International Journal of Psychology

International Journal of Psychology, 2020
DOI: 10.1002/ijop.12715

Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes

Dennis Küster[1] and Aleksandra Swiderska[2]

[1]Department of Computer Science, University of Bremen, Germany
[2]Department of Psychology, University of Warsaw, Poland

According to moral typecasting theory, good- and evil-doers (agents) interact with the recipients of their actions (patients) in a moral dyad. When this dyad is completed, mind attribution towards intentionally harmed liminal minds is enhanced. However, from a dehumanisation view, malevolent actions may instead result in a denial of humanness. To contrast both accounts, a visual vignette experiment ($N = 253$) depicted either malevolent or benevolent intentions towards robotic or human avatars. Additionally, we examined the role of harm-salience by showing patients as either harmed, or still unharmed. The results revealed significantly increased mind attribution towards visibly harmed patients, mediated by perceived pain and expressed empathy. Benevolent and malevolent intentions were evaluated respectively as morally right or wrong, but their impact on the patient was diminished for the robotic avatar. Contrary to dehumanisation predictions, our manipulation of intentions failed to affect mind perception. Nonetheless, benevolent intentions reduced dehumanisation of the patients. Moreover, when pain and empathy were statistically controlled, the effect of intentions on mind perception was mediated by dehumanisation. These findings suggest that perceived intentions might only be indirectly tied to mind perception, and that their role may be better understood when additionally accounting for empathy and dehumanisation.

Keywords: Moral typecasting theory; Dehumanisation; Benevolent intentions; Mind perception; Robots.

Robots have been present in industrial workplaces since the mid-1970s (Gunkel, 2018). Today, they are no longer machines designed to perform a limited set of automated functions but are becoming increasingly visible in everyday life (Dautenhahn, 2007). Some even take on the role of social interaction partners (Duffy, 2003; Jones et al., 2015). As reported by the International Federation for Robotics (International Federation of Robotics (IFR), 2019), worldwide sales of domestic service robots (DSRs), that is, robots used in the household or for entertainment purposes, have grown dramatically in recent years. Following a growth in sales by 15% in 2018, and 27% in 2019, sales of DSRs reached a total value of 4.6 billion USD in 2019, with projections up to 35% annual growth by 2022 (IFR, 2019). Beyond this, independently operating *autonomous robots* have come into the spotlight not only for disaster control (see Nagatani et al., 2013), but also in the context of the recent coronavirus crisis. For example, robots are expected to be soon deployed in hospitals to perform disinfection or distribute medical supplies (IFR, 2020).

Robots as social interaction partners

How will the rising presence of robots in the social sphere shape our interactions with them in the future? For one, we attribute fundamental social motives even to simple geometric shapes (Heider & Simmel, 1944). Consistent with media equation theory (Nass et al., 1997; Nass & Moon, 2000; Reeves & Nass, 1996), autonomous virtual agents represented by an avatar elicit social responses that are very similar to human–human interaction (von der Pütten et al., 2010). Similarly, watching robots being mistreated or subject to painful stimulation has been shown to elicit empathic concern and increased physiological arousal (Rosenthal-von der Pütten et al., 2013;

Correspondence should be addressed to Dennis Küster, Department of Computer Science, University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany. (E-mail: kuester@uni-bremen.de).

This research was supported by a grant Sonata (2016/23/D/HS6/02954) from the Polish National Science Centre (Narodowe Centrum Nauki), and by the German Research Foundation (DFG) through the project "PALMS" (no. 402779631), which is a subproject of the DFG Priority Program "The active self" (SPP 2134; no. 360292909).

Suzuki et al., 2015). This suggests that robots can at times be perceived *almost as if* they were human.

The tendency to imbue non-human entities with human characteristics is called anthropomorphism (Epley et al., 2007). At the core of anthropomorphism lies the attribution of human-like mental capabilities, as it allows us to explain non-humans' observable behaviours, predict their future actions and facilitate effective interactions with our immediate social environment (Waytz et al., 2010a). Perception of non-humans in terms of their inner capacities therefore corresponds to how we perceive other human beings, and fulfils an analogous function (Epley et al., 2007; Waytz et al., 2010a). Human-like appearance, and especially human-like face, is an exceptionally strong social cue that facilitates this process (e.g. Duffy, 2003; Fink, 2012). Indeed, human-likeness has been demonstrated to enhance emotional engagement and boost empathy towards robots (Riek et al., 2009). Conversely, when artificial social partners become too human-like, they may be rejected (Rosenthal-von der Pütten et al., 2019).

