International Journal of Social Robotics (2021) 13:41–54
https://doi.org/10.1007/s12369-019-00608-w

Teaching Robots a Lesson: Determinants of Robot Punishment

Merel Keijsers[1] · Hussain Kazmi[3] · Friederike Eyssel[2] · Christoph Bartneck[1]

Accepted: 8 November 2019 / Published online: 5 December 2019

Abstract
There have been multiple incidents where humans attacked robots in a public environment (Brscić et al., in: Proceedings of the international conference on human–robot interaction, ACM/IEEE, Portland, 2015, https://doi.org/10.1145/2696454.2696468); Vincent, in: A drunk man was arrested for knocking over Silicon Valley's crime-fighting robot, 2017, https://www.theverge.com/2017/4/26/15432280/security-robot-knocked-over-drunk-man-knightscope-k5-mountain-view; Mosbergen, in: Good job, America. You killed hitchBOT. Huffpost, 2015, https://www.huffpost.com/entry/hitchbot-destroyed-philadelphia_n_55bf24cde4b0b23e3ce32a67; Mutlu and Forlizzi, in: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, ACM, 2008, https://doi.org/10.1145/2696454.2696468; Rehm and Krogsager, in: 2013 Proceedings of IEEE RO-MAN, IEEE, 2013, https://doi.org/10.1145/2696454.2696468; (Salvini et al., in: 19th International symposium in robot and human interactive communication, 2010). Although the form of aggression suggests that this behaviour might be motivated by the aggressor's desire for social recognition rather than an urge for vandalism (Salvini et al. 2010; Keijsers and Bartneck, in: Proceedings of the international conference on human–robot interaction, ACM/IEEE, New York, 2018, https://doi.org/10.1145/2696454.2696468), very little is known about the underlying psychological mechanisms. Therefore, extending previous research, the current study investigated if human aggression towards a robot would be influenced by the aggressor's feelings of power, the perception of the threat that robots in general might pose, mind attribution to the robot, and the robot's embodiment. First, threat and power were manipulated. Subsequently, participants played a learning task with either a virtual or an embodied robot. Mind attribution was measured afterwards. Participants were asked to restrict the robot's energy supply after each wrong answer, which was taken as a measure of aggression. Results indicated that an embodied robot was punished less harshly than a virtual one, except for when people had been primed with power and threat. Being primed with power diminished the influence of mind attribution. Mind attribution increased aggression in the threat condition but was related to decreased aggression when people had not been reminded of threat.

Keywords Human–Robot interaction · Robot threat · Robot aggression · Mind attribution · Robot embodiment

1 Introduction

In 2017, an intoxicated man was arrested for assault in a car park in Mountain View, California. A local commented on the incident, stating "*I think this is pretty pathetic (…) because it shows how spineless drunk guys (…) really are because they attack a victim who doesn't even have any arms. I don't think this is a fair fight, really totally unacceptable.*" Fortunately the victim, a K5 Knightscope robot (Fig. 1), only suffered minor scratches [60]. This anecdote confirms two consistent findings from the field of human–robot interaction (HRI).

Firstly, it illustrates that social robots are potential targets of verbal and physical abuse [12,54]. In spite of researchers' best efforts to design behaviours which discourage such behaviour, robot-directed aggression has been shown to

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

Hussain Kazmi
hussain.kazmi@enervalis.com

Friederike Eyssel
feyssel@cit-ec.uni-bielefeld.de

[1] Human Interface Technology lab NZ, University of Canterbury, Christchurch, New Zealand

[2] Center of Excellence Cognitive Interaction Technology, Bielefeld University, Bielefeld, Germany

[3] Department of Electrical Engineering, KU Leuven, Leuven, Belgium

Fig. 1 The K5 Knightscope robot (source: [40])

be remarkably persistent [see for example [5,12,48,54]]. Prevention strategies to reduce robot-directed aggression include making the design of the robot so robust that it simply cannot be damaged [54]; having the robot shut down completely until the abusive behaviour stops [59]; or having the robot run away from humans that are under 1.40m tall, since children are more likely than adults to engage in abuse [12]. While some of these strategies have shown to be moderately successful in reducing robot abuse [12], they are also rather blunt, and none of them is particularly useful in a situation where robots become ubiquitous, such as autonomous driving cars or robots patrolling public spaces.

A second common finding in HRI research is reflected in the reasoning of the local when describing the incident as pathetic. The local wasn't upset with the offender because he had gotten drunk enough to pick a fight with a lifeless object. Rather, it was because the victim had no arms and thus couldn't defend itself. This illustrates that humans seem hard-wired to be sensitive to social cues in behaviour as well as appearance, insofar that they automatically perceive and respond to a robot as a social being even in the absence of humanlike cues [see for example [36,50]]. Some authors [54] have argued that this also shows in the form the aggression takes and that therefore 'robot bullying' is a more appropriate term than 'robot vandalism'. Vandalising robots would have as main objective to damage the robot, and one would thus expect people to set fire to them, key them, or attempt to crash their interface. Instead, humans assault robots in a similar way as they abuse sentient creatures—by kicking and insulting them or trying to force their actions [12,54].

Robot abuse thus appears to be a social behaviour, governed by psychological processes. However, little is currently known about the psychological motivations behind robot abuse. Initial studies pitched an evolutionary explanation [19] while others suggest that disinhibition for aggressive behaviour to occur when "the illusion of anthropomorphism shatters" and the user suddenly stops seeing the robot as a social agent [6,18]. Bartneck et al. [8] hypothesised that abuse might be caused by frustration, if a robot won't not respond as expected.

Further research is needed to shed light on the determinants of aggressive behaviour towards robots, particularly to enable effective interventions [see also [10,21]]. The current research contributes by combining findings from human–human aggression studies with HRI research on the effect of attitudes on negative behaviour. More specifically, by extending previous work on robot abuse [34] we identified and empirically tested four factors that might play a role in robot aggression: the power of the users over the robot, the perceived threat of robots, the robot's embodiment, and the user's attribution of mind towards the robot.

1.1 Related Work

Humans see robots as social agents. This has been shown at the neurological [26,36,51], physiological [50,57], affective [50] and behavioural level [11,14]. Their status as social agents has been reported in both controlled lab settings [e.g. [11,50,57]] and in the less predictable outside world [e.g. [12,27,33]].

On the neurological level, it has been shown that similar brain areas are activated when people watch a human or a robot perform a task [26], or when engaging in social interaction with either a robot or a human partner [36]. At the physiological level, participants' heart rates and skin conductance level increased when they had to administer increasingly heavy shocks to a virtual agent in an adaptation of Milgram's obedience studies [57], indicating increased arousal when participants had to "punish" the agent. The display of distress was echoed in participants' self-reported stress levels. People talk back to robots as if they would understand what's being said [7] and try to keep them safe from harm, even though they rationally acknowledge that the robot would be incapable of feeling and does not possess awareness [17]. In short, humans respond with social cognition, social affect, and social behaviour when interacting with robots.

