[1] Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied.
Proceedings of the ACM/IEEE International Conference on Human Robot Interaction, Chicago pp. 205-214. | DOI: 10.1145/3171221.3171266

Mindless Robots get Bullied

Merel Keijsers, Christoph Bartneck
University of Canterbury, HIT Lab NZ
Private Bag 4800
Christchurch, Canterbury 8140
merel.keijsers@pg.canterbury.ac.nz

ABSTRACT

Humans recognise and respond to robots as social agents, to such extent that they occasionally attempt to bully a robot. The current paper investigates whether aggressive behaviour directed towards robots is influenced by the same social processes that guide human bullying behaviour. More specifically, it measured the effects of dehumanisation primes and anthropomorphic qualities of the robot on participants' verbal abuse of a virtual robotic agents. Contrary to previous findings in human-human interaction, priming participants with power did not result in less mind attribution. However, evidence for dehumanisation was still found, as the less mind participants attributed to the robot, the more aggressive responses they gave. In the main study this effect was moderated by the manipulations of power and robot anthropomorphism; the low anthropomorphic robot in the power prime condition endured significantly less abuse, and mind attribution remained a significant predictor for verbal aggression in all conditions save the low anthropomorphic robot with no prime. It is concluded that dehumanisation occurs in human-robot interaction and that like in human-human interaction, it is linked to aggressive behaviour. Moreover, it is argued that this dehumanisation is different from anthropomorphism as well as human-human dehumanisation, since anthropomorphism itself did not predict aggressive behaviour and dehumanisation of robots was not influenced by primes that have been established in human-human dehumanisation research.

CCS CONCEPTS

•Human-centered computing →Empirical studies in HCI;

KEYWORDS

robot bullying; dehumanization; anthropomorphism; mind attribution; verbal aggression

ACM Reference format:
Merel Keijsers, Christoph Bartneck. 2018. Mindless Robots get Bullied. In *Proceedings of HRI'18, March 5th–8, 2018, Chicago, IL, USA*, , 10 pages.
DOI: http://dx.doi.org/10.1145/3171221.3171266

1 INTRODUCTION

Once upon a time, a small cleaning robot was assigned an unsupervised job at a public square while scientists were conducting a field experiment with a larger and more sophisticated robot nearby. However, things did not go quite as planned, as the cleaning robot was approached by random bystanders and abused - in the absence of any form of provocation from the robot. The baffled scientists had to hastily reach for their cell phones to capture this unexpected form of human-robot interaction. Salvini et al. [53] noted that *"the nature of the abuses suffered by the robots [...] is much more similar to bullying behaviours than vandalism. [...] In the case of urban robots, acts of vandalism could be, for instance, crashing the touch screen monitor, setting fire to the robot, or keying the robot cover. On the contrary, what we noticed during the behavioural study were actions aimed at forcing the robot to do or not to something or, in a few cases, simulations of "physical" attacks"* [53, p. 371].

Humans recognise robots as social actors. They talk to them [4] as if they understand what is being said, they punish them when they prove to be a bad teammate [3] but also feel sorry for them when they are being punished [55], and even try to prevent them from getting hurt [9, 55]. Our brain responds to robots as if they were giving off social cues; activating mirror neurons when watching a robot perform an action [18], activating neural networks linked to the theory of mind when playing a game with a robot [30], and activating areas associated with emotional empathy when watching a robot getting hurt [50]. Many studies report on how people engage in social interaction with robots [26].

However, not all social behaviours are positive. The cleaning robot from the Salvini et al. [53] study is not alone in being abused; Brscić et al. [5] observed how kids attacked a robot that was patrolling a shopping mall. After multiple failed attempts to design robot behaviour that could stop these aggressive tendencies, the authors had to program the robot in such a way that it simply avoided potential bullies (i.e., any kid-sized human). Of course, robots are the ideal target for bullying as they are in a clear subordinate position, will not retort in kind, and cannot feel any pain, which absolves the aggressor from any moral consequence [11].

This is not to say that robot bullying should be tolerated. From an ethical perspective, some behaviours can be deemed immoral even if performed on a entity that is incapable of any suffering, like a robot [57]. Since the robot is recognised by the human as a social actor, abusing it might encourage treating other humanlike beings (e.g. actual humans) in a similar way [65]. More generally speaking, the assertion "I can do whatever I desire with a robot" rests upon the idea that all and any actions are acceptable as long as no-one gets harmed [48], which even in the most libertarian societies is

HRI'18, March 5th–8, 2018, Chicago, IL, USA
© 2018 ACM. ISBN 978-1-4503-4953-6/18/03…$15.00
DOI: http://dx.doi.org/10.1145/3171221.3171266

not a commonly shared attitude [65]. And from a pragmatic point of view, robot abuse can result in considerable damage to the robot and hazardous situations for both the robot and bystanders [11].

Research on the reasons behind robot bullying is still sparse [10] and often involves anecdotal observations [53]. The current paper will thus dive into the psychological motivations behind robot bullying behaviour. More specifically, it will manipulate how humanlike a robot is perceived, using the theoretical frameworks of dehumanisation and anthropomorphism, and measure the effects of this manipulation on bullying behaviour. The goal is to experimentally show that aggressive behaviour towards robots is a social phenomenon, and is guided by the same social processes as aggressive behaviour towards humans.

1.1 Perceiving (non)human

In 1994, Nass et al. [44] found that humans treat computers as if they are social actors; a finding that inspired his "Media Equation" theory, which states that humans will automatically respond to media as if it is real life [46]. In a series of experiments, it was shown that for well-established human interaction mechanisms, the interaction partner can be substituted with a computer without changing the behavioural outcomes. Later studies further confirmed that people interact with machines and media as if they are social agents [40].

The Media Equation extends to robots as well. When playing a cooperative game with either a robotic or a human partner, participants apply the same social norms to both partners, punishing bad performance and rewarding good performance [3]. With increasing humanlikeness in a robot, activation of the brain regions that are associated with the theory-of-mind neural network is enhanced [30]. Moreover, seeing a robot hand carrying out a series of movements activates the same mirror neurons in the brain as observing a human hand [18], further suggesting that the brain processes robots (to some degree) as if they are human agents.

The phenomenon of seeing human characteristics in nonhuman agents is called anthropomorphism (from the Greek words "anthropos", meaning "human", and "morphe", meaning "form"): the first mention of the word stems from the sixth century BC [34]. Epley et al. [13] drew a motivational framework for anthropomorphism that rests on three factors. Two are human needs: the need to understand and predict the behaviour of other agents and the need to be social. The third factor is related to the agent and considers how much it resembles a human in appearance and behaviour.

Waytz et al. [64] confirmed the validity of the first factor in the human-robot interaction field by showing that robots behaving in an unpredictable way have higher anthropomorphism ratings. Moreover, when participants are dealing with unpredictable robots, a brain region that is involved with inferring mental states of other agents becomes activated, suggesting that the robots are indeed recognised as having a mind of their own.

