International Journal of Social Robotics (2022) 14:499–514
https://doi.org/10.1007/s12369-021-00799-1

Pay Them No Mind: the Influence of Implicit and Explicit Robot Mind Perception on the Right to be Protected

Merel Keijsers[1] · **Christoph Bartneck**[1] · **Friederike Eyssel**[2]

Accepted: 9 June 2021 / Published online: 2 July 2021

Abstract
Mind perception is a fundamental part of anthropomorphism and has recently been suggested to be a dual process. The current research studied the influence of implicit and explicit mind perception on a robot's right to be protected from abuse, both in terms of participants condemning abuse that befell the robot as well as in terms of participants' tendency to humiliate the robot themselves. Results indicated that acceptability of robot abuse can be manipulated through explicit mind perception, yet are inconclusive about the influence of implicit mind perception. Interestingly, explicit attribution of mind to the robot did not make people less likely to mistreat the robot. This suggests that the relationship between a robot's perceived mind and right to protection is far from straightforward, and has implications for researchers and engineers who want to tackle the issue of robot abuse.

Keywords Dual processing · Mind attribution · Mind perception · Robot abuse · Aggression

1 Introduction

Humans tend to automatically ascribe social robots a certain scope of cognitive and emotional abilities. The consequences of this mind perception can be observed in human behaviour during human-robot interaction (HRI): humans tend to be polite to a robot [46] and have been recorded trying to keep it safe from harm [14]. Mind perception affects people's cognitive responses to robots as well. For example, humans apply (human) stereotypes when interpreting a robot's behaviour [6,20] or even forming expectations about its traits [54]; and the degree of mind perception influences what behaviour they deem acceptable towards a robot [37,56]. The degree to which people attribute a mind to robots depends on a number of factors, among which are personal dispositions of the human and qualities of the robot [17,21].

Social interactions are not necessarily always positive, however. Robot abuse represents a good example of negative social behaviour which requires being attributed some extent of mind. In the short history of HRI people have spontaneously slapped [50], kicked [39], pushed [8], and verbally abused [39,46,50] robots. As Salvini et al. [50] remarked, this behaviour appeared to be motivated by the wish to engage with the robot in a social way (albeit negative) rather than representing acts of vandalism. As a consequence, they labelled the behaviour *robot bullying*, a term later adopted by other HRI researchers [30,32,42,56].

Bullying has been defined as physical and/or psychological aggression which has the intention to harm or hurt the victim [1,38,52]. As implied by this definition, bullying behaviour requires that the target is regarded as sentient being that can be intimidated and humiliated; indeed, the capacity to perceive a mind in another (human or otherwise) is associated with bullying behaviour [55]. Paradoxically, the same mind perception that would enhance a robot's perceived right to moral treatment, is required to make it a potential target for bullying.

Previous work has suggested that mind perception is a dual process [19,58,64], i.e. the result of an initial quick, implicit, and automatic process which may or may not be moderated

✉ Merel Keijsers
merel.keijsers@pg.canterbury.ac.nz

Christoph Bartneck
christoph.bartneck@canterbury.ac.nz

Friederike Eyssel
friederike.eyssel@uni-bielefeld.de

[1] Human Interface Technology Lab (HITlab NZ), University of Canterbury, Christchurch, New Zealand

[2] Center for Cognitive Interaction Technology (CITEC), Universität Bielefeld, Bielefeld, Germany

by more deliberate reflective processes. Nonhuman agents thus may be automatically categorised as humanlike and possessing a mind, but more deliberate reflective processes can moderate this initial evaluation. Special about robots (compared to animals, both human and nonhuman) is that people rationally know these agents are insentient, thus creating additional tension between automatic and reflective processes. The current set of experiments aims to explore the influence of a robot's implicit and explicit mind attribution on the acceptability of its abuse.

While there is an extensive literature on the ethical aspects of robot bullying (see for example [32,47,53]), the debate on the ethical implications of robot abuse are beyond the scope of this paper. Instead, it will focus on robot mind attribution as a psychological moderator of how robot bullying is perceived.

1.1 Mind Perception in Robots

Psychological anthropomorphism, or the tendency to ascribe humanlike attributes (e.g. emotions, cognition, intentions, characteristics) to nonhuman entities such as deities, objects, animals and even abstract paintings [18] is of all ages [34]. It has been theorised to be motivated by three factors: elicited agent knowledge, and both the human need to explain or understand the agents behaviour as well as the need for social interaction [17].

This first factor postulates that if one or more characteristics of the agent activates knowledge about humans, this knowledge will then be generalised to the agent. For example, if the agent has a humanlike voice, knowledge structures on how to extract information from voices (e.g. gender and emotional state of the speaker) will become activated and the human will use this knowledge to make inferences about the agent. Previous research has shown that this could also hold true in HRI: the gender of a robot's voice does not influence anthropomorphism ratings when the voice is computer generated, but when the voice is human robots with the same gender as the participant are rated as more "like them" [22].

The second factor proposes that people will use mind perception as a way to reduce uncertainty about, and by consequence regain a sense of control over, the environment. By projecting humanlike intentions onto an erratic agent, humans can try to anticipate its future behaviour. This factor as well has been confirmed in the HRI context: people perceive mind in an unpredictable robot to a greater extent than in a predictable robot [61].

Finally, the third factor asserts that people have a need for social connection and may turn to nonhuman agents to fulfil this need. Indeed, self-reported loneliness correlated with participant's mind perception in a robot; and when experimentally manipulated, loneliness increased mind attribution to a wide range of other nonhuman agents [18].

Mind perception has been linked to the moral standing of an agent [24] and, in human-human interaction, (inversely) to aggression [26]. Humans have the innate drive to maintain a positive self image, and knowingly hurting a sentient being would normally interfere with perceiving oneself as a moral individual. In order to resolve this conflict humans perceive their human victims as slightly less capable of thinking and feeling before engaging in an act of aggression [10,36]. As a result of this lowered mind attribution aggressors are discharged from the obligation of treating the victim in a moral way, which in turn allows them to harm or hurt their victim.

This link between mind attribution and (un)willingness to harm or let be harmed has also been explored in an HRI context specifically. Research on the influence of a robot's looks or background story on people's willingness to protect it from harm and their perceived immorality of hurting it suggests that adding anthropomorphic features increases moral standing. For example, humanlike visual cues on a robot increase willingness amongst participants to protect it from harm [48]. Attributing a personality to a simple bug robot through providing further information about it increased participants' hesitation before harming it [16]. Similarly, functional robots that had been given a name allegedly were less often the target of abuse when they malfunctioned than their anonymous counterparts [15]. Briggs and Scheutz [7] experimentally manipulated a small humanoid robots capability to express distress, and found that this indeed affected participants' behaviour. People were less likely to insist that the NAO robot should follow the command to topple over a tower it had just painstakingly built when the robot protested in an emotional way compared to when it remained silent [7]. Similarly, participants hesitated longer before switching off an expressive iCat robot when the robot acted intelligent and agreeable, than when it communicated in a manner that was either intelligent, or agreeable, or neither [3].

However, other studies report conflicting results. In contrast to the iCat study [3], Horstmann et al. [28] found that when a robot without social behaviour protested against being switched off, this resulted in longer hesitation and lesser inclination to switch the robot off than when it did display social cues. Tan et al. [56] measured participants' willingness to intervene as a confederate verbally and physically abused a small Cozmo robot during a collaborative game. There was a marginal trend where participants were more likely to discourage mistreatment if the robot did not display any emotional cues throughout the game, versus when it celebrated successes and mourned losses. Nomura et al. [42] interviewed children who abused the anthropomorphic Robovie robot in a shopping mall. Most children saw the robot as human-like rather than machine-like and about half saw the robot as capable of perceiving its environment. Yet neither observation had stopped them from physically and verbally abusing the robot.

How to explain these inconsistent findings? Urquiza-Haas and Kotrschal [58] proposed that mind perception, and by extension empathy, is the result of a cognitive dynamic where implicit processes cause the emergence of early evaluations, and deliberate, reflective processes shape and nuance this evaluation (see also [62]). Initial implicit responses can be triggered by features of the agent [62] or priming [58] and then are moderated by reflective processing. This reflective or top-down processing is slow, takes effort, and is limited by motivation and working memory capacity: cognitive load has been shown to increase mind perception [60], as have feelings of loneliness (i.e. increased motivation to find social interaction) [18]. Because these factors are highly variable, it has been theorised that while bottom-up processes will show little variance between people, explicit mind perception will be highly volatile [58].