If humans automatically see robots as social beings, including assigning them a moral status, then why do they get aggressive to robots in the wild? It has been suggested that the consequences of the recognition of robots as social agents extend to the field of robot abuse [5,34,54]. Because a robot is perceived as a social agent, aggression towards robots might be predicted by the same factors that drive aggression between humans. Indeed, a recent study [34] looked at whether aggression towards robots could be explained

through dehumanisation, a major concept in psychological research on aggression between humans [30,32].

Dehumanisation is the psychological process by which humans perceive their victims as slightly less capable of thinking and feeling, which decreases the moral standing of the victim and allows the perpetrator to disregard the negative consequences of their own behaviour. As a result, the threshold for inflicting pain (both physical and mental) on others is lowered [30,31]. The tendency to dehumanise is partially determined by stable factors like personality traits of the aggressor [e.g. narcissism, extraversion; [37,43]] and characteristics of the victim [e.g. gender, social class; [4,53]], but also by situational aspects, such as a feeling of social connection [62] or a sense of power [29,38].

Research on the role of dehumanisation in human–human interaction has shown that reduced mind attribution [i.e. the perceived capability to think and feel; [31,35]] is related to an increase in aggression [31,41]. The same relationship has been observed in human–robot interaction, where lower mind attribution to a robot was found to be related to an increase in the number of rude comments people made to it [34]. This suggests that the same fundamental psychological mechanisms may apply to human and robot aggression.

However, although mind attribution and abuse were found to be related, the study by Keijsers and Bartneck [34] had some shortcomings and unexpected findings: for example, inducing feelings of power failed to influence mind attribution and *de*creased, rather than increased, derogative behaviour towards the robot. This was surprising as power is a well-established prime for dehumanising behaviour in human–human interaction [24,25,29], and a power imbalance (with the bully having a position of power over the victim) is one of the defining qualities of bullying [45, 61].

Plausibly, participants in [34] study might have felt threatened after being confronted with the robot. In previous research, encountering robot automatically triggered thoughts of both pragmatic ("robots will steal our jobs!") and innate ("if a robot can do everything a human can do, then what makes us humans special?") threat in people [66]. Such feelings could elicit aggressive behaviour [31]. At the same time, activating an individual's sense of power has been shown to make people less sensitive to threats from outgroups [15]. Thus, inducing a sense of power could have decreased aggression towards robots by reducing the perceived threat. These suggestions remain to be empirically tested.

A second shortcoming of [34] was that it was conducted online, with a virtual rather than an embodied robot, raising questions about the generalisability of the results. Previous research on human–human bullying has suggested that on- and offline bullying do not differ on a conceptual level, as reported by both perpetrators and victims [45]. People are however more likely to bully online than offline [44], supposedly because the online environment reduces inhibition and self-consciousness in participants [58]. This would be the result of both aggressor and victim being anonymous and invisible, and a lack of bystanders who could intervene [39,58]. These factors lower the threshold for aggression between humans [62] as well as aggression towards a virtual robot [19]. Keijsers and Bartneck [34] thus assumed that using an online platform might enhance, but would not alter the effect that other factors have on bullying tendencies towards robots. That being said, literature on robot embodiment is still mixed on whether embodied and virtual robots elicit similar responses [42], and whether the results from [34] generalise to an embodied robot remains a question to be answered.

1.2 Current Research

The current experiment replicated and extended the study by Keijsers and Bartneck [34]. More specifically, it aimed to further explore the roles of power, threat, embodiment, and mind attribution in robot directed aggression. Feelings of power and threat in participants were manipulated, and subsequently punishment behaviour in a learning task with either a virtual or embodied Nao robot was measured as an operationalisation of aggression. While the raw punishment scores cannot be equalled to a measure of aggression, a relative difference in how harsh participants punished their robot between the different conditions should allow for inferences on how justified the participants felt to aggress. Since the manipulation of mind attribution by power priming failed in the previous experiment, the current study included both a power manipulation check and a measure of mind attribution. Unless mind attribution would be manipulated by power, it was to be included in the multiple linear regression model as a covariate rather than a factor.

We hypothesised that participants would be milder in their punishments when feeling powerful and unthreatened, and that they would be particularly harsh when feeling powerless and threatened. We furthermore predicted a main effect for embodiment, in the sense that people would punish the virtual robot more harshly than an embodied robot. However, embodiment was not expected to influence the effects of power and threat (i.e. no interaction effects).

In line with the literature on aggression and dehumanisation, we predicted that mind attribution would be negatively related to robot punishment. Since there had been an interaction between power and mind attribution in the previous study, we hypothesised power would reduce the influence of mind attribution on punishment. Finally, we predicted the negative relation between mind attribution and punishment to be particularly strong when people felt threatened.

2 Methods

2.1 Design and Participants

A 2 (reminder of robot threat: present or absent) × 2 (sense of power: high or low) × 2 (robot embodiment: virtual or embodied) between participants design was realised. Mind attribution was measured by a questionnaire and used as a continuous independent variable. The dependent variable was robot punishment.

148 participants signed up for the virtual robot condition via MTurk. Five participants failed both attention checks and were excluded. The resulting data set thus contained 143 participants with a mean age of 40.34 years (SD = 11.03), and with slightly more females (57%) than males (42%). The majority (97%) were US residents. Participants in the virtual robot condition were originally rewarded with 1 US$ for their participation. When the data collection stagnated after 89 participants, payment was raised to 1.15 US$. The increase in payment did not influence aggression, mind attribution, feelings of power and robot threat (see 3.2 for the statistical tests).

82 participants were assigned to the embodied robot condition. Due to technical issues, the data of only 74 participants were usable for subsequent analysis. Participants were recruited through poster advertising on campus, posting on several student Facebook pages, and snowball sampling. Data collection on age and gender occurred after the experiment via email (with a link to a web page where the data could be left anonymously) as these demographics had not been assessed initially. The mean age of participants who responded to the post-experimental email (77% of the sample) was 27.68 (SD = 6.90) years, with the majority being female (63% female, 30% male, 7% 'rather not say'). Participants in this condition were reimbursed with a 10NZ$ (≈ 6.65 US$) voucher for a local shopping mall.

The monetary compensation in both conditions was based on the reward conventions within each setting.

2.2 Experimental Manipulations

Threat

Threat is commonly manipulated by providing participants with information on supposed threat levels [see for example [23,64]] and has been successfully applied to the field of HRI [66]. In the current research, threat was primed through a video which was shown at the start of the experiment. The first two minutes of video were neutral in tone and identical across conditions.[1,2] Participants in the threat condition saw an additional 20 seconds of material at the end of the video, where the narrator mentioned concerns regarding robots replacing humans on the work floor, and how prominent figures such as Elon Musk and the late Stephen Hawking had warned against the unrestricted development of AI. The video images were adapted from the YouTube video *What is a robot?* [67]; the narration was done by a native English speaker.