Commonalities in responses to humans and non-humans have been supported by neuroscientific evidence that similar brain regions are activated when we think about both kinds of agents, and robots in particular (Gazzola et al., 2007). As more recent work in this field suggests, human-like agents sometimes also produce similar *cognitive* effects, for example, during categorical perception and cognitive conflict processing tasks in response to spectra of images morphed between different types of artificial agents (e.g. robot–human and robot–animal; Wiese & Weis, 2020), or in a Go/No-Go task played against either another human, or a Cozmo (Anki) toy robot (Currie & Wiese, 2019). Nevertheless, some of this research also points to possible differences, such as subtly different strategies employed when playing against a robot (Currie & Wiese, 2019), and differences in the *extent* of conflict processing for subjects responding to morph continua containing fully human agents (Wiese & Weis, 2020).

Moral interactions between humans and robots

Moral interaction with members of other social groups is a highly complex phenomenon. For example, factors such as social distance or perceived similarity, that is, relevant attributes when distinguishing between robots and humans, are known to impact our willingness to grant or deny equal levels of humanness even to other members of the human species (e.g. Haslam & Loughnan, 2014; Waytz & Epley, 2012). At the same time, research into the social-cognitive underpinnings of dehumanisation, that is, the denial of humanness to others, has yielded surprising insights into how *mechanistic* notions may play a role in its certain forms (Haslam, 2006). Thus, a human-like robot might be subject to both: dehumanisation, and perhaps also its reverse, that is, anthropomorphism (Waytz et al., 2010b). Therefore, the question of how we might respond to signs of moral intentions towards a human-like robotic other (see Gunkel, 2018), for good or ill, still requires further investigation.

Some anecdotal evidence suggests that inhibitions towards harming robots might be low. For example, in 2015, a hitchhiking robot called *Hitchbot* was vandalised on its way across the United States (Darling, 2015; Gunkel, 2018). Conversely, however, people might sometimes develop surprisingly strong feelings for robots that are taking harm on their behalf. For example, this was the case for soldiers and explosive ordnance disposal robots designed to clear minefields (Garreau, 2007). Here, the seemingly selfless and benevolent actions of the robot appeared to sensitise the soldiers when they witnessed how the robot was torn apart. In a similar vein, the story about the travels of *Hitchbot* (Darling, 2015) is itself a rhetorical artefact that could hardly have gathered as much attention if there was no reason to care about the robot in the first place. Taken together, these real-life stories suggest that it may be non-trivial to decide whether robots might merit some kind of moral standing from an ethical point of view (see Gunkel, 2012, 2018). Furthermore, it appears worth investigating from a psychological perspective, under what conditions robots are granted or denied mental capacities in the minds of human perceivers.

Perceiving less mind: Moral typecasting or dehumanisation?

Our minds are often seen as the hallmark of humanity (Gray & Wegner, 2012a). Interestingly, however, we do not necessarily construe someone's mind as a single unified structure. Instead, we tend to perceive the mind of other entities along two broad dimensions, labelled *experience* and *agency* (Gray & Wegner, 2012b). *Experience* is about the capacity to feel and to sense (Gray & Wegner, 2012a), which includes feelings of hunger, pain or pleasure, as well as personality and consciousness (Gray et al., 2007; Gray et al., 2012). The second dimension, *agency*, is about the capacity to intend and to act (Gray et al., 2012), entailing capacities for self-control, communication, and memory (Gray et al., 2007). According to moral typecasting theory (MTT), moral interactions emerge within a moral dyad, which consists of a moral patient and a moral agent (Gray & Wegner, 2009). In the relationship between moral agents and patients, a moral patient is subject to moral actions, both good and evil, carried out by a moral agent (Gray, 2010). However, the relationship between both is asymmetrical. That is, agents require mental capacities that support *doing* to carry out an action (i.e. agency; Waytz et al., 2010c), whereas patients suffer the effects of the agents' actions, and therefore need a capacity to *feel* (i.e. experience; Waytz et al., 2010c). Attributions of experience differentiate the perceptions of humans and robots particularly well (Gray et al., 2007; Gray & Wegner, 2012a).