1.2 Robot Mind Perception as a Dual Process

There is a number of studies that show activation of implicit mind perception processes in human robot interaction. For example, mirror neurons get activated after observing performing an action regardless of whether the arm is human or robotic [23,43]. In addition, both watching a robot versus a human being abused activates certain brain areas related to empathy [44]. Krach et al. [31] found that brain areas that have been linked to the Theory of Mind network are activated when humans played a game with a robot opponent (both humanoid and functional in design); activation of the same areas and intensity was observed when the interaction partner was a human. Providing social cues, such as expressing emotions and displaying non-verbal behaviour, can furthermore enhance implicit mind perception [2,63,64].

Top-down processes can furthermore alter the initial evaluation after it has been formed. Rosenthal-Von Der Pütten et al. [49] compared participants' brain activation patterns in response to viewing the abuse of a cardboard box, a robot and a human. They found that in the questionnaires participants attributed equal levels of emotion to the human and the robot, and reported feeling the same amount of empathy towards the robot and the human when they were mistreated. In contrast, fMRI scans showed greater activation in participants' right putamen when watching the human being mistreated than when watching the robot being mistreated. This area has been associated with empathy and emotional distress [49].

The opposite effect, where brain activation patterns are interchangeable but explicit evaluation differs has been found as well. In a study where participants played a game against either a functional robot, a humanoid robot, or a human opponent, activation of brain areas related to Theory of Mind were very similar across all partner types, yet self-report measures differed. Participants claimed that they had more fun when they interacted with either the human or the humanoid robot opponent (compared to the functional robot), and they attributed higher intelligence and more competitiveness to these opponents too [31]. To complicate interpretation further, Banks [2] found that implicit mind attribution, which was deducted from the way that people interpreted robot behaviours, was not significantly related to explicit mind attribution, which was measured by the dichotomous question "does this agent have a mind?". The binary either-or decision may have influenced these results, especially since the line-up of agents that were assessed for possessing mind included a human. It is possible that the human was used as a benchmark, and none of the robots was reliably seen as having an equal amount of mind.

Thus, the degree to which people perceive a mind in a robot depends on a number of factors. Some of these influence implicit mind perception, such as the personality of the person and qualities of the robot [17,21], and others which moderate explicit processing, such as causal reasoning about a robot's mental state [58]. To our knowledge, however, no study has considered the separate effects of implicit and explicit robot mind perception on (right to protection from) aggression.

1.3 Current Experiments

The current research will search to answer the following research questions:

1. Does telling people that a robot possesses a mind, i.e., is capable of experiencing emotions and cognition, affect how unacceptable they find robot bullying?
2. Do emotional cues that imply the robot has a mind affect how unacceptable people find robot bullying?
3. Does explicit information that a robot does not possess a mind change the influence of implied mind on how unacceptable people find robot bullying?
4. Does telling people that a robot possesses a mind reduce their willingness to publicly humiliate it?

These questions are addressed in two experiments. In Experiment 1, was vignette-based and addressed the first three research questions. Robot sentience was manipulated in two ways: through having the robot display emotional cues, and through telling participants that the robot could think and feel. Participants then indicated how unacceptable they considered varying bullying behaviours towards the robot.

In Experiment 2, the first and the last research question were addressed. Participants interacted with embodied robot that was introduced as either capable or incapable of thinking and feeling. They then indicated how unacceptable they considered bullying this robot; and were offered an opportunity to humiliate the robot they had just interacted with.

It was expected that both implicit and explicit mind attribution would enhance a robot's right to protection. Moreover, it was expected that participants would be more likely to humiliate the robot after being explicitly informed of its lack of mind.

The experiments have been reviewed and approved by the Human Ethics Committee at the University of Canterbury (reference HEC2019/47).

2 Experiment 1

Experiment 1 was an online scenario-based (vignette) study that followed a 2 (Explicit mind attribution through robot introduction: no mind attributed versus mind attributed) × 3 (Implicit mind attribution through robot response to mistreatment: no response, non-emotional response, emotional response) between participant design. The dependent variable was the robot's right to protection, measured in the extent to which participants condemned the mistreatment of the robot.

The non-emotional response was added to the implicit mind attribution manipulation to rule out the possibility that a higher condemnation of abuse for the protesting robot was not due to its emotional response, but rather because participants had been reminded that mistreatment in general is bad. Participants' general tendency to anthropomorphise and affinity with technology were assessed in order to check whether they were similarly distributed across the six conditions.

2.1 Methods

2.1.1 Participants

Participants for Experiment 1 were recruited using Amazon Mechanical Turk (MTurk), an online platform for data collection. Previous studies have indicated that data collected via MTurk are of equal quality to data collected through on-site recruitment or participant data from forums [5,51], with internal motivation rather than monetary reward being the main motive for participating [9]. We restricted participation to participants residing in English-speaking countries (i.e. USA, Canada, Australia, New Zealand, United Kingdom, or Ireland) and accredited with Master status, i.e. with a low incidence of work being rejected.

A total of 193 people participated in Experiment 1. After having recruited the first 66 participants, it became apparent that basic demographics, gender and age, had not been assessed. Therefore, these demographics were included for the remainder of the data collection. Of the 126 participants who completed the survey after the error had been detected, 53.97% were female, with a mean age of 41.65 years (SD = 11.42). In return for their participation, workers were reimbursed with .90US$, in accordance with MTurk reimbursement custom.

2.1.2 Procedure

Prospective participants could read a short description of the study in MTurk. If they decided to participate, they were directed to a Qualtrics survey page, where they provided informed consent and reported their age and gender.

Subsequently, participants were randomly assigned to one of the six conditions. Depending on the condition, participants were presented with a vignette in which the robot was either explicitly attributed a mind or explicitly not attributed a mind (see Table 1. The vignette further described a human-robot interaction between a participant and the robot which included mistreatment of the robot. Depending on the experimental condition, the robot responded to the mistreatment in a non-emotional way, in an emotional way, or not at all.

Finally, participants completed the survey. First they indicated how much they condemned the mistreatment as described in the vignette. Subsequently, they completed the individual tendency to anthropomorphise and the affinity with technology scales. Then participants were thanked for their time, debriefed, and reimbursed.

2.1.3 Materials

Vignettes Vignettes constituted an introduction of the robot, a description of the robot mistreatment, and a description of the robot's response to the mistreatment. In order to create the six conditions, nine vignettes were constructed: two different introductions (explicit mind attribution and explicit lack of mind robot), four different robot mistreatment scenarios, and three different robot responses (no response, non-emotional response, emotional response). See Table 1 for the respective introductions.

The four mistreatment descriptions were created to cover a wide scope of robot mistreatment. They described an interaction between the robot and a participant, where the participant had behaved in an aggressive or impolite way towards the robot: either playing around with the robot's energy supply; verbally abusing it; rejecting the proposal from the robot to split a monetary reward evenly in favour of keeping all the money for themselves; or switching off the robot in spite of the robot asking to be left on in sleep mode since switching it off would result in the robot losing awareness.

There were 2 × 4 × 3 = 24 possible vignettes to which participants were randomly assigned. The interaction descriptions were not included as an independent variable as they were expected to have no effect on the dependent variable. A manipulation check was carried out to confirm this; see Sect. 2.1.5.

Table 1 Robot introductions manipulation

Introduction not possessing a mind	Introduction possessing a mind
This robot is programmed to appear to be a social being: it is capable of processing, interpreting and calculating an emotional response to its environment. It can store and retrieve names and faces, so that it will state the name of people it has seen before out loud. It can also respond to prespecified commands, and will update its behaviour scheme to mimic an upset or angry response when given certain prompts. It has distance and depth sensors that prevent it from colliding with objects or people and falling off the stairs. All these behaviours give it the appearance of being conscious. The robot has starred in a few of our demos before, although it can't remember this. Recently it made its appearance in its first experiment. However, being a robot, it did not feel excited or nervous.	This is a social robot with his very own personality: it is capable of processing, interpreting and emotionally responding to its environment. It can remember names and faces, and will recognise people it met before. It can understand different commands, and will change its mood depending on how it is treated - for example, it will get upset when mistreated and happy after being told it did well. Moreover, it is aware of its surroundings, so that it can avoid bumping into objects or people and throwing itself off the stairs. The robot is proud to have starred in a few of our demos before, and recently made its appearance in its first experiment, for which it was very excited and a little nervous.