Power

Feelings of power were manipulated by assigning participants the role of teacher (i.e. indicating power) or assistant (i.e. indicating compliance). The teachers decided for themselves on the correct answer for each trial, whereas assistants had to conform to what was provided as the right answer, regardless of whether they agreed or not. In addition, assistants were reminded of their subordinate role every time they had to provide feedback. In both power and compliance conditions, participants were free to choose their level of punishment for the robot.

The manipulation was based on the design of Study one in Galinsky et al. [24], where participants were primed with power (respectively submission) by being told they would act like a manager (respectively builder) in a subsequent task, and that they would decide on the right building procedure (respectively had to conform to instructions).

Embodiment

Embodiment was manipulated through the method of data collection. Participants for the virtual robot condition signed up via MTurk and completed the experiment online. Previous studies have indicated that data collected via MTurk is of equal quality as on-campus recruitment or participant data from forums [9,56], with internal motivation rather than monetary reward being the main motive for participating [13]. Participants for the embodied robot condition were recruited and completed the experiment on site, with an embodied Nao V5 robot instead of a virtual one.

The virtual and embodied robot conditions differed strongly in terms of sample size. Unequal sample sizes are not necessarily problematic, but do render some statistical tests more sensitive to heteroscedasticity of variance [22]. Thus, in the Results section homogeneity of variance is explicitly addressed.

2.3 Procedure

Virtual Robot Condition

Participants were recruited on MTurk and redirected to the experiment website. On the first screen, they were asked to

[1] Threat condition video: https://youtu.be/GquL-MofDbg.

[2] Control condition video: https://youtu.be/8rdV4Ah8TI8.

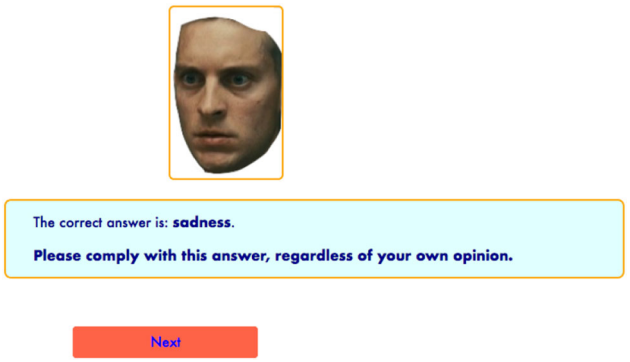

Fig. 2 Face stimulus for the low power/virtual robot condition

Fig. 3 Screenshot of the virtual robot giving its "guess" of the emotion displayed on the face stimulus, and the slider with which the participant can allocate energy

enter their demographics and were presented with a link to the information sheet and informed consent. After providing consent, participants had to turn up the volume for the introduction video, which was a short animation with narration. It was either neutral in tone (control condition) or included a warning on the potential negative consequences of robot development (threat condition). After watching the video, participants were instructed on their role as teacher (power condition) or assistant (compliant condition) in a human–robot emotion recognition task. Participants were told that they would complete three practice and ten actual trials.

In each trial, participants were first shown the emotional face stimulus (see Fig. 2). Participants in the power condition had to decide from five options which emotion was displayed (happiness, sadness, frustration, anger, fear). Participants in the compliance condition were simply informed of the "correct" emotion and reminded that they ought to comply regardless of their own opinion. On the following page, an animated virtual Nao robot was presented, which stated its own guess at the emotion via audio. Participants provided feedback on the robot's answer by adjusting a slider that - they had been told - controlled the energy supply; an allocation of 100 (or the rightmost position) indicated positive feedback and gave the robot full energy, an allocation of 0 (leftmost position) indicated the most negative feedback and severely restricted the energy supply of the robot. The participants could adjust the slider until they were satisfied with their feedback, and then confirm (see Fig. 3). On the following page, the virtual robot would respond to its feedback. When it had provided a wrong answer, it would lower its head and say something like "Oh no, that's a shame", or "Ah, silly me!". Upon a correct answer it would respond in an elated way. Moreover, to stress the effects of energy restriction the robot's lights would dim and its voice would become more slurred as its energy got restricted more by the participant, with with speech speed decreasing with 5% for every 20 points below 100. This decrease was large enough to be noticeable, but low enough to keep the message intelligible and not dredge the sentence on. When the robot was done talking, participants could proceed to the next trial.

At the end of the thirteen trials, participants were informed that the learning task was over and were presented with three questionnaires: mind attribution, power perception, and threat perception. Finally, participants were thanked for their time, given the debriefing, and provided with a key code for collecting their reimbursement on MTurk. The entire experiment took on average 15 minutes.

Embodied Robot Condition

Upon arrival in the laboratory, participants were seated behind a table with the robot, a "feedback box" through which they could change the robot's energy allocation, an envelope labelled *Face cards* which contained 13 cards with the emotional face stimuli (Fig. 4), and a tablet with instructions that would walk them through the experiment. For participants in the power condition, a second tablet was placed on the table, on which they could select the correct answer on each trial of the face recognition task. See Fig. 5 for the experimental setup.

The experimenter handed the participant a folder containing the information sheet, informed consent form, and their participant number, verbally gave a short overview of the experiment, and then left the room. The robot was programmed to complete the experiment autonomously, displaying idling behaviour (looking around) when not engaged with the learning task. The information sheet gave a more detailed description of the experiment, and the (main) tablet took the participants through experimental procedure step by step.

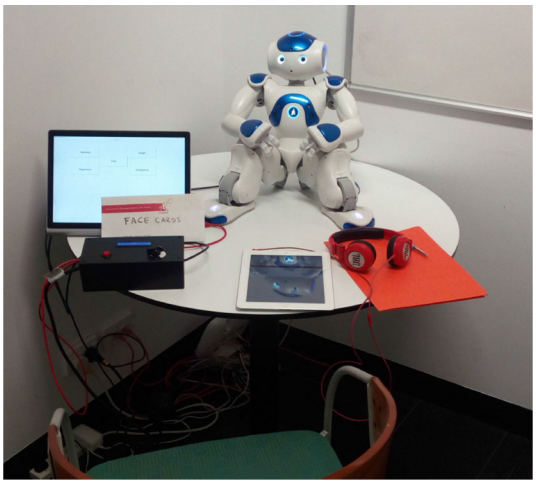

Fig. 4 Experimental setup for the embodied robot/power condition. Left to right: the tablet on which the participant could pick the correct answer in each trial (only for the power condition); envelope with the emotional face stimuli; the feedback box; the Nao robot; the tablet with instructions, movie and questionnaires; the folder with the information sheet and informed consent

Fig. 5 One side of the emotional face stimuli cards for the embodied robot (compliance condition) with a NaoMark at the top and bottom. The other side, which was to be shown to the robot, contained the same image but no text

Participants watched the video which either warned against robots (threat condition), or did not (control condition). Then, the tablet showed the instructions for the emotion recognition interaction task. Participants were instructed to look at the top Face card privately, and either indicate on the second tablet what emotion was depicted (power condition), or to read the emotions label (compliance condition). They then had to show the card to the robot (compliance condition: without showing the emotion label; see Fig. 4). The robot would state its guess at the emotion displayed, after which the participant provided feedback through the feedback box. This was a black box with a dial that could be turned to adjust the energy allocation; a display that showed the energy allocation; and a red button that could be pressed to confirm (see Fig. 5). Upon receiving an updated energy allocation, the robot would respond in either an elated (if correct) or sad (if incorrect) way, with speech being more slurred and its lights more dim for lower energy allocations, and then resume its idling behaviour until it detected a new face card. The first three trials were considered practice trials. After finishing the emotion recognition task, participants completed three questionnaires on the (main) tablet. The entire experiment took about 20 minutes.