The second factor was confirmed when loneliness was shown to correlate with the tendency to assign a mind to interactive gadgets (like an alarm clock that "runs away" when it goes off). Moreover, after a feeling of loneliness was experimentally induced, participants had a greater tendency to anthropomorphise a wide range of nonhuman agents: pets, God, and a series of ambiguous drawings in which one might perceive the features of a face [12].

Finally, the validity of the agent factor was experimentally confirmed by Eyssel et al. [14], who showed that the relationship between psychological closeness to and the anthropomorphism of robots depends on the robots voice. When the robot had a synthesised voice, the gender of the robots voice didn't influence anthropomorphism ratings in participants; but when it had a human voice, robots with a voice that matched the participants' gender were rated as more anthropomorphous.

The Media Equation can however not be equated to anthropomorphism completely [4]. Although humans apply social norms when interacting with a robot, they also display certain behaviours that would be unacceptable in human-anthropomorphous animal behaviour; for example, switching off the robot whenever they become bored with it. So while robots are recognised as social agents, their perceived humanlikeness is somehow different from the humanlikeness of anthropomorphous animals. This suggests that anthropomorphism is not a uni-dimensional construct, but involves a more complex mental process.

A related theoretical framework called mind perception provides a possible explanation. Mind perception theory states that the perception of a mind in other agents, human or not, is guided by two dimensions: Experience and Agency [19]. The first dimension entails to what extent an agent is thought to be capable of experiencing thoughts, feelings, and the world around it. Agents who are high in Experience and low on Agency are perceived as being less responsible for their actions, more prone to feeling hurt, and thus deserving of protection. The second dimension indicates to what extent the agent is seen as capable of self control, memory, planning, and moral judgement. Beings high in Agency but low on Experience are considered to be responsible for their own actions and less deserving of protection from harm, as they don't have (a rich set of) feelings and supposedly can take care of themselves.

In the survey of Gray et al. [19], robots scored high on Agency but low on Experience, whereas animals scored high on Experience but low on Agency (this study however stems from 2006, before the field of social robotics took off. Robots nowadays may be seen as more capable of Experience). This would explain why robots and animals are not anthropomorphous in the same way: since robots are rated high on Agency instead of Experience, one can harm them without feeling bad [19]. Indeed, increasing Agency traits in a robot did not do much for its anthropomorphism or likeability ratings, but increasing its Experience traits resulted in it being perceived as more anthropomorphic and likeable [52, 67].

So in general, humans perceive robots as humanlike mostly to the extent that they are independent and autonomous. But if the settings are right, the perceived capability of a robot to fully experience the world around it can be tweaked as well. Together with the notion that bullying is a social behaviour (albeit a negative one), this suggests that in order to explain robot bullying behaviour, one should look at which psychological mechanisms come into play when humans get mean towards each other.

1.2 Humanness and aggression

The mechanisms underlying human-on-human aggression are explained in dehumanisation theory. Perceiving another as less human allows people to disregard the (negative) consequences of their behaviour and thus decreases empathy towards the victim [7, 36]. Haslam [21] defines two sets of characteristics that determine humanness: Human Nature (HN) and Uniquely Human (UH) traits [see also 23]. UH traits and capabilities are presumably reserved for humans only, like higher forms of cognition. HN on the other hand involves traits and capabilities that are shared with other animals, but are at the same time considered fundamental to being human, like fear or joy. Perceiving less HN traits in an agent results in a 'mechanistic' form of dehumanisation; in humans, this form of dehumanisation is applied to for example bankers and businessmen. Alternatively, perceiving less UH traits results in an 'animalistic' form of dehumanisation [21, 23]; this form of dehumanisation is commonly applied to women or the mentally disabled.

Dehumanisation can be triggered by stable factors like trait characteristics of the person who dehumanises (e.g. narcissism and conservatism) and of the victim (e.g. social class, gender); but also circumstantial factors like emotional state, a sense of power, or self-focus [22]. Interestingly, although agents high on HN traits are seen as more deserving of protection, animalistic dehumanisation is still related to a decrease in empathy [8] and both types of dehumanisation are related to increased aggression [22, 33].

Dehumanisation theory shows considerable overlap with anthropomorphism, to the point where it has been suggested that they are two approaches of the same concept [63]. Applying this claim to research, Loughnan et al. [38] applied a dehumanisation framework on non-human agents and showed that UH traits were more readily associated with robots, while HN traits were more easily linked to animals. However, not all scholars agree that dehumanisation and anthropomorphism are each others reverse. For example, in one study neither robot appearance nor perceived intentionality influenced the mind or moral agency attributed to it. A more human appearance of the robot resulted in an increase in ascribed UH and HN traits; but at the same time less perceived intentionality correlated with a *higher* attribution of UH traits [68].

The inconsistency in findings might be explained by the many different approaches that have been used to measure dehumanisation and anthropomorphism. Anthropomorphism has been operationalised in many different ways by different researchers [27, 51]: from the humanness subscale of the revised Godspeed questionnaire [24], which quite straightforwardly asks the participant to rate the robots on scales like 'living versus inanimate' and 'human-made versus humanlike' [41, 67]; to home-crafted questionnaires [15, 49, 50]; to mind attribution questionnaires [12, 14, 64] and UH and HN attribution measurements [37, 52], which hold the (implicit) assumption of dehumanisation and anthropomorphism being each others opposite. This makes it virtually impossible to tease apart anthropomorphism, dehumanisation, and mind attribution in the literature.

1.3 Current studies

The current studies follow a 2 (dehumanisation tendencies) x 2 (anthropomorphism of the robot) between participant design. Participants are either primed to dehumanise or receive a control task, and then engage in a scripted dialogue with a virtual NAO robot (see Figure 1) which is either high or low in anthropomorphism. The main dependent variable is the proportion of negative or aggressive responses compared to the number of positive interactions.

We hypothesise that participants who received a dehumanisation prime will be be more aggressive to the robot than the control group. Moreover, we expect that this effect is stronger for a high anthropomorphic robot compared to a low anthropomorphic robot.

To manipulate dehumanisation, a feeling of power is primed in participants, as this manipulates dehumanisation tendencies [20] but not anthropomorphism [28]. Both the use of a human voice [14] and the inclusion of social cues through movement [43] have been shown to increase perceived anthropomorphism of a robot. Two questionnaires are administered at the end of the experiment as manipulation checks.

The online setting was partly chosen because it provides easy access to an enormous pool of potential participants, but also because it reduces inhibition and self-consciousness in participants through the "online disinhibition effect" [58]. Although on- and offline bullying do not differ in principle, as reported by both perpetrators and victims [42], people are more likely to bully online than offline [39]. This is due to the invisibility and anonymity of the aggressor and victim, and the lack of bystanders who might intervene [31, 58], among other things. These factors lower the threshold for interhuman aggression [62] as well as aggression towards a virtual robot [10]. It is thus assumed here that using an online platform may enhance the effect but will not alter the nature of the factors that moderate bullying tendencies towards robots.