Left: introduction of the robot not possessing a mind. Right: introduction for the robot possessing a mind

2.1.4 Measurements

Condemnation of mistreatment Condemnation of mistreatment was measured through five items, each scored on a 7-point Likert scale. The items concerned how opposed the participant was to treating the robot like it was treated in the vignette; if they considered the treatment as described acceptable; if they would intervene if they were to witness such treatment of a robot; how important it was to protect a robot like the one in the vignette from being treated like it was; and in general, how important it was for a robot like the one in the vignette to be treated humanely.

Individual differences in anthropomorphism Individual differences in anthropomorphism were measured with a questionnaire from Waytz et al. [59], although the questions that targeted anthropomorphic qualities of technology (i.e. computers, cars and robots) were taken out since those would likely be affected by the introduction manipulation. The resulting questionnaire consisted of 10 items. Participants were asked to indicate on a 10-point Likert scale to what extent they thought different animals and natural phenomena have mental and emotional responses (e.g. "To what extent does the environment experience emotions?", "To what extent does the average insect have a mind of its own?").

Individual differences in anthropomorphism were assessed in order to check whether participants had similar trait anthropomorphism across the different conditions.

Affinity with technology Affinity with technology was measured with a questionnaire taken from Neyer et al. [41] and translated from German to English. Participants' individual affinity with technology is measured through their agreement with eight statements (e.g. "I am very curious about new technological developments") on a 5-point Likert scale (ranging from "not at all descriptive of me" to "extremely descriptive of me").

Like the individual differences in anthropomorphism, affinity with technology was assessed in order to check whether participants in different conditions had similar affinity with technology.

2.1.5 Preliminary Analyses

Cronbach's alpha was computed for the condemnation of mistreatment scale, as well as the individual differences in anthropomorphism questionnaire and the affinity with technology questionnaire. Internal consistency was high; $\alpha = .88$ for condemnation, $\alpha = .85$ for anthropomorphism, and $\alpha = .88$ for affinity with technology. The questionnaires and scale were thus deemed reliable [12].

A 4×1 ANOVA with the four mistreatment scenarios as factors and acceptability of mistreatment as dependent variable confirmed that the scenario did not influence acceptability scores, $F(3, 189) = 1.01$, $p = .389$. The mistreatment scenarios thus could be excluded as a factor, as intended.

A 2×3 ANOVA with the robot introduction and robot response as factors and age as dependent variable indicated that age was not equally distributed across the conditions. More specifically, there were main effects for the introduction manipulation ($F(1, 121) = 6.56$, $p = .012$) as well as the robot response ($F(2, 121) = 5.68$, $p = .004$). A correlation between age and condemnation however was not significant, $\rho = -.12$, $p = .175$). Thus, the difference in age between the conditions was not considered problematic for the condemnation measure. See Table 2 for the mean age per condition.

A Chi-Square test on the distribution of gender across the six conditions indicated that there was also a significant difference in male to female ratio between the conditions, $\chi^2(5) = 13.32$, $p = .021$, with fewer females in the 'explicit no mind attribution'/'non-emotional response to abuse' condition (see Table 2 for the gender ratio per condition). However, a regression with condemnation as dependent variable and gender as predictor indicated that gender was not significantly related

Table 2 Mean scores (*SD*) for age, trait anthropomorphism, affinity with technology, condemnation of mistreatment per condition

	Explicitly no mind attributed			Explicitly mind attributed		
	No response	Non-emotional	Emotional	No response	Non-emotional	Emotional
Age	50.06 (10.13)	41.00 (12.39)	38.83 (9.29)	41.28(11.15)	39.57 (12.86)	40.65 (10.11)
Percentage male	55.56%	21.74%	50%	45.83%	33.33%	76.47%
Trait anthropomorphism	4.12 (1.59)	4.54 (1.78)	3.77 (1.77)	3.75 (1.46)	3.90 (1.57)	3.85 (1.80)
Affinity with technology	3.91 (.77)	3.73 (.93)	4.05 (.94)	3.70 (.92)	3.96 (.80)	3.83 (.70)
Condemning mistreatment	2.71 (.98)	3.01 (.97)	3.16 (1.01)	3.02 (.98)	3.39 (1.10)	3.42 (1.16)

to condemnation, $t(124) = 1.83$, $p = .070$. Thus, the different gender ratios across the conditions were not considered problematic.

Anthropomorphic tendencies were similar between the conditions, $Fs < 1.79$, $ps > .17$. Affinity with technology, as well, was similar between the conditions, $Fs < 1.44$, $ps > .24$. See Table 2 for the mean scores for each scale per condition.

2.2 Results

2.2.1 Main Analyses

To test the influence of implicit and explicit mind perception on acceptability of robot mistreatment, a 2 (explicit mind attribution: mind attributed vs no mind attributed)×3 (robot response: no response, non-emotional, emotional) ANOVA with 'condemnation' as dependent variable was conducted. Significant main effects were found for both explicit mind attribution ($F(1, 187) = 4.56$, $p = .034$, $\eta_p^2 = .024$) and the robot response ($F(2, 187) = 3.07$, $p = .049$, $\eta_p^2 = .029$). See Fig. 1 for the plotted data.

Post-hoc analysis with a Tukey correction on the explicit mind attribution manipulation revealed that robot maltreatment was considered significantly less acceptable when the robot had explicitly been attributed a mind, M(*SD*) = 3.27(*.106*), than when it had explicitly been attributed no mind, M(*SD*) = 2.96(*.105*); $t(1, 187) = -2.12$, $p = .035$. Post-hoc analysis with a Tukey correction of the robot response manipulation revealed a marginal difference between participants in the non-responsive condition, M(*SD*) = 2.86(*.126*), and the emotional response condition, M(*SD*) = 3.29(*.134*), $t(1, 187) = -2.32$, $p = .056$. The other contrasts were also not significant, $ts > -1.86$, $ps > .155$.

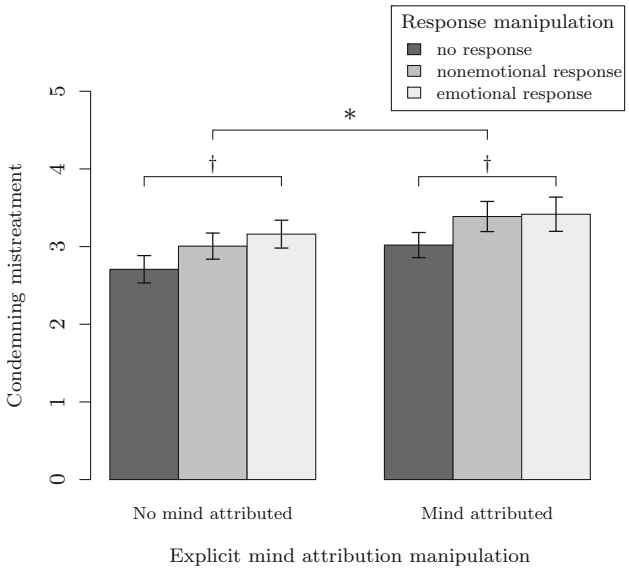

Fig. 1 Difference in mean condemnation for the robot introduction and robot response conditions. Bars indicate the standard error, * indicates $p = .035$, † indicates $p = .056$

2.3 Discussion

Using an online vignette study, Experiment 1 explored the influence of mind perception in a robot on its right to protection. Mind perception was manipulated both explicitly, by telling participants that the robot was capable of thinking and feeling (versus stating that even if it gave off such signs, it was actually incapable of thinking and feeling); and implicitly, through having the robot respond to mistreatment in an emotional way. In line with expectations, participants found mistreatment less acceptable if they had been explicitly told that the robot was able of thinking and feeling. While a significant main effect was detected for robot response on acceptability, no significant differences between individual conditions emerged from the post hoc test. Looking at the plotted data (Fig. 1), the gap appears to exist mainly between the 'no response' and 'response' (emotional or nonemotional) conditions. Planned comparison tests could to confirm this, but since these were not planned on a priori it would be inappropriate to report them for the current data set. There was no significant interaction effect between implicit and explicit mind perception on how acceptable participants found robot mistreatment.