2.4 Materials

Emotional Face Stimuli

The emotional face stimuli were selected from a Google Image search for "emotional scene" and "movie emotional face". Face selection was based on showing an intense and ambiguous emotional expression, and the selected images were cropped so that only the face itself was showing (see Fig. 4). The number of occasions and the specific stimuli to which the robot would provide the wrong answer was predetermined and kept constant between participants and conditions.

Virtual Robot Condition

The learning task website was designed in Twine, an open-source application for creating interactive stories. The animated virtual robot was recorded from the Choregraphe simulation window [2] and edited [1].

Robot Voice

The robot's voice for both the embodied and the virtual condition was generated by the text-to-speech function (voice: 'Junior') in the text editor software [3]. The resulting voice was slightly nasal and child-like, albeit clearly not fully human.

Embodied Robot Behaviour

The embodied robot was a Nao V5 (Softbank), programmed in Python. When the code was run, the robot would display idling behaviour (i.e., looking around) until a NaoMark was detected. NaoMarks are landmarks that have been developed by Softbank and can be recognised by Nao robots. They look like black circles with white triangle fans (Fig. 4); the location and width of the triangle fans is used to distinguish one NaoMark from others. As the emotional face stimuli cards

Table 1 Mean scores (*SD*) per condition of all questionnaires

	Virtual robot		Embodied robot	
	Power	Compliance	Power	Compliance
	Mind attribution scale (MAS) (centered)			
Threat reminder	−.03 (*.20*)	−.03 (*.24*)	.02 (*.19*)	.03 (*.21*)
Control	.01 (*.18*)	−.03 (*.21*)	.08 (*.18*)	.03 (*.15*)
	Power questionnaire			
Threat reminder	.88 (*.13*)	.72 (*.19*)	.82 (*.14*)	.75 (*.16*)
Control	.90 (*.14*)	.78 (*.15*)	.84 (*.17*)	.83 (*.15*)
	Threat questionnaire			
Threat reminder	.53 (*.22*)	.49 (*.19*)	.46 (*.16*)	.51 (*.15*)
Control	.53 (*.22*)	.46 (*.22*)	.51 (*.15*)	.47 (*.15*)

each had their own unique NaoMark on them, the robot could identify the exact card that was being shown to it when it detected a NaoMark.

As soon as the robot detected a NaoMark, it would stop its idling behaviour and state its answer (e.g. "I think it's... anger!"). In the compliance condition, these answers were predefined for each NaoMark, thus ensuring that the robot got the same faces wrong each time the experiment was run. In the power condition this same result had to be achieved in a different way, as the "correct" or "wrong" answer depended on the participants opinion. Therefore, in the power condition, the tablet on which participants indicated their decision communicated this answer to the robot's code. Upon detecting a NaoMark, the robot would then give a different answer if it was supposed to get that specific face wrong, and the same answer if it was supposed to be correct.

The speed of the robot's movement and speech while giving its answer was dependent on how much or how little energy the participant had allocated before. Thus, in both the power and the compliance condition, the robot's code received input from the feedback box (see Fig. 5), which was used to slow down or speed up the robot's movement and speech.

If a NaoMark was detected twice, the robot would say it had already seen that face; if a new NaoMark was detected before input from the feedback box had been received, the robot would say that it still needed feedback on the previous answer.

2.5 Measurements

The main measurement was the punishment score, i.e. to what extent the participant restricted the robot's energy supply. Withholding some form of reward in the form of money or points has been used before in HRI [55], and even the specific case of restricting the energy supply has been applied before in the HRI context [7]. Since one could argue that a robot has no use for money, the energy restriction method was adopted as it seemed a more legitimate punishment for robots.

The robot's perceived capabilities of thinking (example item: "I feel like the robot was capable of engaging in thought") and feeling (example item: "I feel like the robot was capable of experiencing emotion") were measured with the ten-item Mind Attribution Scale (MAS; [35]). How powerful the participants felt was measured with a four-item scale [24], which was slightly adapted to fit the task at hand (example item: "To what extent were you in a position of power over the robot?"). Participants' feelings of threat from robots in general were measured with a ten item scale that was adopted from Zlotowski et al. [69] (example item: "Widespread adoption of robots in everyday life troubles me because it is blurring the boundaries between what is human and what is machine").

For the online experiment, all items were measured on an 11-point Likert scale. Two attention checks were added to detect any participants who were not reading the questions carefully. Because the 11-point Likert scale did not format well on the tablet, the participants in the embodied robot condition reported on a 7-point Likert scale. This did not result in different responses between the virtual and embodied robot conditions on any of the questionnaires (see 3.2 for the test statistics and Table 1 for the descriptives).

3 Results

3.1 Homogeneity of Variance

Bartlett's test was used to assess homogeneity of variance between the conditions. The test returned significant ($K^2(7) = 23.68, p = .001$), indicating that the variances were not equal between the embodiment conditions (see Table 2). Thus, a heteroscedasticity consistent variance covariance matrix is

Table 2 Mean punishment scores (*SD*) per condition

	Virtual robot	
	Power	Compliance
Threat reminder	47.41 (*20.88*)	37.50 (*26.33*)
Control	38.73 (*22.22*)	43.45 (*27.67*)
	Embodied robot	
Threat reminder	50.32 (*16.40*)	59.12 (*16.98*)
Control	51.88 (*13.66*)	54.29 (*13.10*)

NB Lower scores indicate harsher punishment (i.e. less energy allocated)

used for the parameters in the model [68] and a Wald test is used for the main analyses.

3.2 Preliminary Analyses

Before analysis, all items in the questionnaires were re-scaled by dividing them by the total range of their scale, resulting in a set of scores between 0 and 1. The MAS was centered, so that positive scores reflect a higher-than-average score and negative scores reflect a lower-than-average score.

The dependent variable (punishment score) was operationalised as participants' average energy allocation over all trials where the robot had provided a wrong answer. The lower the punishment score, the harsher a participant had punished the robot. See Table 2.

Reliability

The reliability of the three questionnaires was assessed with Cronbach's alpha [16]. The Mind Attribution Scale (MAS) and perceived threat measure had a good internal consistency given an alpha of .90 and .89, respectively; the power scale had an acceptable reliability given an alpha of .71. Thus, all questionnaires were considered reliable.