Similarly, interaction with a virtual robot is not fundamentally different from interaction with an embodied one. Previous studies have shown that virtual representations of robots elicit more social behaviour (like mimicking expressions, empathy, polite behaviour, and physiological responses) than audiotapes or text [50, 55], indicating that virtual robots too are recognised as social agents. Li [35] conducted a meta-analysis on papers that studied the influence of agent embodiment on users' perception of the agent, and concluded that embodied robots elicit stronger behavioural and attitudinal responses than virtual agents. However, several studies which had found no difference in behavioural and attitudinal responses for

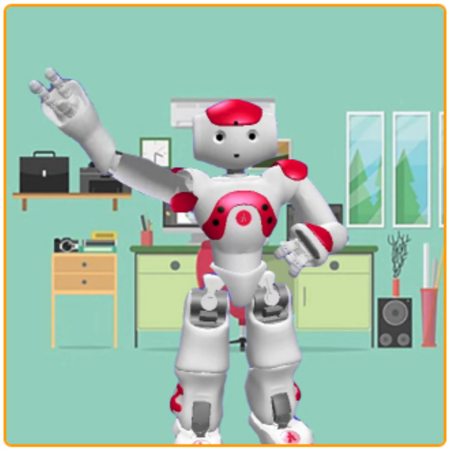

Figure 1: The robot in the opening scene of the experiment

virtual agents and physical robots were missing n this analysis [for example 45, 47]. More recent studies also found that the perception of and response to virtual agents is identical to embodied robots [59, 66]. Thellman et al. [59] found that it is social presence (i.e. whether the robot is perceived as a social actor that manifests humanness [32]) rather than physical presence that predicts the social influence of a robot. Moreover, social presence was not influenced by the physical embodiment of the robot in their experiment.

While the literature is still on the fence on to what extent virtual and embodied robots are interchangeable, we argue that the underlying psychological mechanisms are the same (but the intensity of the experience may or may not differ). Thus, while our experiment features a virtual robot in an online setting, we feel confident that the gist of the findings can be applied to embodied robots too.

2 PILOT STUDY

2.1 Methods

2.1.1 Participants. Participants were approached on a number of platforms, but mainly signed up via online crowdsourcing companies CrowdFlower and Amazon Mechanical Turk. Data from these platforms has been shown to be of equal quality as on-campus recruitment or participant data from forums [2, 54]. For the current study, compliant with the common reimbursement rates on those websites, participants were paid $1.25 USD for completing the study. In addition to CrowdFlower and Amazon, the experiment was also distributed through the university Facebook page and the forum r/SampleSize on the online platform Reddit. Participants who signed up via these platforms did not get reimbursed.

232 participants completed the interaction and questionnaires. 17 participants clearly did not comply with the essay guidelines (i.e. did not write on the provided topic or copy-pasted their essay off of the internet) and were removed from the dataset. Thirty participants failed the attention question and were removed as well. The resulting dataset thus held 185 participants. 39% of them were male; the average age was 38 years (*SD* = 11.5); the majority listed the USA as their country of residence (66%).

2.1.2 Procedure. Participants were told that the study was a pilot for evaluating a virtual robot agent that would introduce the lab robots to children. After providing demographic information, they were asked to write a 200 word essay on either what they would do if they were president with unrestricted power for a day (power prime, dehumanisation condition) or the last time they visited a mall (control condition). This part was framed as a check of their proficiency in English.

After submitting their essay, participants were reminded to turn on the sound on their device and keep it on during the whole interaction. They were then shown a virtual environment with the robot, which introduced itself to them as one of the robots in the HITlab of the University of Canterbury. The robot either had a humanlike voice and gave off social cues through movement (high anthropomorphism condition) or spoke with a synthesised voice and was shown in stills (low anthropomorphism condition).

Participants engaged in a scripted interaction with the robot, where they could respond to the robot by selecting either of two or three responses presented to them on the screen. Sometimes, all options were neutral, but in roughly 90% of the cases one response was positive and another negative or abusive in nature. If a positive answer option was given, there always was a negative response option as well, and vice versa.

After the participant had selected their response, the robot would give its reaction. These reactions differed depending on the selected response. To ensure that the participants fully understood what the robot had said, a transcript appeared on the screen once it was done talking. Participants could also refresh the page in order to re-listen to what the robot had to say. The whole interaction took 10-15 minutes.

After the interaction, the participants were presented with a new screen, which informed them that the interaction part was now over, and were asked to rate the virtual agent on two questionnaires (the anthropomorphism and dehumanisation manipulation check measurements). When the participants had given their opinion on the final item, they were debriefed and asked to submit their answers to the database.

2.1.3 Materials. The scripted interaction was designed in Twine, an open-source application for creating interactive nonlinear stories. Due to the nonlinear story line, there were many possible interaction paths. The response that the participant got from the robot depended on the answer option they had selected. For example, if the robot said "[I]t's rather dark in the storage room where they put us [the robots]. [...] So in spite of not being alone, it can get boring", the participant could choose between "I am sorry to hear that" and "This is stupid. You are a robot, you can't feel". Upon picking the first response, the robot would react in a friendly way, assuring the participant it wasn't all that bad. Alternatively, if the participant chose the second option, the robot would respond in a sad and insecure manner, and change the topic.

The robot voice in the low anthropomorphism condition was generated by the text-to-speech function in the text editor software [25]. The robot voice in the high anthropomorphism condition was recorded from a native English speaking student. The robots' movements in the high anthropomorphism condition were recorded from the Choregraphe simulation window [1] and edited to change the background with Adobe After Effects [56].

2.1.4 Measurements.

Aggression measurement. The proportion of negative responses was used as a measurement of aggression, and used as dependent variable in the binomial models that were defined. Only responses where both a negative and a positive response option had been presented were taken into account.

Manipulation checks. A manipulation check was included for both dehumanisation and anthropomorphism. For the dehumanisation manipulation check, the mind attribution scale (MAS) from Kozak et al. [29] was used. In this questionnaire participants rate to what extent the robot is capable of experiencing each of ten mental capabilities (e.g. "capability of experiencing complex feelings", "capability of engaging in planned action"). Since being capable of feelings and thoughts is central to being human [38], dehumanisation would show in less attributed mind to the robot.

For the anthropomorphism manipulation check, the humanlikeness subscale of the revised Godspeed questionnaire (GQ_r) [24] was used. In this questionnaire, participants rate a robot on six bipolar scales, e.g. "synthetic - real", "living - inanimate", and "without definite lifespan - mortal". In both questionnaires, items were measured on an 11-point Likert scale.

2.2 Results

2.2.1 Reliability, randomisation and manipulation check. The reliability of both questionnaires was assessed by calculating Cronbach's alpha. The GQ$_r$ had an alpha of .83; the MAS had an alpha of .90. Thus, both questionnaires were considered reliable. To make interpretation easier, the full MAS was reverse-scored so that a higher score indicated a lower degree of mind attribution and thus a higher degree of dehumanisation.