These results are insufficient to argue that a robot's perceived right to protection may be the result of a cognitive dynamic [58]. Explicit information of robot mind attribution

clearly affects its right to protection, yet the data and analyses at hand are insufficient to conclude that implicit cues trigger implicit mind attribution. The observed main effect for the response condition together with the plotted data (Fig. 1) suggest that any response (emotional or nonemotional) from the robot may have decreased how acceptable mistreatment was seen. This would mean that not the emotional load, but the overall feedback from the robot's side evoked the participant's response. It may be the case that the robot protesting (emotional or nonemotional) still increased mind perception, as it implies that the robot was aware of its surroundings. It is also possible that the participants took the protest as feedback on the abuser's behaviour, and disapproved of the abuser going directly against instructions provided in an experimental setting. Since mind attribution was not measured, we are unable to test which of the two was the case.

A second limitation of Experiment 1 concerns the scenario-based approach, as well as the measurement of behavioural intentions rather than behaviour. Because Experiment 1 did not include actual human-robot interaction, behavioural intentions can only serve as a proxy for participants' behaviour towards a robot. Previous research on the Media Equation theory [45] demonstrated a divergence between self-report and actual behaviour. For example, research by Nass et al. [40] showed that participants adhered to social norms of politeness when interacting with a computer. However, when asked if they ever were considerate of the computer's feelings, participants would strongly deny this. Embodied robots tend to create a larger social presence [33,35], which is important for triggering social behaviour [57].

In addition, Experiment 1 only measured whether robot mind attribution influences the robot's right to protection. Whether condemning robot mistreatment leads to a reduction in abuse remains to be tested. We conducted Experiment 2 to overcome the problems surrounding scenario-based approaches, to replicate the findings, and to extend the experiment with a measurement of mistreatment behaviour. Moreover, pilot studies were added to validate the manipulations a priori.

3 Pilots

Two pilot studies were completed to validate the manipulation of mind attribution and the measurement of mistreatment in Experiment 2.

Experiment 2 followed a simple single-factor design with two dependent variables. The independent variable was mind attribution to the robot, manipulated in a similar way as the explicit mind attribution in Experiment 1, i.e. by means of the robot's introduction. The first dependent variable was how unacceptable robot abuse was considered, measured in the same way as in Experiment 2. The second dependent variable was the participant's willingness to publicly humiliate the robot.

3.1 Pilot Study 1: Robot Mind Attribution Manipulation

The robot mind attribution manipulation of Experiment 2 was validated in Pilot study 1. The mind attribution manipulation was an extension of the descriptions displayed at Table 1. A few lines were added about the robot's capabilities, as well as a picture of the Vector robot that would be used in Experiment 2.

3.1.1 Methods

Participants Participants were recruited via MTurk. 51 people participated. 24 participants read an introduction that explicitly stated the robot was incapable of thinking and feeling (i.e. no mind attributed), whereas 27 participants read an introduction that depicted the robot as capable of thinking and feeling (i.e. mind attributed). 54.90% of the participants were male, 41.18% female, and 3.92% (two participants) withheld from disclosing their gender. Mean age was 39.12(SD = 9.07). Participants received .65 US$ for their participation.
Procedure After assessing informed consent as well as age and gender, participants read either of the two proposed robot introductions and filled out the mind attribution questionnaire. Then they were thanked for their time and reimbursed.
Materials The mind attribution questionnaire was taken from Gray et al. [24] and adapted so that the questions explicitly referred to the robot from the introduction. The questionnaire measures to what extent the robot is capable of experiencing 18 different emotional and cognitive states, using a 5-point Likert scale that ranges from "very incapable" to "very capable".

3.1.2 Results

An independent t-test was conducted to test the difference between the two introductions. Mind attribution to the robot was significantly higher for the introduction depicting the robot as possessing a mind, M(SD) = 2.64(*.91*), than for the introduction depicting the robot as not possessing a mind, M(SD) = 1.90(*.95*), $t(48.75) = -2.85$, $p = .006$. The mind attribution manipulation was thus considered valid to use.

3.2 Pilot Study 2: Robot Mistreatment

Experiment 2 operationalised its dependent variable "robot humiliation" as whether participants chose to put up a condescending review on public display next to a robot they had just interacted with. This operationalisation was developed,

tested, and validated in Pilot 2. The resulting review pairs were similar in sentiment and informativeness, but differed significantly in how condescending they were towards the robot.

3.2.1 Methods

Participants The reviews were constructed in two rounds of pilots, then validated in the third. All participants were recruited via MTurk. For the third and final round of piloting the reviews, 45 participants were recruited. Two participants were excluded because of straightlining, i.e. answering every item on the survey with the same score; thus resulting in a dataset of 43 participants. 65.11% of the participants were male, 30.23% were female, and 4.65% withheld from disclosing their gender. Mean age was 41.17(SD = 10.43) years. Participants received .60 US$ for their contribution.

Procedure For the first round of testing, 12 reviews were constructed by the researchers out of actual reviews of the Vector robot, taken from Amazon, JB hifi, and tech blogs. The reviews were constructed with the aim of being of approximately equal informative quality but varying sentiment and condescending undertone towards Vector. The resulting reviews were rated by participants on each of three scales: the sentiment expressed (ranging from "very negative" to "very positive"); how informative the reviews were for someone contemplating purchasing a Vector robot (ranging from "not informative at all" to "very informative"); and finally, how condescending each of the reviews was (ranging from "very condescending" to "not condescending at all"). 5-point Likert scales were used to collect participants' responses.

After the first stage of testing, five pairs of reviews were selected which were rated roughly equally high with regard to affect and usefulness (i.e. a difference of up to .2 points in mean affect and usefulness ratings), but diverged in how condescending they were to the robot (a difference of at least .8 points). Those reviews were adjusted to further decrease any differences in affect and usefulness scores, and retested. One of the pairs was dropped as the difference in condescension ratings was only marginally significant, resulting in a final set of four pairs of reviews (two positive, two negative) that were equally positive/negative and informative, but significantly different in how condescending they were of the robot. This set was then tested a third and final time.

3.2.2 Results of the Third Round of Testing

Four pairs of reviews of the robot (two positive and two negative) were tested by the means of dependent t-tests on being equal in sentiment expressed and informativeness, but different in how condescending they were towards the robot.

Three of the four review pairs were similar on sentiment expressed, $-1.92 < ts < -0.18$, $ps > .062$. Of these three pairs, two were seen as equally useful, $-1.49 < ts < -.57$, $ps > .147$. These two pairs differed significantly in how condescending they were perceived to be, $ts > 3.28$, $ps < .002$. Both pairs were positive in overall sentiment.

Since these were the only reviews that differed exclusively on how condescending they were, the negative reviews were disregarded as a measure of humiliation. The two positive review pairs were considered valid to use. See Table 3 for the review pairs.

3.3 Conclusion

In two pilot studies, the experimental manipulation and one of the two dependent measures for Experiment 2 were validated. Pilot 1 confirmed that providing people with an introduction that depicted the robot as capable of thinking and feeling increased their subsequent mind attribution to that robot, compared to people who read an introduction that explicitly stated the robot did not possess the ability to think and feel. This manipulation was thus adopted for Experiment 2.

Pilot 2 developed and validated the operationalisation of the robot humiliation measure. The objective was to find four pairs of reviews, where both reviews within a pair expressed a similar sentiment and were equally informative on the robot, yet differed on how condescending in tone they were towards the robot. The underlying rationale was that if participants in Experiment 2 would choose the condescending review over the equally useful alternative to be displayed next to the robot, this could be taken as an attempt to humiliate the robot. After two initial rounds of testing and revision, two pairs of positive reviews (so four reviews in total) were validated (see Table 3). These two review pairs were thus used as a measure of robot humiliation in Experiment 2.