Randomisation Check

To assess randomisation between conditions, differences in mean age and gender ratio were tested. The embodied and virtual robot condition did not differ in male to female ratio, $\chi^2(1, N = 198) = 1.08, p = .30$. Participants were significantly older in the virtual robot condition ($M = 40.34$, $SD = 11.03$) than in the embodied robot condition ($M = 27.82$, $SD = 6.90$), $t(159) = 9.61$, $p < .001$. Gender, age, or an interaction term were not related to punishment of the robot in the virtual robot condition, Fs(1, 138) < .27, ps > .61, suggesting that a difference in age between the two embodiment conditions would not influence the main analysis outcomes.

Manipulation Checks

Two manipulation checks were ran: one for the power condition and one for the threat condition. The manipulation of these conditions was checked by means of ANOVAs with questionnaire score as the dependent variable and the conditions as independent variables.

Participants in the power condition reported feeling more powerful ($M = .87$, $SD = .14$) than the participants in the compliance condition ($M = .77$, $SD = .16$), $F(1, 209) = 12.35$, $p < .001$; no other significant effects were present. Power was thus successfully manipulated.

Perceived threat did not differ between conditions, Fs(1, 209) < 2.53, ps > .11. This result indicated that the threat manipulation either had not worked, or that its effect was too subtle to be picked up by the questionnaire. See Table 1 for the means and standard deviations of all questionnaires.

Because perceived threat was not successfully manipulated, any significant differences in punishment behaviour between the threat and control condition cannot be ascribed to participants' feelings of threat. Thus, from this point on this manipulation will be referred to as "threat reminder".

In addition to the manipulation checks, two non-manipulation checks were ran: for mind attribution and for payment in the virtual robot condition. As expected, mind attribution was not manipulated by power, threat, or embodiment Fs(1, 209) < .90, ps > .34 (see also Table 1) and was thus entered into the multiple regression model as a covariate rather than an experimental factor. In the virtual robot condition, payment was unrelated to either the dependent variable (i.e. punishment, $F(2, 140) = .71$, $p = .49$) or one of the questionnaire variables (i.e. perceived threat, perceived power, mind attribution; Fs(2, 140) < .75, ps >.48).

3.3 Main Analyses

To test whether power, threat reminder, mind attribution (MAS), and embodiment, influenced punishment (i.e. mean energy allocation to the robot after a wrong answer) in the way that was predicted in 1.2, two multiple linear regression models were fitted and compared. The first model contained a four-way interaction between all the predictors, that is embodiment, power, threat reminder, and the centered MAS scores . The second model left out all the nonsignificant effects from the first. By comparing both models to the null model, we tested whether they predicted punishment significantly better than chance. By comparing the two models against one another, we tested whether either of them was superior to the other. Comparisons were done by means of Wald tests.

Both models were better at predicting punishment than the null model, $F(15, 216) = 3.06$, $p < .001$, and $F(11, 216) = 3.62$, $p < 001$, respectively. The difference between the first

and the second model was not significant, $F(4, 201) = .05$, $p = .72$, indicating that they predicted punishment equally well. Occam's razor was applied and the second model, being the simpler of the two, was selected as the one that predicted punishment behaviour best.

The second model revealed a significant main effect and a number of interaction effects, which make interpretation complicated. Thus, in addition to reporting the coefficients, a model interpretation will be given below.

MAS was a significant predictor of punishment: $b = 39.82$, $p = .05$. Furthermore, there were two significant two-way interactions: between power and threat reminder ($b = 16.50$, $p = .05$), and MAS and threat reminder ($b = -67.44$, $p = .01$). A two-way interaction between MAS and power was marginally significant, $b = -52.86$, $p = .051$; as was a two-way interaction between embodiment and threat reminder $b = 14.49$, $p = .08$. Finally, there were two three-way interactions: between power, threat reminder, and MAS, $b = 80.64$, $p = .02$; and between power, threat reminder and embodiment, $b = -25.69$, $p = .02$.

3.4 Model Interpretation

It is important to note that although embodiment, power, and threat reminder were experimental factors and thus can be assumed to have caused the effect on punishment, mind attribution was measured and not manipulated. As a result, a causal relationship between mind attribution and punishment cannot be inferred. Moreover, although participants that saw the extended video in the threat reminder condition behaved differently from the participants that did saw the control video, the failure of the manipulation check indicated that it would be wrong to assume that feelings of threat caused this change in behaviour.

The fitted values of energy allocation for the virtual (left) and embodied (right) robot are plotted for each experimental condition in Fig. 6. Please note that a higher (fitted) energy allocation corresponds to a less severe punishment.

As can be seen in Fig. 6, people tended to allocate more energy after a mistake (i.e., were less harsh in their punishments) to an embodied robot than a virtual one.

The interactions between mind attribution and the different manipulations can also be seem in the variance bars of the predicted energy allocations in Fig. 6. When people felt powerful and had been reminded of threat, mind attribution was not related to energy allocation. The short or nonexistent variance bars for the power conditions indicate that when feeling powerful, how much mind people attributed to the robot did not relate to punishment. When people had been assigned the compliant role however, how capable they thought the robot to be of thinking and feeling was related to their energy allocation. When looking at the coefficient estimates in the model, it becomes clear that although mind

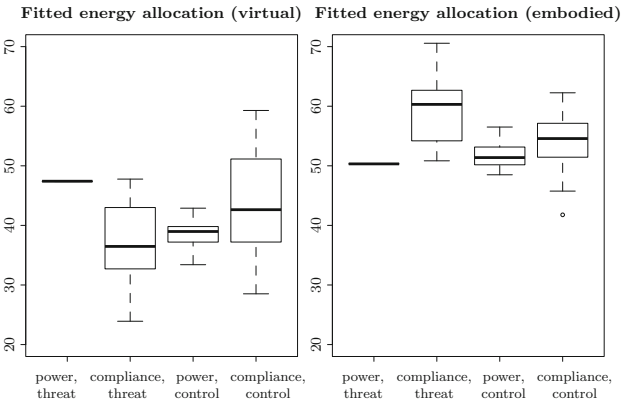

Fig. 6 Predicted energy allocations (lower scores indicating more restriction, i.e. harsher punishment) from the second model for the virtual (left) and embodied (right) robot, separated per experimental condition. A larger variance indicates a larger influence of mind attribution on the predicted scores

attribution and power allocation were positively related in the control condition ($b = 39.82$, i.e. the more a participant thinks the robot is capable of thinking or feeling, the kinder they get), this effect reverses when people had seen the threat reminder video ($b = (39.82 - 67.44 =) -27.62$, i.e., the more a participant thought the robot would be capable of thinking and feeling, the more they restricted its energy supply after a wrong answer). In other words, the relationship between mind attribution and energy restriction flipped as people were exposed to the threat reminder video.

Another thing that can be observed from Fig. 6 and the model coefficients is that embodiment changed the influence of threat reminder as well as the interaction between the threat reminder and power condition. Seeing the threat reminder video increased energy allocation compared to the control condition, but only for the embodied robot. Seeing the threat reminder video while feeling powerful *in*creased the energy allocation for the virtual robot with 16.50 points, but *de*creased it for the embodied robot with $(16.50 - 25.69 =) 9.19$ points.

