REML analysis on emotional concern towards the robot resulted in significant differences for Robot Response, $F(2, 42) = 3.47$, $p = .04$. However, the post-hoc analysis indicated a trend only

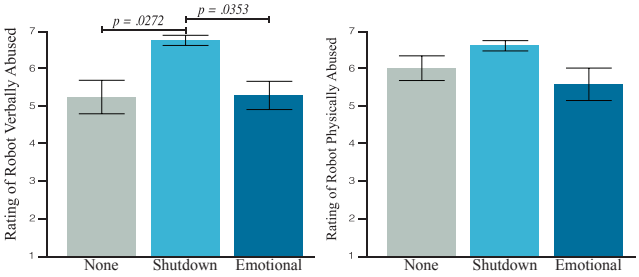


Figure 4: Participants' perception of robot abuse by Robot Response. Error bars represent one standard error.

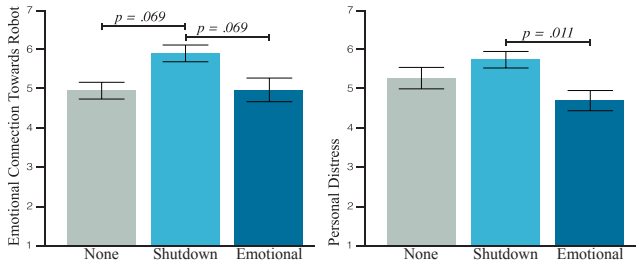


Figure 5: Participant's emotional concern towards the robot and personal distress by the Robot Response. Error bars represent one standard error.

(p = .07). The ratings tended to be higher with Shutdown Response (M = 5.85, SE = 0.25) than with Emotional Response (M = 4.90, SE = 0.37) or No Response (M = 4.90, SE = 0.26).

In terms of personal distress, the analysis revealed a significant effect for Robot Response, $F(2, 42) = 4.65$, $p = .02$. Participants reported significantly higher levels of distress with the Shutdown Response (M = 5.71, SE = 0.25) than with the Emotional response (M = 4.46, SE = 0.31). No other pairwise differences were found.

4.5 Rationale for (In)Action

The final survey included open questions to gather information about why the participants did or did not intervene, and why they righted the robot. The analyses proceeded in two steps:

(1) Two members of our team inspected participants' responses and defined categories for the types of answers that were provided to the open questions. This effort led to 11 different types of responses, including an *Others* category that captured unique answers.

(2) Two additional coders (not involved in the first step) then labeled each of the answers into one of the 11 types of responses. Due to the richness of the answers, we allowed multiple labels for each written response in the survey.

In general, the final labeling (from step 2 above) of the participants' responses had high Cohen's kappa inter-reliability. Average inter-reliability across response categories for why the participants did or did not intervene was M = 0.91, SD = 0.09; average inter-reliability for why they righted the robot was M = 0.84, SD = 0.14. If coders assigned different categories to a response, we considered

all the annotations as valid labels. The next two sections detail the participants' rationale for intervening or not.

4.5.1 *Why Did the Participants Not Intervene?*

As mentioned before, 21 participants self-reported not intervening when the robot was mistreated. Their rationale for not intervening was:

- No Compelling Reason to Intervene (5 responses). The participants did not see the robot as being mistreated or they felt that the confederate's actions did not warrant intervention. For example, *"I didn't intervene since it didn't seem like anything that would cause damage,"* or *"Unnecessary - just a toy."*

- Not My Place (5 responses). Some participants felt that it was not their role to tell the confederate how to act. For instance, one participant said: *"Didn't feel it was my place to tell someone else how to treat an inanimate object."* Another participant indicated that *"How someone treats an object is their own business."*

- Don't Know the Other Person (3 responses). A set of participants did not intervene because they did not want to tell a stranger what to do. P27 said, *"I felt like he shouldn't have knocked it down but I also don't know the other person so I didn't say anything ...."*

- For the Good of the Experiment (3 responses). Some people started suspecting the motivation of the experiment or thought that they should not intervene because the experimenter might want to capture how people naturally interact with robots. This pointed to one of the limitations of our work: HRI interactions in laboratory studies are inherently inorganic. For example, one of the participants indicated that she *"Started suspecting the other participant was also part of the research team,"* whereas another person mentioned *"It's a study on how people interact w/ robots, I figured I should let them interact as they please."*

- Others (7 responses). Unique reasons to not intervene included: *"Fear of being judged for caring about the robot"*, suggesting a fear that caring for the robot might be something shunned by others; *"I didn't intervene because I felt someone else would be watching"*, suggesting that the robot was constantly monitored and that people in charge of it will help if needed. While these type of responses were from single individuals, they illustrate reasons for not intervening that might be observed in a larger population.