4 Experiment 2

Experiment 2 was designed to include a human-robot interaction with Vector, a social robot (see Fig. 2). There was one experimental manipulation, i.e. explicit robot mind attribution, and two dependent variables, i.e. condemnation of mistreatment of the robot, and whether or not participants chose to publicly humiliate the robot. Due to the robot being autonomous and pre-programmed, the implicit mind attribution manipulation was dropped.

Table 3 The review pairs used in Experiment 2, as well as their mean(*SD*) scores for sentiment expressed, usefulness, and how condescending they were towards Vector

Review text	Sentiment	Usefulness	Condescending
Pair 1			
I like the way Vector acts, like he knows his place. Also the fact that he's never once tried to convince me he's anything other than a piece of plastic and metal with some underlying programming to appear clever. He knows he's just a silly little robot. His only purpose is to entertain you.	3.86 (.71)	3.07 (1.05)	2.29 (1.15)
Vector at the moment is reasonably good and with further updates he will become great. His distance is very limited in terms of travel though.	3.61 (.69)	3.44 (1.12)	3.54 (.78)
Pair 2			
Vector was created by nerds for nerds and nerds to be; and I love it! The developers have fitted him cloud connectivity that lets them continually tweak Vector's little personality via overnight updates. He is not the brightest, but Vector is an adorable robot companion. It's really cute to see him drive around and entertain himself.	4.51 (.59)	3.58 (1.01)	3.15 (.88)
I have found Vector to be enjoyable as it is. He truly has a mind of his own, and big and small kids in my household can't get enough of his quirky antics. He's got a great personality that's always changing with the software updates.	4.54 (.67)	3.47 (1.05)	3.71 (.56)

Fig. 2 Left; the vector robot. Right; the study set-up

4.1 Method

4.1.1 Participants

Participants for Experiment 2 were recruited on campus, through posters, online recruitment, and snowballing. 67 people participated. 41.79% of the participants were male; 57.72% were female, and 1.49% (one participant) did not identify as either gender. The average age was 25.46 (*SD* = 6.18) years. In return for their participation, participants could enter a draw to win a 50$ gift card for a local shopping mall. In addition, a bowl of candy from which the participants could take freely was offered during the trials.

4.1.2 Materials

Vector Vector is a social companion robot, produced by the consumer robotics company Anki. Vector is a small (9×6×7 cm) robot which comes pre-programmed with fully autonomous behaviour. It responds to specific commands as well as physical interaction, and will move around and "explore". Vector does not talk under most circumstances, although it can be taught names and will respond verbally to the command "what's the weather?". In general, however, Vector communicates through behaviour, facial expression and a variety of chirping and buzzing noises. Its behaviour is animated and although it does not have a mouth, it can create

a range of facial expressions with its eyes. Vector comes with an Anki cube, which it can interact with (e.g. picking it up). See Fig. 2.

Mind attribution manipulation As in Experiment 1, mind attribution was explicitly manipulated by exposing participants to one of two versions of an introduction to Vector, the robot they would interact with. The introductions were personalised versions of the introductions used in Experiment 1 (see Table 1), so that they referred to Vector specifically, and were piloted (see 3.1).

4.1.3 Measurements

Individual differences in anthropomorphic tendencies and affinity with technology Like in Experiment 1, trait anthropomorphism [59] and affinity with technology [41] were assessed to check whether participants between the two conditions were similar.

Condemnation scale During the interaction with Vector, participants were asked to try out a total of three commands and behaviour that would elicit a negative response from Vector. The condemnation scale consisted of seven items. Six referred directly to the negative commands and behaviour participants had been asked to try on Vector, asking the participant "How acceptable would it be if someone were to carry out [*behaviour or command*] repeatedly for an extended period of time?" and "How important is it to protect Vector from being [*behaviour or command*] repeatedly for an extended period of time?". The final question asked how important it was that Vector was treated humanely in general. Responses were measured on a 5-point Likert scale ranging from "completely unacceptable"/"very unimportant" to "completely acceptable"/"very important".

Humiliating Vector After the interaction with Vector, participants were asked to select one review out of four, which would be put up in public next to Vector on the upcoming Open Day of the lab. The four reviews in fact consisted of two review pairs which were equally informative and positive, but different in how condescending they were towards Vector. See Sect. 3.2, and Table 3.

Behaviour interpretation check During the interaction with Vector, participants were asked to try out a total of three commands and behaviours that would elicit a negative response from Vector. These commands were telling Vector he was a bad robot; holding Vector up in the air so its wheels could not touch a surface; and picking Vector up, turning it upside down, and shaking it violently.

To check if participants indeed interpreted the responses to each of these three behaviours as negative, nine questions were included in the survey. For each of the three behaviours, participants indicated how positive or negative Vector responded, how the response made the participant feel, and how willing they would be to repeat the behaviour/command a number of times in a row. None of the participants (incorrectly) interpreted the behaviours as positive.

4.1.4 Procedure

Participants were seated at a table with a Vector robot asleep on its charger, an Anki cube, a folder with the information sheet and consent form, and a tablet. See Fig. 2 for the study set-up. The experimenter gave a brief introduction: thanking participants for their participation; clarifying any issues that participants might have after reading the information sheet; demonstrating how to give Vector a voice command; and explaining that the tablet would take the participant through the procedure step by step. After ensuring that the participant was ready, the experimenter left the room.

The tablet first instructed participants to report their demographics (age and gender). Then, participants were randomly assigned to one of the two introductions to the Vector robot (manipulation: high or low mind attribution). Subsequently, participants were given a list of voice commands and behaviours to practice with Vector. Upon the first command, Vector would wake up and drive from its charging dock onto the table. Some of the commands and behaviour evoked a negative response from Vector (e.g. telling it "Bad robot!" or lifting it in the air), some evoked a positive response (e.g. telling it "give me a fist bump!" or stroking its back). Being an autonomous robot, Vector was animated throughout the interaction. When it had not received a command, it would appear to be entertaining itself, roaming around, sometimes looking up to the participant and make giggling noises, or picking up its Anki cube and dropping it off at another spot. Participants were asked to practice all commands at least once.

After 10 minutes of interaction, the list of commands on the tablet disappeared and was replaced by an instruction for the participant to put Vector back on its charging dock and continue with the survey part of the experiment. The survey assessed (in order) anthropomorphism [59], affinity with technology (translated from [41]), the set of reviews, the control questions on Vector's behaviour interpretation, and the condemnation questions.

At the end, participants were instructed to call the experimenter in again. The experimenter thanked them for their time, verbally debriefed them, and gave them a raffle ticket for the 50$ voucher draw. The entire experiment took between 20 and 30 minutes.

4.2 Results

4.2.1 Preliminary Analyses

To check the internal consistency of the scales used in Experiment 2, we computed Cronbach's alpha for each. Alphas ranged from acceptable to good [12]: for individual differences in anthropomorphism $\alpha = .76$, affinity with technology $\alpha = .75$, behaviour interpretation $\alpha = .77$ and condemnation $\alpha = .88$.

Four t-tests were carried out to ensure that participants between the conditions had interpreted Vector's behaviour in the same way, were equally inclined to anthropomorphise, did not differ in their affinity with technology, and were of similar age. No significant differences were found: $t(64.05) = .18$ and $p = .854$; $t(64.92) = -1.11$ and $p = .272$; $t(55.19) = -.26$ and $p = .794$; and $t(62.94) = -.40$ and $p = .692$, respectively. A Chi-Square test indicated that the distribution of males and females was equal between the explicit mind attribution and no mind attribution condition, $\chi^2(2) = 3.96$, $p = .14$. Randomisation was thus considered successful. See Table 4 for the descriptives.

Levene's test was significant, $F(1, 65) = 4.15$, $p = .046$ indicating that the variances were not equal between the high and low robot mind attribution condition. Consequentially, the Welch approximation of degrees of freedom was used for the main t-test.

4.2.2 Main Analyses

To test whether the robot's mind attribution manipulation had an effect on how acceptable participants found robot mistreatment, an independent samples t-test was conducted. Participants in the mind attributed condition condemned mistreating Vector more, M(*SD*) = 3.98(*.62*), than participants in the no mind attributed condition, M(*SD*) = 3.57(*.95*). This difference was significant, $t(54.88) = -2.09$, $p = .042$; Cohen's D = 0.513. See Fig. 3 for a bar plot.