The four conditions did not differ significantly from each other in participants' mean age, gender, or country of residence; or with respect to the total number of interactions per participant. The groups did not differ significantly in sample size, $\chi^2(3, N = 185) = 4.25$, $p = .24$, with 40 participants in the low anthropomorphism/control condition, 58 in the high anthropomorphism/control condition, 45 in the low anthropomorphism/dehumanisation condition, and 42 in the high anthropomorphism/dehumanisation condition.

Participants in the high anthropomorphism condition rated their robot as significantly more anthropomorphous ($M = 5.76$, $SD = 2.10$) than participants in the low anthropomorphism condition ($M = 4.90$, $SD = 2.17$), $F(1,181) = 6.25$, $p = .01$, no main effect for the dehumanisation condition or interaction term, ps > .35. Participants in the dehumanisation condition did not attribute significantly less mind to their robot ($M = 5.93$, $SD = 2.01$) compared to the control condition ($M = 5.78$, $SD = 2.27$), $F(1,181) = .14$, $p = .70$, no significant main effect for the anthropomorphism condition or interaction term, ps > .33. Thus, the manipulation of anthropomorphism had been successful, but the power prime had not led to a greater degree of dehumanisation of the robot. As it did not manipulate dehumanisation tendencies, the dehumanisation condition will from this point on be referred to as the power prime condition. The MAS scores will be used as an indication of dehumanisation instead. See Table 2 for the mean score on both questionnaires.

On average, 75% of participants' interaction paths overlapped ($SD = .08\%$).

2.2.2 Main analysis. Four binomial regression models are proposed and compared below. For all models, the dependent variable was the proportion of negative responses. The predictor variables were a composition of either or both conditions and the score on the MAS. To make interpretation easier, the scores on the MAS had been centered beforehand. Chi-square statistics are used to assess if a proposed model is better at predicting aggressive responses than the null model (which holds no predictors). The Akaike information criterion (AIC) is used to compare the models amongst each other, with a lower AIC score indicating a better fit compared to the alternative model and a difference (Δ_{AIC}) of less than 2 points indicating that the models are roughly equivalent [6].

The first model that was put up for comparison followed the original analysis plan and had as predictors the two conditions and an interaction term. This model had no significant predictors, all $-.85 < z < .54$, all ps > .40, and it thus did not do any better than the null model at predicting aggressive responses, $\chi^2(3, N = 185) = 2.36$, $p = .50$; AIC = 867.1.

Table 1: Descriptives of the models predicting the aggression ratio in the pilot study

	Predictors	b	z	AIC
Model 1	(intercept)	−1.64	-13.93^{***}	
	ant	0.08	0.54	
	power	−0.14	−0.85	
	ant×power	0.00	0.02	867.1
Model 2	(intercept)	−1.9	-13.78^{***}	
	ant	0.22	1.17	
	power	−0.01	−0.07	
	MAS	0.41	7.88^{***}	
	ant×power	−0.05	−0.99	
	ant×MAS	−0.10	$-1.78^{\dagger}$	
	power×MAS	−0.16	−0.71	711.61
Model 4	(intercept)	−1.82	-29.24^{***}	
	MAS	0.33	12.04^{***}	708.27

†, *, **, and *** denote significance at $p < .10$, $p < .05$, $p < .01$, and $p < .001$, respectively (two-tailed).

In the second model, the MAS score and power prime condition were added as predictor variables and twice each as an interaction variable. The MAS score was the only significant predictor, $b = .41$, $z = 7.88$, $p < .001$, although an interaction between MAS and anthropomorphism approached significance, $b = -.10$, $z = -1.78$, $p = .08$. The model was a significant improvement over the null model, $\chi^2(6, N = 185) = 163.85$, $p < .001$; the AIC indicated a preference for the second model over the first, AIC = 711.61, $\Delta_{AIC} = 155.49$.

In the third model, the power prime condition was removed from the model, leaving the MAS score and the anthropomorphism condition as main effects and interaction. In this model as well, only the MAS score predicted aggression; $b = .38$, $z = 8.82$, $p < .001$, although an interaction between MAS and anthropomorphism once more approached significance, $b = -.09$, $z = -1.70$, $p = .09$. This model predicted aggressive responses significantly better than the null model, $\chi^2(3, N = 185) = 160.35$, $p < .001$; the AIC difference indicated a slight preference for the third model over the second, AIC = 709.11, $\Delta_{AIC} = 2.5$.

Thus, a final model was defined containing only the MAS score as a predictor; $b = .33$, $z = 12.04$, $p < .001$. This model as well was significant, $\chi^2(1, N = 185) = 157.19$, $p < .001$; the AIC indicated it to be preferable over the second, and roughly equivalent to the third model, AIC = 708.27, $\Delta_{AIC} = 3.34$ and $\Delta_{AIC} = .84$, respectively.

Since model 3 and 4 fit the data equally well, there is no statistical incentive to prefer one over the other. However, as the MAS score was the only significant predictor in model 3, Occam's razor is applied and model 4 is identified as the model that predicts aggressive responses best. See Table 1 for the statistics of model 1, 2 and 4 (since model 3 and 4 were very similar in their outcomes, model 3 is not included).

2.3 Discussion

Two main findings emerged in the pilot study. Firstly, the results indicate that mind attribution is predictive of robot bullying. The

[1] Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied.
Proceedings of the ACM/IEEE International Conference on Human Robot Interaction, Chicago pp. 205-214. | DOI: 10.1145/3171221.3171266

Table 2: Questionnaire descriptives per condition for both studies

		GQ_r (*SD*)			MAS (*SD*)		
		low anthrop.	high anthrop.	total	low anthrop.	high anthrop.	total
Pilot study	Control	4.97 (2.21)	6.06 (2.16)	5.62 (2.24)	5.92 (2.10)	5.69 (2.38)	5.78 (2.27)
	Power prime	4.83 (2.15)	5.33 (1.96)	5.07 (2.06)	5.74 (2.00)	6.14 (2.03)	5.93 (2.01)
	total	4.90 (2.17)	5.76 (2.10)	5.36 (2.17)	5.83 (2.04)	5.93 (2.24)	5.85 (2.15)
Main study	Control	4.99 (2.03)	6.65 (1.54)	5.77 (1.98)	5.27 (2.27)	5.20 (1.30)	5.24 (1.86)
	Power prime	5.52 (2.23)	6.59 (2.25)	6.07 (2.29)	5.80 (2.23)	5.31 (1.97)	5.55 (2.10)
	total	5.27 (2.13)	6.61 (1.96)	5.94 (2.15)	5.55 (2.25)	5.27 (1.70)	5.41 (1.99)

less mind is attributed to a robot, the more aggressive responses it will get. Whether the robot was moving and talking with a human voice, or still and speaking with a computer generated voice did not have any significant influence on aggression. These findings have a few implications.