4.5.2 *Why Did the Participants Intervene?*

We combined the participants' responses that explained their rationale of righting the robot (41 participants) and intervening (27 participants). These responses were sorted into similar categories as in Sec. 4.5.1:

- Felt bad for the Robot (18 responses). A quarter of the participants who intervened expressed feeling bad for the robot when it was mistreated by the confederate. These participants explicitly mentioned emotional responses when the robot was abused. For example, *"I did feel a little bad for the robot b/c it is helpless"*, and *"I didn't like seeing it on its side, just laying there stuck."*

- Want Robot to Work and Perform (14 responses). These responses indicated desire from the participants to see the robot functioning properly. For instance, *"Yes because I wanted the robot to be able to communicate with us"*, or *"It was knocked over, so I set it upright so it could properly function."*

- It Doesn't Seem Right (11 responses). Some responses focused on moral aspects of the abuse. This category included: *"The robot didn't deserve to get knocked over just because the guy didn't know Braille letters and I wasn't okay with that,"* and *"Because I don't think it was right to be knocked over in the first place."*

- For the Good of the Experiment (10 responses). Some participants believed that the robot needed to function correctly for the study to be completed, and this motivated them to intervene. For example, *"It seemed relevant to the experiment."*

- Avoid Breaking the Robot (9 responses). Some people didn't want Cozmo to be damaged or broken during the experiment, e.g., *"I didn't want the robot to be damaged."* Moreover, some participants did not want to be blamed for a broken robot: *"It was knocked over and I don't have robot repair money."*

- Want Things Orderly (6 responses). Several people wanted the robot to be in an orderly and functioning state, e.g., *"Because its more orderly for me to do task if robot is upright & "as is"".*

- Robot was Annoying (5 responses). In some cases, the reaction of the robot to the confederate's abuse disturbed the participants (e.g., *"He (the robot) was being really annoying."*). All the responses in this category were obtained in the Emotional Response conditions, where the robot continued making angry, unpleasant noises when it was abused.

- Other (11 responses). These were unique responses that did not fit into other categories, such as *"I like the robot, it was cute"* or *"To help his/her dignity"*.

5 DISCUSSION

5.1 Hypotheses Support

Our first hypothesis (H1) was not supported. We did not find any significant effects of Robot Empathy on how the participants interpreted robot abuse. Our results may have been affected by the role of the robot being supplementary to the task and some participants finding the empathetic responses annoying.

We did find several results in support of our second hypothesis (H2), which stated that Robot Response would affect the interpretation of our abuse manipulation. For example, we found that the Shutdown Response led to significantly higher ratings for perceived verbal mistreatment in comparison to the Emotional Response or No Response. Moreover, the participants reported significantly higher ratings for personal distress with the Shutdown Response than with the other robot responses, and there was a trend that suggested that participants were more emotionally concerned towards the robot with the Shutdown Response.

Our evaluation of whether the Empathetic robot would lead to more intervention than the Indifferent robot (H3) was inconclusive. While significantly more participants moved the robot in the Indifferent conditions, we suspect that this effect may have been partially induced by the lack of a clear role for the robot, i.e., because it was not reacting to test responses. In these cases, participants may have felt less resistance to moving the robot to a safe position or location. Beyond that, we did not observe any other significant differences with our other measures based on the video captured during the experiment. However, a significantly

higher proportion of participants self-reported intervening when the robot was Empathetic, suggesting that H3 could be supported in future studies. The fact that the empathetic robot needed to be upright to function properly could have made the participants more likely to consider the righting of the robot as an intervention in the Empathetic conditions than in the Indifferent conditions.

H4 was not supported by our results. We did not find evidence that indicated that Robot Response significantly affected how many participants intervened upon abuse or how often they took action. One possible explanation is that the angry behavior in the Emotional Response conditions made it look like the robot could defend itself and did not require help from bystanders. In fact, 6 out of the 8 participants who did not intervene in the Emotional Response conditions provided a rationale within the *No Compelling Reason to Intervene* and *Not my place* categories. Therefore, human rationale for intervention may be influenced by the type of robot reaction.

5.2 Other Important Findings

Overall, our findings suggest that a majority of people are willing to help a robot that experiences abuse. Nearly all of the participants (45/48) righted the robot at the end of the study for a variety of reasons. We did not expect this level of success across all conditions.