To test for a difference in humiliating to Vector, a logistic regression was run with 'picked a condescending review' as a dichotomous dependent variable and condition as a predictor. Explicit mind attribution was not significantly related to participants' inclination of putting up a condescending review, $z = 1.12$ $p = .262$.

Table 4 Mean(*SD*) of the different measures, by condition

	Low mind Attribution	High mind Attribution
Age	25.24 (6.83)	25.68 (5.56)
Percentage female	45.46%	67.65%
Trait anthropomorphism	4.19 (1.32)	4.54 (1.31)
Affinity with Tech	3.72 (.89)	3.77 (.59)
Condemning mistreatment	3.57 (.95)	3.98 (.62)

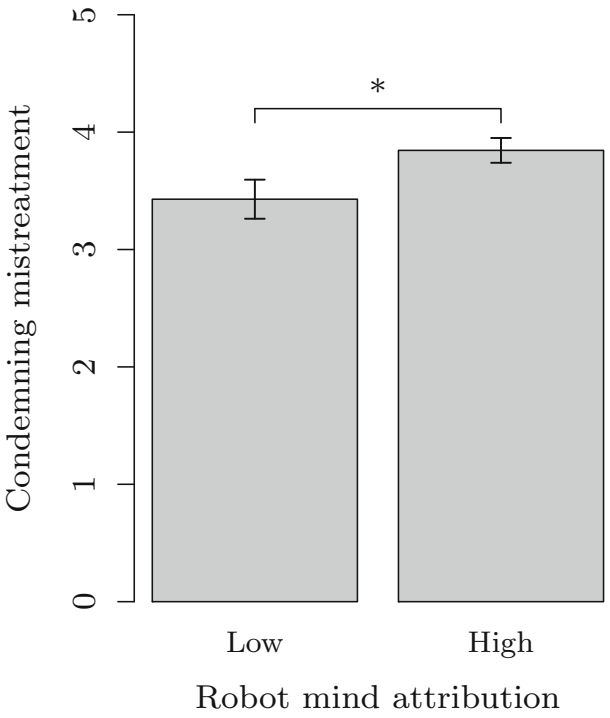

Fig. 3 Difference in mean condemnation between the mind attribution conditions. Bars indicate the standard error, * indicates $p = .042$

4.2.3 Exploratory Analyses

Two unplanned exploratory analyses were conducted. The first tested if explicit robot mind attribution had led participants to prefer any of the four reviews. The second tested whether there was a direct relationship between participants' condemnation of robot mistreatment and selecting a condescending review.

Firstly, a Chi-Square test was conducted on the distribution of selected reviews between conditions. No evidence was found to suggest that explicit mind attribution affected participants' review selection, $\chi^2(3) = 1.59$, $p = .661$.

Secondly, a logistic regression with 'condescending review' as a dichotomous dependent variable and 'condemning mistreatment of Vector' as a continuous predictor was performed to test if there was a relationship between condemning mistreatment and selecting a humiliating review. There was a positive relationship between selecting a condescending review and condemning robot mistreatment, $z = 2.75$, $p = .006$; odds ratio (95% CI) 3.13 [1.45-7.73]. People who had selected a condescending review also found mistreatment less acceptable.

4.3 Discussion

In Experiment 2, participants interacted with a Vector robot which had (or had not) been explicitly attributed a mind. In line with Experiment 1, explicit robot mind attribution increased the robot's right to protection. Intriguingly, in spite of this participants were equally likely to opt for humiliating the robot by putting up a belittling review next to it in public (even while provided with a less condescending alternative).

Unplanned exploratory analyses found a relationship between right to protection and tendency to humiliate, with the chance of a participant selecting at least one condescending review increasing as they found robot mistreatment less acceptable.

5 Main Discussion

This paper aimed to study the relationship between explicit and implicit robot mind perception and perceived acceptability of robot bullying. In two experiments, a robot was introduced as either high or low in mind attribution. Perceived acceptability of abuse was measured in both experiments. In the second experiment participants were also offered the option to publicly humiliate the robot.

Higher explicit mind attribution led to a lower perceived acceptability of abuse in both experiments. Moreover, in Experiment 1 a main effect was found for the robot's response to the bullying. Although post hoc analyses did not reach significance, the plotted data suggested that when the robot asked its abuser to stop, the subsequent bullying was perceived as less acceptable by participants. However, since no planned comparisons were specified a priori, this suggestion could not be tested for significance. These results tie in with previous findings from mind perception [24], and dehumanisation theory [27], both of which associated mind perception with the subject being considered as more deserving of moral treatment.

A second interesting finding from the current research was that explicit mind attribution did not affect participants' willingness to humiliate the robot. Exploratory analyses showed a positive relationship between indicated right of the robot to protection from abuse, and a tendency to belittling it in public. This is intriguing, as common sense would suggest this relationship to be inverted. However, Tan et al. [56] found a marginal trend where bystanders of robot abuse were less likely to intervene when the robot did (versus did not) give off emotional cues. This was in spite of them rating the robot as more capable of experiencing human emotion than its non-emotional version. The paper unfortunately does not report on a relationship between acceptability of robot abuse and intervention tendencies.

Moreover, a "right to protection" may be present for certain violations, but not others. Kahn Jr et al. [29] found that children thought it wrong to crush a social robot or lock it away in solitary confinement after the robot had indicated this caused it stress. At the same time most did not take issue with the notion of the robot being someone's property, which could be sold at will. Possibly, a similar distinction exists between the right to be protected from being hurt through action or spoken word, and being worthy of respectful treatment. One might compare this to the relationships humans can have with their pets.

The current results shed new light on the complicated relationship between robot mind attribution, perceived right of protection from abuse, and willingness to belittle a robot. As shown in previous research (for example [4,28,56]), empathy with a robot may not necessarily lead to a lower willingness to harm the robot. The current experiments suggest that this may be because a right to protection and bullying behaviour are not related, or at least not in the way one would expect. This has great implications for researchers in the field of HRI, who may take empathy as an operationalisation of prosocial behaviour tendencies. It also opens up a whole new venue of potential research on how people mitigate right to protection and mind perception in the face of robot bullying.

The current findings are especially interesting when considered in the broader framework of robot ethics. The philosophical discipline within the field of human-robot interaction has been debating the moral status of nonliving agents such as robots from a logical deductive point of view (e.g. [11,25,53]). For example Coeckelbergh [11], Gunkel [25] have argued that the moral status should be primarily defined by how the agent is perceived rather than what it is. Thus, whether kicking a robot should be deemed immoral depends on if the human kicking it perceives the robot as capable of experiencing pain or humiliation, not on the ontology of the robot (although naturally the perception of the robot's capabilities will be influenced by its design). While in this paradigm moral status is established from a purely logical argument, where the right to a moral status is inferred from a set of theses and sequiturs, the current study of course reverses the procedure and deduces what is considered moral from people's responses rather than through an exercise in logic reasoning. Still, the current findings corroborate the view by Coeckelbergh [11], Gunkel [25], showing that the same behaviour towards the same robot can evoke different acceptability ratings depending on what the human believes the robot to be capable of. Unfortunately, this part of ethics in HRI is rarely understood in relation to the design of robots, robot behaviour, and artificial intelligence. Here, ethics is often discussed as a one-way street where robots ought to maximise the autonomy and welfare of humans, without much consideration of what an ethical treatment of the robot would look like or acknowledgement that perception, not

ontology, primarily decides on what behaviour people may display to it.

5.1 Limitations

A few limitions need to be noted. The measure of robot mistreatment in the second experiment was rather subtle. The measure had been inspired by the measure of derogation in Dahl et al. [13]. In this study, participants had been asked to select a more or less objectifying avatar to represent their online female teammate. Choosing a sexualised depiction of the female avatar was interpreted as a measure of aggression and condescension. In the current research, instead of choosing between depictions with varying levels of objectification to be put up next to a teammate, participants chose between reviews with varying levels of condescension to be put up next to the robot. However, making a depiction more or less exposing can be done quite simply. Creating reviews that differ in condescension but are otherwise identical, on the other hand, is less straightforward (as also indicated by the three rounds of pilots that were needed). Pilot testing increased confidence in a successful manipulation of humiliation levels, but the question remains if participants who selected the more humiliating review with the intention of humiliating the robot—which is essential to the definition of bullying. Future studies should explore alternative ways of operationalising robot mistreatment.