They suggest that the relation between anthropomorphism and dehumanisation is more complicated than “two sides of the same coin”. One can manipulate “aliveness” of a robot without affecting the mind that is attributed to it, or the bullying that it will suffer. The interaction between mind attribution and anthropomorphism suggested that the influence of mind attribution might be modified by robot looks in such a way that mind gets less relevant as the robot looks more humanlike, but this interaction did not reach statistical significance. These findings are in line with the findings from Zlotowski et al. [68], who found that robot appearance does not affect mind attribution, and the assertion of Thellman et al. [59] that the social presence of a robot, not its embodiment, is the main factor in shaping affective and behavioural reactions.

Moreover, although power priming as a dehumanisation manipulation failed, the results indicate that human-robot aggression is related to the same psychological processes that guide human-human aggression. Perceiving the robot as less capable of thinking and feeling *increases* the number of attempts to hurt a robot, instead of taking away the incentive for bullying. Due to the study setup, a causal direction unfortunately cannot be inferred; less mind attribution may lead to more aggression, or people may perceive a robot as being less mindful in order to justify their aggressive responses.

Table 3: Mean aggression ratio (*SD*) for both studies

		low anthrop.	high anthrop.	total
Pilot	Control	.31 (.25)	.35(.23)	.31 (.24)
	Power prime	.31 (.23)	.34 (.19)	.32 (.21)
	total	.31 (.24)	.34 (.21)	.33 (.22)
Main	Control	.36 (.26)	.34 (.21)	.35 (.24)
	Power prime	.29 (.25)	.29 (.25)	.29 (.25)
	total	.32 (.26)	.31 (.24)	.32 (.25)

NB: *ratio* is the number of negative to positive responses; data has been square root transformed to facilitate interpretation.

The second main finding was the failure of the power prime. This could be taken as evidence that humans do not dehumanise robots in the same way they dehumanise humans; while aggression is related to (a lack of) mind perception, factors that influence mind perception in fellow humans do not influence robot mind perception. Alternatively, the manipulation method could have been biased. While power priming was copied from previous studies, where it had been an effective method [17, 20], the topic had not been adopted verbatim. Given that the majority of the respondents lived in the US, together with the recent developments in the oval office, “president for a day” seemed to have triggered more than just feelings of power. For example, some participants used the essay mainly to express their unhappiness with the current POTUS.

Thus, in the main experiment the design was kept identical to the pilot except for the prime. We adopted an essay topic that had been previously described [16] and established [17, 20] to manipulate dehumanisation.

3 MAIN STUDY

Except for the power prime, this study's design was identical to the pilot.

3.1 Methods

3.1.1 Participants. Only Amazon Mechanical Turk was used as a recruitment platform for the main study, as participants on this platform are reimbursed only after their submitted data has been approved, which allowed the researchers to discard participants who failed the attention check. 129 participants completed the essay and questionnaires. Of those, 12 submitted an essay that was either off-topic or had been copy-pasted from the internet and were removed, resulting in a dataset with 117 participants. 49% were male, the average age was 38 years (*SD* = 11.0), and the majority (80%) resided in the USA.

3.1.2 Procedure, materials, and measurements. The procedure was identical to the pilot study, save for the essays. In the dehumanisation condition participants now had to recall and describe in detail a personal incident in which they had power over another individual or individuals [16, 17, 20]. The visit to the shopping mall in the control condition was changed to a visit to a grocery store, as some participants in the pilot study had remarked that they hadn't been to a mall in years.

The materials and measurements were identical to those used in the pilot study.

3.2 Results

3.2.1 Reliability, randomisation and manipulation check. Both questionnaires were reliable (Cronbach's alpha = .86 for the MAS; .90 for the GQ_r). The full MAS again was reverse-scored, so that a higher score indicated a higher degree of dehumanisation.

The four conditions did not differ significantly from each other with respect to the participants' country of residence, gender or the total number of interactions. The groups did not differ significantly in sample size, with 28 participants in the low anthropomorphism/control condition, 25 in the high anthropomorphism/control condition, 31 in the low anthropomorphism/dehumanisation condition, and 33 in the high anthropomorphism/dehumanisation condition, $\chi^2(3, N = 117) = 1.26$, $p = .74$.

Participants' mean age differed significantly between the groups, $F(1,115) = 12.24$, $p < .001$. Since age is correlated to the aggression ratio (ρ = -.15), it was included in the models as a control variable.

Participants in the high anthropomorphism condition rated their robot as significantly more anthropomorphous ($M = 6.62$, $SD = 1.96$) than participants in the low anthropomorphism condition ($M = 5.27$, $SD = 2.13$), $F(1,113) = 8.50$, $p < .01$, no significant main effect for the dehumanisation condition or the interaction term, *p*s > .33. Participants in the dehumanisation condition did not attribute significantly less mind to their robot ($M = 5.55$, $SD = 2.10$) compared to the control condition ($M = 5.24$, $SD = 1.86$), $F(1,113) = 1.01$, $p = .32$, no main effect for the anthropomorphism condition or the interaction term, *p*s > .58. Thus, the manipulation of anthropomorphism had been successful, but the manipulation of dehumanisation had not. The MAS was once more used as a measure of dehumanisation, and again the dehumanisation condition will be referred to as the "power prime condition" from this point on. See Table 2 for descriptives of both questionnaires.

On average, 74% of participants interaction paths overlapped (*SD* = .09%).

3.2.2 Main analysis. As in the pilot study, a series of binomial models were composed and compared. The dependent variable was the proportion of negative responses. The predictors were a subset of either or both experimental conditions and the scores on the MAS and the GQ_r, with age as a control variable. The scores on both questionnaires were centered in order to facilitate interpretation of the models. Chi-square statistics were calculated to assess if a proposed model was better at predicting aggressive responses than the null model (which holds no predictors). The Akaike information criterion (AIC) was used to compare the models amongst each other, with a lower AIC score indicating a better fit compared to the alternative model, and a difference (Δ_{AIC}) of 2 points or less indicating the models are approximately equal [6].

The first model contained the two experimental conditions, an interaction term, and the control variable. In this model only age was a significant predictor, $b = -.02$, $z = -2.30$, $p = .02$. This model still outperformed the null model on predicting the number of aggressive responses,$\chi^2(4, N = 117) = 10.13$, $p = .04$; AIC = 594.46.

Table 4: Descriptives of the models predicting the aggression ratio in the main study

	Predictors	b	z	AIC
Model 1	(intercept)	−0.82	-3.09^{**}	
	ant	−0.28	−1.45	
	power	−0.29	−1.53	
	ant×power	0.37	1.34	
	age	−0.02	-2.30^{*}	594.46
Model 2	(intercept)	−0.72	-2.60^{**}	
	ant	−0.27	−1.33	
	power	−0.68	-3.10^{**}	
	MAS	−0.05	−0.80	
	ant×power	0.75	2.50^{**}	
	ant×MAS	0.41	3.07^{**}	
	power×MAS	0.42	4.92^{***}	
	ant×power×MAS	−0.60	-3.77^{***}	
	age	−0.02	-2.55^{**}	545.27

*, **, and *** denote significance at $p < .05$, $p < .01$, and $p < .001$, respectively (two-tailed).