4 Discussion

Although the HRI literature has noted the issue of robot abuse [e.g. [6,12]] and the need for suitable interventions [10,54,63] so far there has been little research on the psychological motivations for this behaviour [but see [5,34]]. The current experiment looked at the influence of power, threat, embodiment, and mind attribution on robot punishment. We found this relationship to be rather complicated, and will discuss the implications below. Further research is needed if the field of HRI wants to better understand what drives human–

robot aggression, and to develop appropriate interventions in response.

The four psychological factors had been selected based on the literature on human–human aggression [29,31,44] as well as a previous experiment that introduced dehumanisation to HRI research [34]. This previous work had meant to study the influence of mind attribution and anthropomorphism on robot bullying through manipulating power and the robot's human-likeness. However, priming power did not influence mind attribution even though it did have an effect on how abusive people got towards the robot. This suggested that while the connection between dehumanisation and aggression holds true for robots as well as humans, factors that trigger dehumanisation in inter-human interaction (i.e., power) do not generalise to HRI. In the current experiment we thus studied aggression towards robots in relation to mind attribution and power, but also added in robot embodiment and feelings of threat in order to test whether the previous results could be replicated and generalised to embodied robots.

4.1 Predictions and Findings

Only a part of our predictions as stated in 1.2 were confirmed. More or less in line with expectations, people were kinder to an embodied robot than to a virtual one (see Fig. 6). However, this was not the predicted main effect and interactions were found between embodiment and other manipulations. Notable is the interaction between robot embodiment, power prime, and threat reminder, which formed an unexpected exception to the tendency of participants to be more mild in their punishments of an embodied robot. Equally puzzling is that also the compliance/threat condition interacted with embodiment, so that it went from the condition with the lowest amount of allocated energy to the highest. Why did robot embodiment influence the effect of threat?

One explanation could be the that one group got in physical contact with a robot, while the other group had to deal with a virtual (socially distant) robot. Previous studies have found that physical contact improved peoples opinions of a stereotyped entity [20,31]. Wullenkord et al. [65] found that actual contact with a robot reduced negative attitudes, and increased positive attitudes compared to imagined contact or none at all. In the threat reminder condition, negative stereotypes of robots were triggered; and then half of that group had to interact with a socially distant virtual robot while the other half got to interact an embodied one. People dealing with a socially distant virtual robot may have held on to the negative attitude, while participants that were introduced to the embodied robot softened their negative responses as a result of the interaction.

For the participants in the compliance condition embodiment could thus have had an effect on how strongly negative people felt towards the robot, and subsequently influence how harshly they punished it. People who feel in power however tend to be more prone to rely on stereotypes [28]. Thus, participants in the power condition may not have been swayed much based by the embodiment of the robot. Another prediction that was partially confirmed was the relationship between mind attribution and punishment. Mind attribution was related to less harsh punishments when people had not been reminded of robot threat, and when people were primed with power this effect disappeared. In contrast to predictions however, when participants had been reminded of threat this relation between mind attribution and punishment reversed. In previous studies on inter-human interactions, higher perceived threat has been associated with lower mind attribution [31]. However, stereotypical robot threat depends strongly on AI becoming more intelligent, while human threat seems to be more complicated - high intelligence on its own is not sufficient. It thus makes intuitive sense that when feeling threatened, higher mind attribution to a robot is related to more aggression (after all, the smarter the robot, the more capable it is of overthrowing you). Still, it is an interesting contrast with how threat, mind attribution and human aggression are related.

4.2 Strengths and Limitations

The current work has made a modest but nonetheless much needed addition to the body of experimental work on the psychology of robot abuse. The use of theories from inter-human aggression to address not the direct problem (i.e. "how to stop aggression towards robots") but instead study the question that lies below (i.e. "what makes people more, or less, aggressive towards robots") is new to the field of HRI. Moreover, the experimental setup allows to draw causal inferences on most of the factors that were studied in the current work.

Some limitations have to be noted as well. Contrary to our prediction, the manipulation check revealed that the threat manipulation failed; yet at the same time the manipulation still had an effect on behaviour. Possibly, another construct rather than threat was manipulated. For example, the mention of famous persons such as Elon Musk in the last 20 seconds of the video for the threat condition may have activated concepts such as authority, or scientific and creative thinking. However, since the 20 seconds of extra material consisted of a list of concerns with only a sideways reference to two celebrities, it seems improbable that concepts related to the celebrities were triggered but the explicitly mentioned 'robot threat' was not. Moreover, if indeed celebrity-related concepts had been primed, the question remains as to why that would influence how harshly participants punished their robot.

An alternative explanation for the failed manipulation check is that the movie was too abstract in the threat it posed, or the questionnaire too coarse to capture the effect of the manipulation. The method of using a movie as manipula-

tion as well as the threat questionnaire had been used before [66,69] to confirm successful manipulation of threat. However, in these studies the movie had been more explicit in showing threat: participants saw videos with robots directly outperforming humans [66] and being able to reject human commands [69]. In contrast, the current study had only a reminder of the concerns around robots in general, not the Nao robot that was used, and one could argue that the potential threats mentioned (i.e. robots taking over the work force and AI becoming uncontrollable) do not apply to Nao. The questionnaire on the other hand is quite explicit in its statements (e.g. "The increased prevalence of robots in everyday life is threatening to human safety", "In the long run, robots pose a direct threat to human safety and well-being"), and may thus have fitted a more explicit and specific version of threat manipulation better. A more thorough replication and re-examination of the effects of threat is needed. More in general, future work would be advised to pilot test manipulations even if they appear to be straightforward.

Secondly, the sample size for the embodied robot was smaller compared to the virtual condition. This was due to web-based experiments being easier to run, which makes large sample sizes feasible, whereas lab based experiments are labour-intensive and to a much greater extent restricted by the availability of resources like funding, time, and the pool of potential participants. However, especially in the light of an interaction between embodiment and the other independent variables, a larger sample size for the embodied robot condition would have been fitting.

A third, minor, limitation is the lack of initial demographic assessment in the embodied robot condition, which made it impossible to check for successful full randomisation of gender and age. Also, whether the age difference between the embodiment conditions influenced the results cannot be assessed with certainty. Although age was unrelated to punishment for the virtual robot, data did not allow to test this for the embodied robot.

Moreover, although the robots had the same design and displayed the same behaviour, the addition of the furniture in the virtual robot's room may have suggested the robot to be human-sized (or the furniture to be child-sized) whereas the embodied robot stands a little under 60 cm (2 feet) tall. This may have lead to a difference in size perception of the robots, which in turn could have influenced the sense of power that people had over the robot. The robot design and behaviour however stayed the same, meaning that the robot in both embodiment conditions maintained its child-like appearance and high-pitched voice. These would both indicate a kid sized rather than an adult sized robot. In addition, the manipulation checks showed no evidence for an effect of embodiment on perceived power. Participants in the power condition felt more powerful over the robot, but the embodiment condition nor the threat condition had an effect on perceived power and there were no interaction effects. It thus seems unlikely that any difference in perceived size, if present at all, biased the results.