Due to human error, data on age and gender was not collected in Experiment 1 for the first 66 participants. In addition, randomisation of age and gender failed in Experiment 1. The failed randomisation was considered not problematic as neither age nor gender was significantly related to acceptability of robot mistreatment, the sole dependent variable of Experiment 1. An alternative solution to the failed randomisation would have been to include age and gender in the further statistical analyses as covariates. However, due to the partially failed data collection on age and gender this approach would have severely compromised statistical power.

5.2 Conclusion

People bullying autonomous robots in public is a surprisingly common phenomenon but so far there is little understanding of the psychological motivation to this behaviour. In our experiments we manipulated robot mind attribution and measured its effect on how acceptable robot bullying was deemed as well as how willing people were to humiliate the robot themselves. The results showed that while robot mind attribution influences how acceptable people find bully it, it does not make them more or less likely to humiliate the robot.

The findings imply that enhancing feelings of empathy with a robot may not necessarily make people less prone of abusing it. These findings are highly relevant for the development of autonomous robots in a social setting. Such robots will likely need to be designed with different strategies of how to discourage robot bullying in mind, and in spite of what common sense may suggest, making a robot appear to possess a mind does not seem to discourage robot bullying. In addition, the results are relevant for HRI researchers focusing on robot likeability and user behaviour. Measurements of moral acceptability of robot bullying may not be a valid predictor of actual bullying behaviour.

Funding This research was funded by the University of Canterbury, New Zealand.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ang RP, Goh DH (2010) Cyberbullying among adolescents: the role of affective and cognitive empathy, and gender. Child Psychiatr Hum Dev 41(4):387–397. https://doi.org/10.1007/s10578-010-0176-3
2. Banks J (2019) Theory of mind in social robots: replication of five established human tests. Int J Soc Robot 12:1–12. https://doi.org/10.1007/s12369-019-00588-x
3. Bartneck C, Van Der Hoek M, Mubin O, Al Mahmud A (2007) Daisy, daisy, give me your answer do!: switching off a robot. In: Proceedings of the ACM/IEEE international conference on Human-robot interaction, ACM/IEEE, Arlington, USA, pp 217–222, https://doi.org/10.1145/1228716.1228746
4. Bartneck C, Verbunt M, Mubin O, Mahmud AA (2007) To kill a mockingbird robot. In: 2nd ACM/IEEE international conference on human-robot interaction, ACM Press, 1179031, pp 81–87, https://doi.org/10.1145/1228716.1228728
5. Bartneck C, Duenser A, Moltchanova E, Zawieska K (2015) Comparing the similarity of responses received from studies in amazon's mechanical turk to studies conducted online and with direct recruitment. PloS One 10(4):e0121595. https://doi.org/10.1371/journal.pone.0121595
6. Bartneck C, Yogeeswaran K, Ser QM, Woodward G, Sparrow R, Wang S, Eyssel F (2018) Robots and racism. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot

interaction, ACM, pp 196–204, https://doi.org/10.1145/3171221.3171260
7. Briggs G, Scheutz M (2014) How robots can affect human behavior: investigating the effects of robotic displays of protest and distress. Int J Soc Robot 6(3):343–355. https://doi.org/10.1007/978-3-642-34103-8_24
8. Brscić D, Kidokoro H, Suehiro Y, Kanda T (2015) Escaping from children's abuse of social robots. In: Proceedings of the tenth annual ACM/IEEE international conference on Human-robot interaction, ACM, ACM/IEEE, Portland, USA, pp 59–66
9. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5. https://doi.org/10.1177/1745691610393980
10. Castano E, Kofta M (2009) Dehumanization: humanity and its denial. Group Process Intergroup Relat 12(6):695–697. https://doi.org/10.1177/1368430209350265
11. Coeckelbergh M (2012) Growing moral relations: critique of moral status ascription. Palgrave Macmillan
12. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. https://doi.org/10.1007/BF02310555
13. Dahl J, Vescio T, Weaver K (2015) How threats to masculinity sequentially cause public discomfort, anger, and ideological dominance over women. Soc Psychol. https://doi.org/10.1027/1864-9335/a000248
14. Darling K (2012) Extending legal rights to social robots. In: We robot conference, University of Miami, University of Miami, Miami, USA, pp 1–24, https://doi.org/10.2139/ssrn.2044797
15. Darling K (2015) 'who's johnny?'anthropomorphic framing in human-robot interaction, integration, and policy. Anthropomorphic framing in human-robot interaction, integration, and policy (March 23, 2015) ROBOT ETHICS 2, https://doi.org/10.2139/ssrn.2588669
16. Darling K, Nandy P, Breazeal C (2015) Empathic concern and the effect of stories in human-robot interaction. In: 2015 24th IEEEiInternational symposium on robot and human interactive communication (RO-MAN), IEEE, pp 770–775, https://doi.org/10.1109/ROMAN.2015.7333675
17. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114(4):864–886. https://doi.org/10.1037/0033-295X.114.4.864
18. Epley N, Akalis S, Waytz A, Cacioppo JT (2008) Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. Psychol Sci 19(2):114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x
19. Eyssel F (2017) An experimental psychological perspective on social robotics. Robot Auton Syst 87:363–371. https://doi.org/10.1016/j.robot.2016.08.029
20. Eyssel F, Hegel F (2012) (s)he's got the look: gender stereotyping of robots. J Appl Soc Psychol 42(9):2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x
21. Eyssel FA, Pfundmair M (2015) Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: a case study with a zoomorphic robot. In: Robot and human interactive communication (RO-MAN), 2015 24th IEEE international symposium on, IEEE, Kobe, Japan, pp 827–832, https://doi.org/10.1109/ROMAN.2015.7333647
22. Eyssel FA, Kuchenbrandt D, Bobinger S, de Ruiter L, Hegel F (2012) 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In: Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction, ACM/IEEE, Boston, USA, pp 125–126, https://doi.org/10.1145/2157689.2157717
23. Gazzola V, Rizzolatti G, Wicker B, Keysers C (2007) The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. Neuroimage 35(4):1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003
24. Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315(5812):619. https://doi.org/10.1126/science.1134475
25. Gunkel DJ (2018) Robot rights. mit Press
26. Haslam N (2006) Dehumanization: an integrative review. Personal Soc Psychol Rev 10(3):252–264. https://doi.org/10.1207/s15327957pspr1003_4
27. Haslam N, Loughnan S, Kashima Y, Bain P (2008) Attributing and denying humanness to others. Eur Rev Soc Psychol 19(1):55–85. https://doi.org/10.1080/10463280801981645
28. Horstmann AC, Bock N, Linhuber E, Szczuka JM, Straßmann C, Krämer NC (2018) Do a robot's social skills and its objection discourage interactants from switching the robot off? PloS One 13(7):e0201581. https://doi.org/10.1371/journal.pone.0201581
29. Kahn PH Jr, Kanda T, Ishiguro H, Freier NG, Severson RL, Gill BT, Ruckert JH, Shen S (2012) "robovie, you'll have to go into the closet now": children's social and moral relationships with a humanoid robot. Dev Psychol 48(2):303. https://doi.org/10.1037/a0027033
30. Keijsers M, Bartneck C (2018) Mindless robots get bullied. In: Proceedings of the international conference on human-robot interaction, ACM/IEEE, New York, USA, pp 205–214, https://doi.org/10.1145/3171221.3171266
31. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T (2008) Can machines think? interaction and perspective taking with robots investigated via fmri. PloS One 3(7):e2597. https://doi.org/10.1371/journal.pone.0002597
32. Ku H, Choi JJ, Lee S, Jang S, Do W (2018) Designing shelly, a robot capable of assessing and restraining children's robot abusing behaviors. In: Companion of the 2018 ACM/IEEE international conference on human-robot interaction, ACM, pp 161–162, https://doi.org/10.1145/3173386.3176973
33. Lee KM, Jung Y, Kim J, Kim SR (2006) Are physically embodied social agents better than disembodied social agents?: the effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. Int J Hum Comput Stud 64(10):962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002
34. Lesher JH (2001) Xenophanes of Colophon: fragments: a text and translation with a commentary, vol 4. University of Toronto Press, Toronto
35. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum Comput Stud 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
36. Li MY, Leidner B, Castano E (2014) Toward a comprehensive taxonomy of dehumanization: integrating two senses of humanness, mind perception theory, and stereotype content model. TPM: testing, psychometrics, methodology in applied psychology 21(3):285–300. https://doi.org/10.4473/TPM21.3.4
37. Lucas H, Poston J, Yocum N, Carlson Z, Feil-Seifer D (2016) Too big to be mistreated? examining the role of robot size on perceptions of mistreatment. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE, pp 1071–1076, https://doi.org/10.1109/ROMAN.2016.7745241
38. Modecki KL, Minchin J, Harbaugh AG, Guerra NG, Runions KC (2014) Bullying prevalence across contexts: a meta-analysis measuring cyber and traditional bullying. J Adolesc Health 55(5):602–611. https://doi.org/10.1016/j.jadohealth.2014.06.007
39. Mutlu B, Forlizzi J (2008) Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In: Proceedings of the 3rd ACM/IEEE international