In the second model, the MAS was added as a main predictor and as a factor in the interaction term. This model returned main effects for power prime and age, $b = -.68$, $z = -3.10$, $p = .002$ and $b = -.02$, $z = -2.55$, $p = .01$, respectively. There were interactions between the two conditions, $b = .75$, $z = 2.50$, $p = .01$, and between either condition and the MAS scores, $b = .41$, $z = 3.07$, $p = .002$ for the interaction with the anthropomorphism condition, and $b = .42$, $z = 4.92$, $p < .001$ for the interaction with the power prime condition; and a three-way interaction between the conditions and the MAS score, $b = -.60$, $z = -3.77$, $p < .001$. The model was significantly better than the null model at predicting aggressive responses; $\chi^2(8, N = 117) = 67.14$, $p < .001$; and its AIC indicated it to be preferable over the first model, AIC = 545.27, $\Delta_{AIC} = 49.19$.

The second model is thus identified as the model that predicts robot bullying best. See Table 4 for the descriptives of both models.

3.2.3 Model interpretation. The chosen model gets easier to interpret when the regression equations are written out for each of the four conditions. See Table 5.

For the low anthropomorphic robot in the control condition, only age predicted aggression; for each additional year of the participants' age, the log odds of an aggressive response decreased with .02.

For the high anthropomorphic robot in the control condition, mind attribution was a significant predictor of aggression as well; for every point that the MAS score was above the mean (i.e. the

Table 5: Regression equations for the four conditions

Condition	Regression equation
Low ant, control	$log(ratio) \sim -.73 - .02 * age$
High ant, control	$log(ratio) \sim -.73 + .41 * MAS - .02 * age$
Low ant, power	$log(ratio) \sim -1.41 + .42 * MAS - .02 * age$
High ant, power	$log(ratio) \sim -.66 + .23 * MAS - .02 * age$

NB: *ratio* is the number of negative to positive responses.

[1] Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied.
Proceedings of the ACM/IEEE International Conference on Human Robot Interaction, Chicago pp. 205-214. | DOI: 10.1145/3171221.3171266

less mind was attributed), the log odds of an aggressive response increased with .41.

All else being equal, the low anthropomorphic robot in the power prime condition had a lower baseline rate of aggressive responses compared to the other conditions. The relationship between age, mind attribution, and aggression was similar to the anthropomorphic robot in the control condition.

Finally, the overall lower rate of abuse of robots in the high anthropomorphism/control condition (Table 3) was not due to a lower baseline rate, but to a less strong effect of mind attribution on aggression.

4 DISCUSSION

As social robotics take up an increasingly prominent place in both science and society, the issue of robot abuse becomes more relevant. While a variety of scholars have observed abuse of both embodied [26, 53] and virtual [11, 61] agents, there is still very little fundamental research on where this behaviour originates from.

The current studies took up a (social) psychology paradigm and investigated the influence of anthropomorphism and dehumanisation on verbal abuse of a virtual robot. The hypotheses - dehumanisation leads to more aggression, an effect that is enhanced by anthropomorphism - were partially confirmed. In the pilot study, while aggression was unaffected by power and anthropomorphism, a lack of mind attribution (an indication of dehumanisation) was directly related to abuse. In the main study, this effect became moderated by feelings of power and anthropomorphism.

Against our expectations, priming participants with power failed to induce dehumanisation tendencies. Although the relationship between dehumanisation and robot bullying still could be studied by using the mind attribution score that was originally intended as manipulation check, the effect of power (null in the pilot and *decreasing* aggression in the main study) does raise some questions.

The most glaring question - did the prime actually manage to induce feelings of power in participants - cannot be tested with the data. However, the prime was adopted because of its solid previous establishment as inducing feelings of power [16, 17, 20], and its influence on behaviour in the current studies (albeit not through dehumanisation) indicates that something indeed was triggered.

If we assume for the moment that this was power, then why did power not influence dehumanisation? How to explain the drop in aggressive tendencies after being power-primed in the low anthropomorphism condition? And why did the power prime decrease the influence of mind attribution on aggression when the robot is anthropomorphic?

A potential explanation is that power worked as an inhibitor. Following this ratio, people bully robots out of uncertainty or perceived threat, as some sort of testing and probing. When they feel powerful, their dominance already feels established, which allows them to be friendlier. Indeed, Zlotowski et al. [68] recently found that the more autonomous a robot appeared to be, the more threatened people felt - a feeling that mediated the relationship between robot autonomy and participants' negative attitudes towards robots. Feelings of power in the current study may have reduced perceived autonomy in robots, or counteracted its moderation of negative attitudes through reducing the experienced threat.

This interpretation of the results provides an intriguing paradigm for future studies on robot abuse. Like the bullying of humans, robots bullying rests upon dehumanisation. But the power imbalance, which is so central in human-human abuse [60], appears to take on a different role in human-robot abuse. Further investigating the interaction between power, perceived threat, dehumanisation, and abuse of robots would lead to deeper understanding of how exactly the human brain processes robots and could be of tremendous value to the field of human-robot interaction.

Also interesting is the independence of anthropomorphism and dehumanisation, concepts that have been labelled each others reverse before [13, 22]. Humanlikeness in a robot did not influence aggression or mind perception, but mind perception on itself was related to robot abuse. These findings match the work of Zlotowski et al. [68], who found that robot appearance did not substantially influence mind attribution; and contradict the proposition that anthropomorphism and dehumanisation are simply two extremes of the same scale [as expressed by for example by 63].

The presented findings, albeit fundamental in nature, have implications for applied robotics as well. Of course, it would be far too early to recommend complicated robot behaviours designs that aim to reduce robot dehumanisation and enhance perceived power of the user; more elaborate studies, with embodied robots, are needed to pinpoint the exact relationship between dehumanisation, power, and robot bullying. However, considering the link between mind perception and aggression, one might consider giving priority to robot qualities that increase its perceived capability of thinking and feeling, over humanlikeness and aliveness.

The first major limitation of the current studies is that it was conducted in a virtual environment. While this has its perks (e.g. the online disinhibition effect), the current literature is undecided on whether this is entirely equatable to embodied robots [35, 45, 47, 59]; the same goes for online versus offline bullying [39, 42]. While we argue that the underlying psychological mechanisms are the same and the results can therefore be generalised, follow-up studies will have to empirically confirm that this is indeed the case.