Previous research has demonstrated that participants who read about intentional infliction of harm onto a liminally minded moral patient (e.g. a vegetative state patient), spontaneously attributed more mind to the patient (Ward et al., 2013). This effect, labelled *the harm-made mind*, has been explained on the basis of the patient's mere participation in a dyadic moral interaction (Ward et al., 2013). That is, if there is a moral agent who acts intentionally on another entity, then the second entity is automatically perceived to fill the position of a patient. The moral patient then completes the moral dyad, and is attributed correspondingly more mind (Ward et al., 2013). The harm-made mind effect has been replicated across different types of social entities, including a "highly complex social robot" (Ward et al., 2013, p. 1442), and harmed faces manipulated to appear robotic (Swiderska & Küster, 2018). As a consequence of this *dyadic completion*, immoral actions may increase perceptions of pain and suffering even when no moral patient is explicitly present (Gray et al., 2014). However, mind perception may not be enhanced in this manner if the victim's base level of mind is already very high (Ward et al., 2013, study 5). In fact, harm in the moral dyad resulted in significantly *less* mind attribution when the victim was a conscious human to begin with. Ward et al. (2013) speculated that this surprising finding might be explained by a dehumanisation account, suggesting that dehumanisation may occur for entities that have a mind (humans)—but not for "entities with absent or liminal minds" (p. 1443), such as robots.

Intriguingly, possession of human-like mental capacities is thus further linked to the ascription of moral status, as discussed by dehumanisation theory. Appearing mindless reduces moral standing, equates an entity with a mere object deprived of all emotional experiences and agency-related competencies, and warrants for it to be treated accordingly (Waytz et al., 2010c). *Dehumanisation* (Haslam, 2006) is a process that can itself be mapped onto the dimensions of agency and experience. Specifically, *experience*-related mechanisms of *mechanistic dehumanisation* have been hypothesised to justify distancing, lack of empathy, and diminished willingness to help others (Andrighetto et al., 2014; Haslam, 2006). Apart from mechanistic dehumanisation, the cognitive underpinnings of *animalistic* dehumanisation, that is, perceiving other people to lack what distinguishes us from animals (Haslam, 2006), can be tied to a denial of the agency dimension of mental competencies.

One of the mechanisms underlying dehumanisation of outgroups has been linked to beliefs in human superiority over other species, while perceptions of similarity contribute to greater humanisation and greater empathy (Costello & Hodson, 2010; Krebs, 1975). In the same vein, as shown by Andrighetto et al. (2014), the two different forms of dehumanisation may affect the relative willingness to help different types of social outgroups after they have been struck by natural disasters. In their study (Andrighetto et al., 2014), Italian participants differently dehumanised Haitians, who were dehumanised as animal-like (i.e. denied mental capacities associated with agency), whereas Japanese were dehumanised as automata (i.e. denied mental capacities associated with experience. However, in both cases, reduced empathy emerged as a mediator that explained the negative effects of dehumanisation on helping. Perceived divisions between species can be shaped even by subtle experimental manipulations, such as apparent threats posed by others (Costello & Hodson, 2014), or by social robots that look too similar to humans - thereby threatening human distinctiveness (Ferrari et al., 2016). Perceived threat amplifies dehumanisation, which in turn has been associated with an unwillingness to help the threatening target (Haslam & Loughnan, 2014). Conversely, signs of love, gratitude and cuteness have been suggested to imbue others with positive social value, resulting in re-humanisation of another through social engagement (Sherman & Haidt, 2011). In consequence, expressions of benevolent intentions by an agent towards a robot might reduce dehumanisation.