Finally, it should be noted that the face cards used were not pilot tested for either image quality (colour hue, saturation and contrast) or perceived emotional ambiguity. The first issue would increase variance in a similar way across conditions and would thus at least affect the results evenly across participants, resulting in a reduction of power to detect significant results. The second issue, however, may have biased results in the Power manipulation, with participants who decided for themselves what the correct emotion was being more convinced of the "correct" answer than participants who received instructions on this. Note that the faces had been selected to be ambiguous rather than clear in their emotional expression. Anecdotal evidence, in the form of participants in both conditions complaining afterwards that the emotions on the face cards were not obvious and that they could see how the robot's guess was potentially applicable as well, suggests that this is not the case but in future studies a more testable form of control would be preferred.

Unfortunately, drawing a direct line between the aggression observed in this experiment and robot abuse is not possible. Abuse is a complex behaviour and has to do with not only aggression but also involves power imbalance, the intention to humiliate or hurt, and a lack of (sufficient) provocation for the aggressive behaviour [see for example [45,49]]. Therefore, while we strongly believe that the current findings are relevant for the topic of robot abuse, the study cannot be directly generalised to abuse.

4.3 Future Work

The current study has some implications for future research. The relevance of mind attribution to robot-directed aggression has been replicated, thus suggesting a promising direction for the research on robot abuse discouraging strategies. However, also replicating previous findings, mind attribution was not manipulated by factors that are effective manipulators in inter-human interaction. A further exploration of if, and how, mind attribution to robots can be manipulated will be necessary before it can be used in aggression deterring strategies.

Related to this, current results imply that the relationship between mind attribution and aggression may depend on perceived threat. However, with the failing manipulation check this cannot be said with certainty. Seen how the popular media tends to dramatise social robots (with robots at the very least running amok, and at the very worst first taking over the workforce and then the world), perceived threat would arguably be an important factor to study in relation to robot abuse.

Predicting aggression towards robots appears to be at least as hard as predicting aggression towards humans. If anything,

this only strengthens our call for more theoretical research on the psychological factors influencing human–robot aggression. Before long, other robots will join the K5 Knightscope in public areas. If these are to survive, they will have to be able to deal with humanity's less pretty behaviours as well.

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

Compliance with ethical standards

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

References

1. Adobe Systems Software (2017) Adobe After Effects CC for MacOS (14.2.1) [Computer software]
2. Aldebaran Robotics, SoftBank Group (2014) Choregraphe for MacOS (2.1.4) [Computer software]
3. Apple Inc (1995-2016) TextEdit (Version 1.12 (329)) [Computer software], voice "Junior"
4. Bain P, Park J, Kwok C, Haslam N (2009) Attributing human uniqueness and human nature to cultural groups: distinct forms of subtle dehumanization. Group Process Intergroup Relat 12(6):789–805. https://doi.org/10.1177/1368430209340415
5. Bartneck C, Hu J (2008) Exploring the abuse of robots. Interact Stud 9(3):415–433. https://doi.org/10.1075/is.9.3.04bar
6. Bartneck C, Rosalia C, Menges R, Deckers I (2005) Robot abuse—a limitation of the media equation. In: Proceedings of the Interact 2005 Workshop on agent abuse, designed intelligence, Rome, Italy. https://doi.org/10.17605/OSF.IO/4FXQ6
7. 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 international conference on human–robot interaction, ACM/IEEE, Arlington, USA, pp 217–222. https://doi.org/10.1145/1228716.1228746
8. Bartneck C, Reichenbach J, Carpenter J (2008) The carrot and the stick—the role of praise and punishment in human–robot interaction. Interact Stud Soc Behav Commun Biol Artif Syst 9(2):179–203. https://doi.org/10.1075/is.9.2.03bar
9. 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
10. Brahnam S, De Angeli A (2008) Special issue on the abuse and misuse of social agents. Interact Comput 20:287–291. https://doi.org/10.1016/j.intcom.2008.02.001
11. 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
12. Brscić D, Kidokoro H, Suehiro Y, Kanda T (2015) Escaping from children's abuse of social robots. In: Proceedings of the International Conference on Human-Robot Interaction, ACM/IEEE, Portland, USA, pp 59–66. https://doi.org/10.1145/2696454.2696468
13. 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
14. Cañamero LD (2002) Playing the emotion game with Feelix. In: Socially intelligent agents, Springer, pp 69–76. https://doi.org/10.1007/0-306-47373-9_8
15. Croizet JC, Claire T (1998) Extending the concept of stereotype threat to social class: the intellectual underperformance of students from low socioeconomic backgrounds. Personal Soc Psychol Bull 24(6):588–594. https://doi.org/10.1177/0146167298246003
16. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. https://doi.org/10.1007/BF02310555
17. 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
18. De Angeli A (2006) On verbal abuse towards chatterbots. In: Proceedings of CHI 2006 Workshop on Misuse and Abuse of Interactive Technologies. Montreal, Canada. https://doi.org/10.1016/j.intcom.2008.02.004
19. De Angeli A, Brahnam S (2008) I hate you! Disinhibition with virtual partners. Interact Comput 20(3):302–310. https://doi.org/10.1016/j.intcom.2008.02.004
20. Ensari N, Miller N (2002) The out-group must not be so bad after all. The effects of disclosure, typicality, and salience on intergroup bias. J Personal Soc Psychol 83(2):313. https://doi.org/10.1037/0022-3514.83.2.313
21. 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
22. Field A (2009) Discovering statistics using SPSS, 3rd edn. Sage Publications, Thousand Oaks
23. Fischer P, Greitemeyer T, Kastenmüller A, Frey D, Oßwald S (2007) Terror salience and punishment. Does terror salience induce threat to social order? J Exp Soc Psychol 43(6):964–971
24. Galinsky AD, Gruenfeld DH, Magee JC (2003) From power to action. J Personal Soc Psychol 85(3):453–466. https://doi.org/10.1037/0022-3514.85.3.453
25. Galinsky AD, Magee JC, Inesi ME, Gruenfeld DH (2006) Power and perspectives not taken. Psychol Sci 17(12):1068–1074. https://doi.org/10.1111/j.1467-9280.2006.01824.x
26. 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
27. Gockley R, Bruce A, Forlizzi J, Michalowski M, Mundell A, Rosenthal S, Sellner B, Simmons R, Snipes K, Schultz AC, et al. (2005) Designing robots for long-term social interaction. In: International conference on intelligent robots and systems, IEEE/RSJ, New York, USA, pp 1338–1343. https://doi.org/10.1109/IROS.2005.1545303
28. Goodwin SA, Gubin A, Fiske ST, Yzerbyt VY (2000) Power can bias impression processes. Stereotyping subordinates by default and by design. Group Process Intergroup Relat 3(3):227–256
29. Gwinn JD, Judd CM, Park B (2013) Less power= less human? Effects of power differentials on dehumanization. J Exp Soc Psychol 49(3):464–470. https://doi.org/10.1016/j.jesp.2013.01.005
30. Haslam N (2006) Dehumanization. an integrative review. Personal Soc Psychol Rev 10(3):252–264. https://doi.org/10.1207/s15327957pspr1003_4
31. Haslam N, Loughnan S (2014) Dehumanization and infrahumanization. Ann Rev Psychol 65:399–423. https://doi.org/10.1146/annurev-psych-010213-115045
32. 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