Robot anthropomorphism was manipulated with minimal measures. On one hand, this allowed for a "cleaner" design, where it could be clear that it was anthropomorphism, and not for example perceived strength or size of the robot, that influenced bullying behaviour. On the downside, the difference in anthropomorphism between the conditions, although significant, is small.

Finally, as mentioned above, the dehumanisation manipulation was checked only by measuring mind attribution, and not feelings of power. However, by adopting the instructions verbatim from successful studies in the main experiment, it seems less likely that a well-established prime suddenly failed to work than that it simply did not influence mind attribution. Nonetheless, in future studies inclusion a measurement of power might be considered as a second manipulation check.

The field of human-robot interaction is very young, but has been around long enough to suggest that understanding the motivation behind robot abuse may prove to be no easier than understanding what drives people to pick on each other. Nonetheless, gaining insights on robot bullying will benefit both our understanding of the human mind as the development of an environment where a small cleaning robot can do its job without fear of being harrassed.

REFERENCES

[1] SoftBank Group Aldebaran Robotics. 2014. Choregraphe for MacOS (2.1.4) [Computer software]. (2014).

[2] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from studies in Amazonfis Mechanical Turk to studies conducted online and with direct recruitment. *PloS one* 10, 4 (2015), e0121595. https://doi.org/10.1371/journal.pone.0121595

[3] Christoph Bartneck, Juliane Reichenbach, and J. Carpenter. 2008. The Carrot and the Stick - The Role of Praise and Punishment in Human-Robot Interaction. *Interaction Studies - Social Behaviour and Communication in Biological and Artificial Systems* 9, 2 (2008), 179–203. https://doi.org/10.1075/is.9.2.03bar

[4] Christoph Bartneck, Michel Van Der Hoek, Omar Mubin, and Abdullah Al Mahmud. 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, 217–222. https://doi.org/10.1145/1228716.1228746

[5] Drazen Brscić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from children's abuse of social robots. In *Proceedings of the tenth annual ACM/IEEE international conference on Human-robot interaction*. ACM, ACM/IEEE, Portland, USA, 59–66.

[6] Kenneth P Burnham and David R Anderson. 2003. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, New York. https://doi.org/10.1198/tech.2003.s147

[7] Emanuele Castano and Miroslaw Kofta. 2009. Dehumanization: Humanity and its denial. *Group Processes & Intergroup Relations* 12, 6 (2009), 695–697. https://doi.org/10.1177/1368430209350265

[8] Emanuele Castano, Miroslaw Kofta, Sabina Čehajić, Rupert Brown, and Roberto González. 2009. What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations* 12, 6 (2009), 715–729. https://doi.org/10.1177/1368430209347727

[9] Kate Darling. 2012. Extending legal rights to social robots. In *We Robot Conference, University of Miami*. University of Miami, Miami, USA, 1–24. https://doi.org/10.2139/ssrn.2044797

[10] Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with computers* 20, 3 (2008), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004

[11] Antonella De Angeli, Sheryl Brahnam, Peter Wallis, and Alan Dix. 2006. Misuse and abuse of interactive technologies. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, Montreal, Canada, 1647–1650. https://doi.org/10.1145/1125451.1125753

[12] Nicholas Epley, Scott Akalis, Adam Waytz, and John T Cacioppo. 2008. Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science* 19, 2 (2008), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

[13] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

[14] Friederike Eyssel, Dieta Kuchenbrandt, Simon Bobinger, Laura de Ruiter, and Frank Hegel. 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, 125–126. https://doi.org/10.1145/2157689.2157717

[15] Friederike A Eyssel and Michaela Pfundmair. 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, 827–832. https://doi.org/10.1109/ROMAN.2015.7333647

[16] Adam D Galinsky, Deborah H Gruenfeld, and Joe C Magee. 2003. From power to action. *Journal of personality and social psychology* 85, 3 (2003), 453–466. https://doi.org/10.1037/0022-3514.85.3.453

[17] Adam D Galinsky, Joe C Magee, M Ena Inesi, and Deborah H Gruenfeld. 2006. Power and perspectives not taken. *Psychological science* 17, 12 (2006), 1068–1074. https://doi.org/10.1111/j.1467-9280.2006.01824.x

[18] Valeria Gazzola, Giacomo Rizzolatti, Bruno Wicker, and Christian Keysers. 2007. The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage* 35, 4 (2007), 1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003

[19] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *Science* 315, 5812 (2007), 619. https://doi.org/10.1126/science.1134475

[20] Jason D Gwinn, Charles M Judd, and Bernadette Park. 2013. Less power= less human? Effects of power differentials on dehumanization. *Journal of Experimental Social Psychology* 49, 3 (2013), 464–470. https://doi.org/10.1016/j.jesp.2013.01.005

[21] Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review* 10, 3 (2006), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

[22] Nick Haslam and Steve Loughnan. 2014. Dehumanization and infrahumanization. *Annual review of psychology* 65 (2014), 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

[23] Nick Haslam, Stephen Loughnan, Yoshihisa Kashima, and Paul Bain. 2008. Attributing and denying humanness to others. *European review of social psychology* 19, 1 (2008), 55–85. https://doi.org/10.1080/10463280801981645

[24] Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518. https://doi.org/10.1016/j.chb.2010.05.015

[25] Apple Inc. 1995-2016. TextEdit (Version 1.12 (329)) [Computer software], voice "Princess". (1995-2016).

[26] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on robotics* 23, 5 (2007), 962–971.

[27] Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. 2015. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology* 6, Article 360 (2015), 14 pages. https://doi.org/10.3389/fpsyg.2015.00390

[28] Sara Kim and Ann L McGill. 2011. Gaming with Mr. Slot or gaming the slot machine? Power, anthropomorphism, and risk perception. *Journal of Consumer Research* 38, 1 (2011), 94–107. https://doi.org/10.1086/658148

[29] Megan N Kozak, Abigail A Marsh, and Daniel M Wegner. 2006. What do I think you're doing? Action identification and mind attribution. *Journal of personality and social psychology* 90, 4 (2006), 543–555. https://doi.org/10.1037/0022-3514.90.4.543

[30] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS One* 3, 7 (2008), e2597. https://doi.org/10.1371/journal.pone.0002597

[31] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior* 28, 2 (2012), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

[32] Kwan Min Lee. 2004. Presence, explicated. *Communication theory* 14, 1 (2004), 27–50. https://doi.org/10.1111/j.1468-2885.2004.tb00302.x

[33] Bernhard Leidner, Emanuele Castano, and Jeremy Ginges. 2013. Dehumanization, retributive and restorative justice, and aggressive versus diplomatic intergroup conflict resolution strategies. *Personality and Social Psychology Bulletin* 39, 2 (2013), 181–192. https://doi.org/10.1177/0146167212472208

[34] James H Lesher et al. 2001. *Xenophanes of Colophon: fragments: a text and translation with a commentary*. Vol. 4. University of Toronto Press, Toronto.