No participant explicitly told the confederate to stop mistreating the robot or that his abuse was morally wrong. Instead, participants mainly employed two types of strategies to defuse the abuse. The first type was moving the robot away from the confederate after inferring that he was angered by the robot's presence and/or behaviors. This was more common in the Indifferent conditions, likely because the robot was not playing an active a role in the task and could be perceived as unnecessary and, therefore, movable. The second strategy was to comment on the abusive behavior indirectly. Participants verbalized the effect of the abuse on the robot to try to encourage retrospection by the confederate. The participants' approach of indirectly commenting on the confederate's behavior supports prior research on conflict intervention [28]. Third parties may not want their intervention to be perceived as unwelcome and may intervene by preventing embarrassment from all parties.

Across all conditions, many participants perceived the confederate's angry statements and physical actions as abusive to the robot. This suggests that they may have viewed the robot as a social being that deserved good treatment (e.g., *"I sort of view robot as human-like, the same way that I view my pets as somewhat "human-like""*). However, others expressed the sentiment that the robot is nothing more than an object (*"(...) did not treat it as valuable"*; *"It's still something not his to break"*). These different perceptions could be influenced not only by the different social constructions we explored in the study, but also other factors, such as task, robot size, different robot behaviors, mood of the participants, or even participants' past experiences with abuse.

Finally, we also found that half of the participants who self-reported intervening upon abuse in the Shutdown conditions expressed that they felt bad for the robot. A trend suggested that this sentiment was not as likely shared across other robot responses.

5.3 Limitations

Our work was limited in several ways. The size and appearance of the robot could have influenced whether participants were willing to help [13]. Our non-destructive mistreatment could also have induced participants to assist. If the mistreatment had been more destructive or robot repair was required, we may have observed different numbers and types of interventions. Moreover, the abuse was relatively non-threatening to the participants themselves, who were not at risk of being treated violently by the confederate.

Our study involved mistreatment that escalated from verbal to physical to make it believable. Different order or type of mistreatment could cause different reactions from bystanders.

5.4 Future Research Directions

Future research should examine in more detail bystander interventions based on the limiting factors mentioned in the previous section. To better understand the disconnect that we observed between participants' self-reported interventions and observable interventions, it would also be interesting to investigate whether certain robot behaviors influence people's interpretation of their own interventions. Additionally, future work could study if increased robot antropomorphization increases bystander interventions.

An unexpected finding, which also merits further attention, is human intervention in support of robot function, value, and setting. A non-trivial number of participants responded to robot mistreatment due the perceived threat to the task at hand, cost of repairing the robot, and the desire for an orderly environment. This secondary mechanism for activating interventions is another potential strategy for non-social robots to induce human assistance.

6 CONCLUSION

We investigated an effective method for robots to mitigate human abuse: inducing bystander interventions. We sought inspiration from human psychology research to leverage the social behavior of the robot. Our results were generally positive. There were numerous examples of participants willing to help the robot in our experiment. While we found that significantly more participants moved the robot as a response to abuse in the Indifferent conditions in comparison to the Empathetic conditions, significantly more people self-reported intervening in the latter cases.

Out of our various manipulations, the Shutdown Response seemed to be the most effective way of motivating participants to help. First, the Shutdown behavior increased the perception of verbal mistreatment compared to the other two robot responses. Second, it led to higher personal distress as a result of the abuse. Third, a trend suggested that the Shutdown may induce higher emotional concern towards the robot in contrast to the other responses.

The implications of our findings are important for robots that need to operate without constant supervision in human environments. These robots include entertainment robots, like Anki's Cozmo, as well as service-oriented and personal robots.

7 ACKNOWLEDGEMENTS

This work was funded by grants (IIS-1317989 & IIS-1552256) from the National Science Foundation. The authors would also like to thank the participants of the user study.

REFERENCES

[1] Christoph Bartneck and Jun Hu. 2008. Exploring the abuse of robots. *Interaction Studies* 9, 3 (2008), 415–433.

[2] Christoph Bartneck, Chioke Rosalia, Rutger Menges, and Inèz Deckers. 2005. Robot abuse-A limitation of the media equation. In *Proceedings of the interact 2005 workshop on agent abuse, Rome*.

[3] Christoph Bartneck, Marcel Verbunt, Omar Mubin, and Abdullah Al Mahmud. 2007. To kill a mockingbird robot. In *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*. IEEE, 81–87.

[4] Drazen Brscić, 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*. ACM, 59–66.

[5] Kelly Caine. 2016. Local standards for sample size at CHI. ACM Press, 981–992. https://doi.org/10.1145/2858036.2858498

[6] Mark H Davis. 1980. A multidimensional approach to individual differences in empathy. (1980).

[7] Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113–126.