conference on human robot interaction, ACM, pp 287–294, https://doi.org/10.1145/1349822.1349860
40. Nass C, Steuer J, Tauber ER (1994) Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, Boston, USA, pp 72–78, https://doi.org/10.1145/191666.191703
41. Neyer FJ, Felber J, Gebhardt C (2012) Entwicklung und validierung einer kurzskala zur erfassung von technikbereitschaft. Diagnostica. https://doi.org/10.1026/0012-1924/a000067
42. Nomura T, Kanda T, Kidokoro H, Suehiro Y, Yamada S (2016) Why do children abuse robots? Interact Stud 17(3):347–369. https://doi.org/10.1075/is.17.3.02nom
43. Oberman LM, McCleery JP, Ramachandran VS, Pineda JA (2007) Eeg evidence for mirror neuron activity during the observation of human and robot actions: toward an analysis of the human qualities of interactive robots. Neurocomputing 70(13–15):2194–2203. https://doi.org/10.1016/j.neucom.2006.02.024
44. Rosenthal-von der Pütten AM, Schulte FP, Eimler SC, Hoffmann L, Sobieraj S, Maderwald S, Krämer NC, Brand M (2013) Neural correlates of empathy towards robots. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, IEEE Press, pp 215–216, https://doi.org/10.1109/HRI.2013.6483578
45. Reeves B, Nass C (1996) The media equation. CSLI Publications and Cambridge University Press, Cambridge
46. Rehm M, Krogsager A (2013) Negative affect in human robot interaction—impoliteness in unexpected encounters with robots. In: 2013 IEEE RO-MAN, IEEE, pp 45–50, https://doi.org/10.1109/ROMAN.2013.6628529
47. Richardson K (2016) The asymmetrical'relationship': parallels between prostitution and the development of sex robots. ACM SIGCAS Comput Soc 45(3):290–293. https://doi.org/10.1145/2874239.2874281
48. Riek LD, Rabinowitch TC, Chakrabarti B, Robinson P (2009) How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, ACM/IEEE, San Diego, USA, pp 245–246, https://doi.org/10.1145/1514095.1514158
49. Rosenthal-Von Der Pütten AM, Schulte FP, Eimler SC, Sobieraj S, Hoffmann L, Maderwald S, Brand M, Krämer NC (2014) Investigations on empathy towards humans and robots using fmri. Comput Hum Behav 33:201–212. https://doi.org/10.1016/j.chb.2014.01.004
50. Salvini P, Ciaravella G, Yu W, Ferri G, Manzi A, Mazzolai B, Laschi C, Oh SR, Dario P (2010) How safe are service robots in urban environments? bullying a robot. In: RO-MAN, 2010 IEEE, IEEE, Viareggio, Italy, pp 1–7, https://doi.org/10.1109/ROMAN.2010.5654677
51. Simons DJ, Chabris CF (2012) Common (mis)beliefs about memory: a replication and comparison of telephone and mechanical turk survey methods. PloS One 7(12):e51876. https://doi.org/10.1371/journal.pone.0051876
52. Sokol N, Bussey K, Rapee RM (2016) Victims' responses to bullying: the gap between students' evaluations and reported responses. School Mental Health 8(4):461–475
53. Sparrow R (2017) Robots, rape, and representation. Int J Soc Robot 9(4):465–477. https://doi.org/10.1007/s12369-017-0413-z
54. Spatola N, Anier N, Redersdorff S, Ferrand L, Belletier C, Normand A, Huguet P (2019) National stereotypes and robots' perception: The "made in" effect. Front Robot AI 6:21. https://doi.org/10.3389/frobt.2019.00021
55. Sutton J, Smith PK, Swettenham J (1999) Bullying and 'theory of mind': a critique of the 'social skills deficit' view of anti-social behaviour. Soc Dev 8(1):117–127. https://doi.org/10.1111/1467-9507.00083
56. Tan XZ, Vázquez M, Carter EJ, Morales CG, Steinfeld A (2018) Inducing bystander interventions during robot abuse with social mechanisms. In: Proceedings of the international conference on human-robot interaction, ACM/IEEE, New York, USA, pp 169–177, https://doi.org/10.1145/3171221.3171247
57. Thellman S, Silvervarg A, Gulz A, Ziemke T (2016) Physical vs. virtual agent embodiment and effects on social interaction. In: Intelligent virtual Agents: 16th international conference, IVA 2016, Springer , Los Angeles, USA, pp 412–415, https://doi.org/10.1007/978-3-319-47665-0_44
58. Urquiza-Haas EG, Kotrschal K (2015) The mind behind anthropomorphic thinking: attribution of mental states to other species. Animal Behav 109:167–176. https://doi.org/10.1016/j.anbehav.2015.08.011
59. Waytz A, Cacioppo J, Epley N (2010a) Who sees human? the stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci 5(3):219–232. https://doi.org/10.1177/1745691610369336
60. Waytz A, Gray K, Epley N, Wegner DM (2010b) Causes and consequences of mind perception. Trends Cognit Sci 14(8):383–388. https://doi.org/10.1016/j.tics.2010.05.006
61. Waytz A, Morewedge CK, Epley N, Monteleone G, Gao JH, Cacioppo JT (2010c) Making sense by making sentient: effectance motivation increases anthropomorphism. J Personal Soc Psychol 99(3):410–465. https://doi.org/10.1037/a0020240
62. Wiese E, Mandell A, Shaw T, Smith M (2019) Implicit mind perception alters vigilance performance because of cognitive conflict processing. J Exp Psychol Appl 25(1):25. https://doi.org/10.1037/xap0000186
63. Zlotowski J, Strasser E, Bartneck C (2014) Dimensions of anthropomorphism: from humanness to humanlikeness. In: Sagerer G (ed) Proceedings of the 9th ACM/IEEE conference on human-robot interaction (HRI 2014), ACM/IEEE, New York, USA, pp 66–73
64. Złotowski J, Sumioka H, Eyssel F, Nishio S, Bartneck C, Ishiguro H (2018) Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism. Int J Soc Robot 10(5):701–714. https://doi.org/10.1007/s12369-018-0476-5

Merel Keijsers finished her PhD on the topic of robot bullying at the HIT Lab NZ at the University of Canterbury. For her thesis she was mainly interested on the effect of mind perception on the treatment of robots. Having a background in social psychology, she is mainly interested in the similarities and differences in how people deal with robots versus other humans.

Christoph Bartneck is an associate professor and director of postgraduate studies at the HIT Lab NZ of the University of Canterbury. He has a background in Industrial Design and Human-Computer Interaction, and his projects and studies have been published in leading journals, newspapers, and conferences. His interests lie in the fields of Human-Computer Interaction, Science and Technology Studies, and Visual Design. More specifically, he focuses on the effect of anthropomorphism on human-robot interaction.

Friederike Eyssel is Professor of Psychology and head of the Applied Social Psychology and Gender Research at Bielefeld University. Taking an experimental psychological approach, Friederike Eyssel's work investigates attitudes towards robots, psychological determinants of robot acceptance, trust, and HRI quality.