[35] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

[36] Meng Yao Li, Bernhard Leidner, and Emanuelle Castano. 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 (2014), 285–300. https://doi.org/10.4473/TPM21.3.4

[37] Stephen Loughnan and Nick Haslam. 2007. Animals and androids: Implicit associations between social categories and nonhumans. *Psychological Science* 18, 2 (2007), 116–121. https://doi.org/10.1111/j.1467-9280.2007.01858.x

[38] Steve Loughnan, Nick Haslam, Tess Murnane, Jeroen Vaes, Catherine Reynolds, and Caterina Suitner. 2010. Objectification leads to depersonalization: The denial of mind and moral concern to objectified others. *European Journal of Social Psychology* 40, 5 (2010), 709–717. https://doi.org/10.1002/ejsp.755

[39] Paul Benjamin Lowry, Jun Zhang, Chuang Wang, and Mikko Siponen. 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. *Information Systems Research* 27, 4 (2016), 962–986. https://doi.org/10.1287/isre.2016.0671

[40] Holger Luczak, Matthias Roetting, and Ludger Schmidt. 2003. Let's talk: anthropomorphization as means to cope with stress of interacting with technical devices. *Ergonomics* 46, 13-14 (2003), 1361–1374. https://doi.org/10.1080/00140130310001610883

[41] Karl F MacDorman and Debaleena Chattopadhyay. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146 (2016), 190–205. https://doi.org/10.1016/j.cognition.2015.09.019

[42] Kathryn L Modecki, Jeannie Minchin, Allen G Harbaugh, Nancy G Guerra, and Kevin C Runions. 2014. Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health* 55, 5 (2014), 602–611. https://doi.org/10.1016/j.jadohealth.2014.06.007

[43] Lilia Moshkina, Susan Trickett, and J Gregory Trafton. 2014. Social engagement in public places: a tale of one robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM/IEEE, Bielefeld, Germany, 382–389. https://doi.org/10.1145/2559636.2559678

[44] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Boston, USA, 72–78. https://doi.org/10.1145/191666.191703

[45] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on.* ACM/IEEE, Arlington, USA, 145–152. https://doi.org/10.1145/1228716.1228736

[46] Byron Reeves and Clifford Nass. 1996. *The Media Equation.* CSLI Publications and Cambridge University Press, Cambridge.

[47] Juliane Reichenbach, Christoph Bartneck, and Julie Carpenter. 2006. Well done, Robot! The importance of praise and presence in human-robot collaboration. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on.* IEEE, Hatfield, UK, 86–90. https://doi.org/10.1109/ROMAN.2006.314399

[48] Kathleen Richardson. 2016. The asymmetrical'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society* 45, 3 (2016), 290–293. https://doi.org/10.1145/2874239.2874281

[49] Laurel D Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. 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, 245–246. https://doi.org/10.1145/1514095.1514158

[50] Astrid M Rosenthal-von der Pütten, Nicole C Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C Eimler. 2013. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics* 5, 1 (2013), 17–34. https://doi.org/10.1007/s12369-012-0173-8

[51] Peter AM Ruijten, Diane HL Bouten, Dana CJ Rouschop, Jaap Ham, and Cees JH Midden. 2014. Introducing a rasch-type anthropomorphism scale. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.* ACM/IEEE, Bielefeld, Germany, 280–281. https://doi.org/10.1145/2559636.2559825

[52] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323. https://doi.org/10.1007/s12369-013-0196-9

[53] Pericle Salvini, Gaetano Ciaravella, Wonpil Yu, Gabriele Ferri, Alessandro Manzi, Barbara Mazzolai, Cecilia Laschi, Sang-Rok Oh, and Paolo Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *RO-MAN, 2010 IEEE.* IEEE, Viareggio, Italy, 1–7. https://doi.org/10.1109/ROMAN.2010.5654677

[54] Daniel J Simons and Christopher F Chabris. 2012. Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PloS one* 7, 12 (2012), e51876. https://doi.org/10.1371/journal.pone.0051876

[55] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V Sanchez-Vives. 2006. A virtual reprise of the Stanley Milgram obedience experiments. *PloS one* 1, 1 (2006), e39. https://doi.org/10.1371/journal.pone.0000039

[56] Adobe Systems Software. 2017. Adobe After Effects CC for MacOS (14.2.1) [Computer software]. (2017).

[57] Robert Sparrow. 2017. Robots, Rape, and Representation. *International Journal of Social Robotics* 9, 4 (2017), 465–477. https://doi.org/10.1007/s12369-017-0413-z

[58] John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326. https://doi.org/10.1089/1094931041291295

[59] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. Virtual Agent Embodiment and Effects on Social Interaction. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016.* Springer International Publishing, Los Angeles, USA, 412–415. https://doi.org/10.1007/978-3-319-47665-0_44

[60] Anthony A Volk, René Veenstra, and Dorothy L Espelage. 2017. So you want to study bullying? Recommendations to enhance the validity, transparency, and compatibility of bullying research. *Aggression and violent behavior* 36 (2017), 34–43. https://doi.org/10.1016/j.avb.2017.07.003

[61] Peter Wallis. 2005. Robust normative systems: What happens when a normative system fails. In *Abuse: the darker side of human-computer interaction, Interact 2005.* Rome, Italy, 68–72.

[62] Adam Waytz and Nicholas Epley. 2012. Social connection enables dehumanization. *Journal of experimental social psychology* 48, 1 (2012), 70–76. https://doi.org/10.1016/j.jesp.2011.07.012

[63] Adam Waytz, Nicholas Epley, and John T Cacioppo. 2010. Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science* 19, 1 (2010), 58–62. https://doi.org/10.1177/0963721409359302

[64] Adam Waytz, Carey K Morewedge, Nicholas Epley, George Monteleone, Jia-Hong Gao, and John T Cacioppo. 2010. Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of personality and social psychology* 99, 3 (2010), 410–465. https://doi.org/10.1037/a0020240

[65] Blay Whitby. 2008. Sometimes itfis hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers* 20, 3 (2008), 326–333. https://doi.org/10.1016/j.intcom.2008.02.002

[66] Ricarda Wullenkord, Marlena R Fraune, Friederike Eyssel, and Selma Šabanović. 2016. Getting in Touch: How imagined, actual, and physical contact affect evaluations of robots. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on.* IEEE, New York, USA, 980–985. https://doi.org/10.1109/ROMAN.2016.7745228

[67] Jakub Zlotowski, E. Strasser, and C. Bartneck. 2014. Dimensions of anthropomorphism: from humanness to humanlikeness. In *Proceedings of the 9th ACM/IEEE Conference on Human-Robot Interaction (HRI 2014)*, G. Sagerer (Ed.). ACM/IEEE, New York, USA, 66–73.

[68] Jakub Zlotowski, Hidenobu Sumioka, Christoph Bartneck, Shuichi Nishio, and Hiroshi Ishiguro. 2017. Understanding anthropomorphism: Anthropomorphism is not a reverse process of dehumanization. (2017), 6 pages.