From an MTT perspective, essentially any type of moral action directed at a patient should increase attributions of experience to the patient (see Ward et al., 2013), and attributions of agency to the agent (Gray & Wegner, 2009). According to MTT, intentional harms should generally be perceived as worse than identical unintended harms (Ames & Fiske, 2013). Intentional pain from electric shocks stings worse (Gray & Wegner, 2008), as do intentional insults (Gilbert et al., 2004). In MTT, such findings could be explained by an even greater asymmetry of mind attributions in the moral dyad if harm is caused intentionally rather than accidentally (cf., Gray & Wegner, 2009). In consequence, even more agency should be attributed to intentionally harmful agents, and an even greater capacity for experience should be granted to intentionally harmed patients. This, however, appears to contrast with certain findings established by dehumanisation theory. Accordingly, harmful agents should rather be perceived to be *less* agentive than benevolent agents (Khamitov et al., 2015). For example, offenders may be *denied* mental capacities closely related to both agency and patiency (Bastian et al., 2013)—and people themselves report feeling less human after having behaved immorally (Kouchaki et al., 2018). Consistent with the dehumanisation account, a series of six vignette studies showed a surprising denial of agency to intentionally malevolent agents (Khamitov et al., 2015). Furthermore, malevolent agents were viewed as less worthy of moral consideration, and this effect was mediated through reductions in perceived agency. Finally, lending initial support to the converse argument of a humanisation effect through benevolent intentions, mind perception was facilitated if participants imagined actively helping a robot in a recent vignette study (Tanibe et al., 2017). This raises the question if mind-imbuing intentions indeed need to be harmful in nature, or if benevolent intentions, as expressed by signals of love, cuteness or affection might be equally, or even more, potent drivers of enhanced mind attribution towards human-like robots or human avatars.

The present research

The present study aimed to contrast predictions made by MTT and dehumanisation theory from a different angle than in the previous study by Khamitov et al. (2015). Notably, we concentrated on mind attribution to moral patients. We chose to examine a human and a robotic avatar to compare two entities of moderate to high human-likeness,[1] wherein the human avatar should be perceived as possessing a mind to begin with. First, in line with MTT and the harm-made mind effect, we hypothesised that the robotic avatar would be attributed more mind (experience) in a harm context, whereas the human avatar should be victimised (i.e. reduced mind attributions; cf. Ward et al., 2013, study 5). Alternatively, from a dehumanisation perspective, the robotic avatar should likewise be dehumanised (attributed less mind) because it is a member of a social outgroup (robots) that is already perceived as less human to begin with.

Our second hypothesis concerned the impact of social intentions: Following dehumanisation theory, we expected benevolent intentions to lead to (more) rehumanisation of its targets, as mediated by empathy. In contrast, malevolent intentions should lead to dehumanisation of both its targets. Here, an MTT account based on dyadic completion would predict no differences in the impact of the two intentions on the perceptions of the avatars. To complement our second hypothesis, we therefore examined the MTT proposition that intentional harms are worse than accidental harm (Ames & Fiske, 2013; Ward et al., 2013). Here, dehumanisation theory would predict greater moral outrage towards the agent (Bastian et al., 2013) but possibly also stronger victim (patient) dehumanisation (Castano & Giner-Sorolla, 2006), and a more important role of empathy (e.g. Čehajić et al., 2009).

Third, we aimed to address the limitation that most previous work in this field has been based exclusively on textual vignettes that require deliberate processing by participants. We therefore operationalised socio-moral intentions by means of a set of visual vignettes. Finally, we aimed to explore the notion, supported by MTT, that enhanced mind perception for moral patients should be mediated by perceived pain (Ward et al., 2013), and possibly by empathy for the patient (Swiderska & Küster, 2018). In addition to manipulating socio-moral intentions, we therefore included a visual depiction of harm in the form of a facial wound depicted on the moral patient. To test our hypotheses, we designed a visual-vignette study that depicted human and robotic avatars as moral patients who were either unharmed (control condition) or harmed (hypothesis 1). The same avatars were furthermore subject to an explicit display of either malevolent or benevolent (reconciliatory) intentions of a moral agent (hypothesis 2).

METHOD

Participants

A power analysis with G*Power (Faul et al., 2007) indicated that 245 participants would be sufficient to detect small to medium sized main and interaction effects (Cohen's $f = 0.18$) in a $2 \times 2 \times 2$ analysis of variance (ANOVA) with 80% power. In a single wave of data collection, 217 participants were recruited via Crowdflower (http://www.crowdflower.com/) and compensated 1 USD each, and another 81 participants were recruited, without compensation, via a free online survey website (http://psych.hanover.edu/research/exponnet.html). Out of these 298 participants, data from 45 had to be excluded,[2] yielding a final sample of 253 participants (162 women; $M_{\text{age}} = 38.36$, $SD = 13.34$). The study was approved by the Ethics Committee at the Department of Psychology, Warsaw University, Poland.

Ethical compliance statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The authors declare that they have no conflict of interest. Informed consent was obtained from all individual adult participants included in the study.