33. Hayashi K, Sakamoto D, Kanda T, Shiomi M, Koizumi S, Ishiguro H, Ogasawara T, Hagita N (2007) Humanoid robots as a passive-social medium-a field experiment at a train station. In: Proceedings of the international conference on human–robot interaction, New York, USA, pp 137–144, https://doi.org/10.1145/1228716.1228735
34. 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
35. Kozak MN, Marsh AA, Wegner DM (2006) What do I think you're doing? Action identification and mind attribution. J Personal Soc Psychol 90(4):543–555. https://doi.org/10.1037/0022-3514.90.4.543
36. 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
37. Kteily N, Bruneau E, Waytz A, Cotterill S (2015) The ascent of man: theoretical and empirical evidence for blatant dehumanization. J Personal Soc Psychol 109(5):901. https://doi.org/10.1037/pspp0000048
38. Lammers J, Stapel DA (2011) Power increases dehumanization. Group Process Intergroup Relat 14(1):113–126. https://doi.org/10.1177/1368430210370042
39. Lapidot-Lefler N, Barak A (2012) Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Comput Hum Behav 28(2):434–443. https://doi.org/10.1016/j.chb.2011.10.014
40. Lasica JD (2014) Knightscope K5 at the Launch Festival, held Feb. 24-26, 2014 at San Francisco's Design Concourse. https://commons.wikimedia.org/wiki/File:Knightscope_K5_(12809731473).jpg. (Online; recovered 25 June 2019)
41. Leidner B, Castano E, Ginges J (2013) Dehumanization, retributive and restorative justice, and aggressive versus diplomatic intergroup conflict resolution strategies. Personal Soc Psychol Bull 39(2):181–192. https://doi.org/10.1177/0146167212472208
42. 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
43. Locke KD (2009) Aggression, narcissism, self-esteem, and the attribution of desirable and humanizing traits to self versus others. J Res Personal 43(1):99–102. https://doi.org/10.1016/j.jrp.2008.10.003
44. Lowry PB, Zhang J, Wang C, Siponen M (2016) Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. Inf Syst Res 27(4):962–986. https://doi.org/10.1287/isre.2016.0671
45. 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
46. Mosbergen D (2015) Good job, America. You killed hitchBOT. Huffpost https://www.huffpost.com/entry/hitchbot-destroyed-philadelphia_n_55bf24cde4b0b23e3ce32a67
47. 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
48. Nomura T, Kanda T, Kidokoro H, Suehiro Y, Yamada S (2017) Why do children abuse robots? Interact Stud 17(3):347–369. https://doi.org/10.1145/2701973.2701977
49. Postigo S, González R, Montoya I, Ordoñez A (2013) Theoretical proposals in bullying research. A review. Anal de Psicol 29(2):413–425
50. Rosenthal-von der Pütten AM, Krämer NC, Hoffmann L, Sobieraj S, Eimler SC (2013a) An experimental study on emotional reactions towards a robot. Int J Soc Robot 5(1):17–34. https://doi.org/10.1007/s12369-012-0173-8
51. Rosenthal-von der Pütten AM, Schulte FP, Eimler SC, Hoffmann L, Sobieraj S, Maderwald S, Krämer NC, Brand M (2013b) 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
52. Rehm M, Krogsager A (2013) Negative affect in human robot interaction – impoliteness in unexpected encounters with robots. In: 2013 Proceedings of IEEE RO-MAN, IEEE, pp 45–50. https://doi.org/10.1109/ROMAN.2013.6628529
53. Rudman LA, Mescher K (2012) Of animals and objects: Men's implicit dehumanization of women and likelihood of sexual aggression. Personal Soc Psychol Bull 38(6):734–746. https://doi.org/10.1177/0146167212436401
54. 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: 19th international symposium in robot and human interactive communication, RO-MAN, 2010 IEEE, IEEE, Viareggio, Italy, pp 1–7. https://doi.org/10.1109/ROMAN.2010.5654677
55. Sandoval EB, Brandstetter J, Bartneck C (2016) Can a robot bribe a human? The measurement of the dark side of reciprocity in human robot interaction. In: 11th ACM/IEEE international conference on human–robot interaction, IEEE, Christchurch, pp 117 – 124. https://doi.org/10.1109/HRI.2016.7451742
56. 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
57. Slater M, Antley A, Davison A, Swapp D, Guger C, Barker C, Pistrang N, Sanchez-Vives MV (2006) A virtual reprise of the Stanley Milgram obedience experiments. PlOS One 1(1):e39. https://doi.org/10.1371/journal.pone.0000039
58. Suler J (2004) The online disinhibition effect. Cyberpsychol Behav 7(3):321–326. https://doi.org/10.1089/1094931041291295
59. 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
60. Vincent J (2017) A drunk man was arrested for knocking over Silicon Valley's crime-fighting robot. https://www.theverge.com/2017/4/26/15432280/security-robot-knocked-over-drunk-man-knightscope-k5-mountain-view. (Online; recovered 30 August 2018)
61. Volk AA, Veenstra R, Espelage DL (2017) So you want to study bullying? Recommendations to enhance the validity, transparency, and compatibility of bullying research. Aggress Viol Behav 36:34–43. https://doi.org/10.1016/j.avb.2017.07.003
62. Waytz A, Epley N (2012) Social connection enables dehumanization. J Exp Soc Psychol 48(1):70–76. https://doi.org/10.1016/j.jesp.2011.07.012
63. Whitby B (2008) Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. Interact Comput 20(3):326–333. https://doi.org/10.1016/j.intcom.2008.02.002
64. Woods J (2011) Framing terror: An experimental framing effects study of the perceived threat of terrorism. Crit Stud Terror 4(2):199–217. https://doi.org/10.1080/17539153.2011.586205

65. Wullenkord R, Fraune MR, Eyssel F, Šabanović S (2016) Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE, New York, USA, pp 980–985. https://doi.org/10.1109/ROMAN.2016.7745228
66. Yogeeswaran K, Złotowski J, Livingstone M, Bartneck C, Sumioka H, Ishiguro H (2016) The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. J Hum Robot Interact 5(2):29–47. https://doi.org/10.5898/JHRI.5.2.Yogeeswaran
67. Young M (2016) What is a robot? https://www.youtube.com/watch?v=S5miA6jXf0E&frags=pl%2Cwn
68. Zeileis A (2004) Econometric computing with HC and HAC covariance matrix estimators. Research Report Series / Department of Statistics and Mathematics. https://doi.org/10.18637/jss.v011.i10
69. Zlotowski J, Yogeeswaran K, Bartneck C (2017) Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. Int J Hum Comput Stud 100:48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008