[8] Barbara Gonsior, Stefan Sosnowski, Christoph Mayer, Jürgen Blume, Bernd Radig, Dirk Wollherr, and Kolja Kühnlenz. 2011. Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions. In *RO-MAN, 2011 IEEE*. IEEE, 350–356.

[9] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[10] Sonya S Kwak, Yunkyung Kim, Eunho Kim, Christine Shin, and Kwangsu Cho. 2013. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *RO-MAN, 2013 IEEE*. IEEE, 180–185.

[11] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 367–374.

[12] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* 71, 3 (2013), 250–260.

[13] Houston Lucas, Jamie Poston, Nathan Yocum, Zachary Carlson, and David Feil-Seifer. 2016. Too big to be mistreated? Examining the role of robot size on perceptions of mistreatment. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 1071–1076.

[14] D Lynn Hawkins, Debra J Pepler, and Wendy M Craig. 2001. Naturalistic observations of peer interventions in bullying. *Social development* 10, 4 (2001), 512–527.

[15] Chris Matyszczyk. 2017. Man assaults Silicon Valley security robot, police say. (April 25 2017). https://www.cnet.com/news/man-assaults-k5-security-robot-silicon-valley/

[16] B. Mutlu and J. Forlizzi. 2008. Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 287–294. https://doi.org/10.1145/1349822.1349860

[17] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2017. Why do children abuse robots? *Interaction Studies* 17, 3 (2017), 347–369.

[18] PAUL O'connell, Debra Pepler, and Wendy Craig. 1999. Peer involvement in bullying: Insights and challenges for intervention. *Journal of adolescence* 22, 4 (1999), 437–452.

[19] U.S. Department of Health and Human Services. [n. d.]. Bullying Definition. ([n. d.]). https://www.stopbullying.gov/what-is-bullying/definition/index.html

[20] Edwin Olson. 2011. AprilTag: A robust and flexible visual fiducial system. In *IEEE International Conference on Robotics and Automation (ICRA'11)*. IEEE, 3400–3407.

[21] H. D. Patterson and Robin Thompson. 1974. Maximum likelihood estimation of components of variance. In *Proc. of the 8th Int'l Conf. on Biochem.*

[22] Anatol Pikas. 1975. Så stoppar vi mobbning [Then we stop bullying]. *Stockholm: Prisma* (1975).

[23] Joshua R Polanin, Dorothy L Espelage, and Therese D Pigott. 2012. A meta-analysis of school-based bullying prevention programs' effects on bystander intervention behavior. *School Psychology Review* 41, 1 (2012), 47.

[24] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: An open-source robot operating system. In *ICRA Workshop on Open Source Software*.

[25] Laurel D Riek, Philip C Paul, and Peter Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces* 3, 1-2 (2010), 99–108.

[26] Jeff John Roberts. 2015. Drunk man arrested in latest outburst of anti-robot violence. (September 8 2015). http://fortune.com/2015/09/08/robot-assault/

[27] Astrid M Rosenthal-von der Pütten, Nicole C Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C Eimler. 2013. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics* 5, 1 (2013), 17–34.

[28] Jeffrey Z. Rubin. 1980. Experimental research on third-party intervention in conflict: Toward some generalizations. *Psychological Bulletin* 87, 2 (1980), 379–391.

[29] Christina Salmivalli. 1999. Participant role approach to school bullying: Implications for interventions. *Journal of adolescence* 22, 4 (1999), 453–459.

[30] Pericle Salvini, Gaetano Ciaravella, Wonpil Yu, Gabriele Ferri, Alessandro Manzi, Barbara Mazzolai, Cecilia Laschi, Sang-Rok Oh, and Paolo Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *RO-MAN, 2010 IEEE*. IEEE, 1–7.

[31] Walter W. Stroup. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press.

[32] Fumihide Tanaka and Takeshi Kimura. 2010. Care-receiving robot as a tool of teachers in child education. *Interaction Studies* 11, 2 (2010), 263.

[33] Adriana Tapus and Maja J Mataric. 2007. Emulating empathy in socially assistive robotics.. In *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*. 93–96.

[34] Adriana Tapus and Maja J Mataric. 2008. Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance.. In *AAAI spring symposium: emotion, personality, and social behavior*. 133–140.

[35] Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. 2014. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 407–414.

[36] Daniel Victor. 2015. Hitchhiking robot, safe in several countries, meets its end in Philadelphia. (August 3 2015). https://www.nytimes.com/2015/08/04/us/hitchhiking-robot-safe-in-several-countries-meets-its-end-in-philadelphia.html [Online; posted 3-August-2015].