[1] Previous work, such as the studies presented by Ward et al. (2013) may have been subject to a certain degree of floor effects (studies 1 to 4) or ceiling effects (study 5) with respect to the potential for raising/lowering mind attribution towards the respective moral patients.

[2] Out of these, 42 failed to complete the experiment, and three participants made ineligible entries.

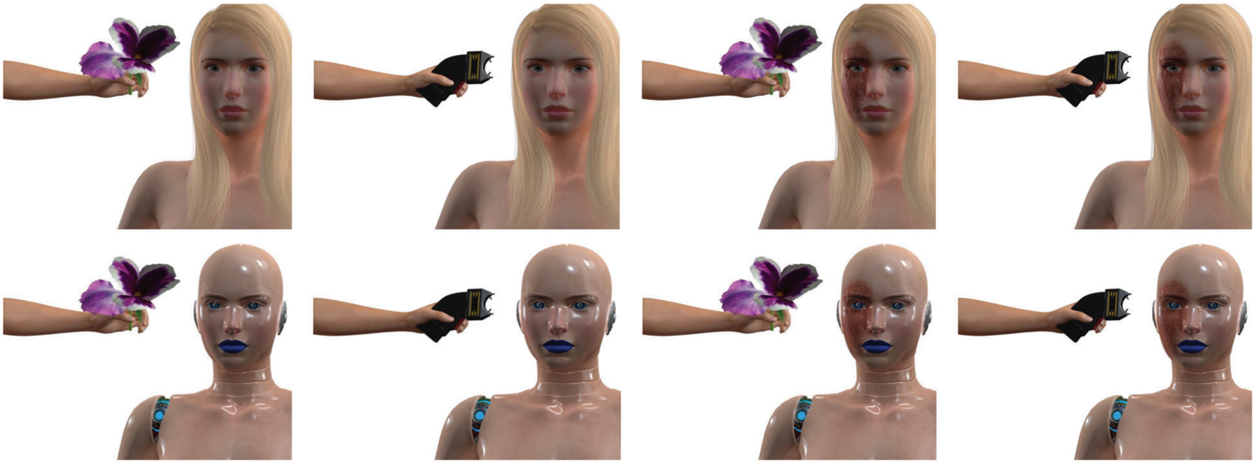

Figure 1. Visual vignettes used in Experiment 1. [Colour figure can be viewed at wileyonlinelibrary.com].

Materials

We created eight visual vignettes using Poser Pro 2014 (Smith Micro). They depicted a malevolent or benevolent intention towards two types of moral patients (human avatar, robotic avatar), who either showed signs of harm or not (harmed, unharmed; Figure 1). The malevolent intention was represented by a taser pointed at the avatar, and the benevolent intention by a flower. The avatars showed a neutral facial expression, and the same base skin textures were used to ensure the expression was identical for both of them. To generate the harmed versions of the faces, a moderately severe burn was added to their right side. The images measured 948 × 711 pixels and were displayed on a white background.

Procedure

Participants were presented with a single, randomly selected, vignette. Their task was to evaluate the intention and the moral patient on a number of characteristics (see the next section for details). The experiment was delivered in English through EFS Survey (Version 9.0, QuestBack AG, Germany).

Dependent measures

We assessed the avatars' perceived *experience* via seven main items (having personality, experiencing desire, feelings, emotions, pleasure, hunger, fear), supplemented by two items on conscious mental experience (being conscious of itself, being conscious of the people and world around it), as adapted from Ward et al. (2013). The two latter items have previously been shown to be closely associated with the experience dimension (Gray et al., 2007), and were included to obtain an overall index of the patient's *mind* (α = .96; cf., Ward et al., 2013[3]). We further explored the avatars' perceived capacity for *pain* as a mediator of mind perception (Ward et al., 2013), and included an adapted measure of state empathy towards the patient (seven items, e.g. "I could show compassion for [Ann]", "I could share [Ann's] feelings"; α = .94; see Shen, 2010) as an alternative potential mediator (Swiderska & Küster, 2018). To compare moral typecasting and dehumanisation accounts, we included a *mechanistic dehumanisation* scale (five items, e.g. treating the patient as an object, as a means to an end; α = .86; adapted from Bastian & Haslam, 2011), and a measure of dehumanisation-induced emotional experiences attributed to the avatars (14 items, e.g. confusion, distress; α = .98; Cuddy et al., 2007). The response scales for all of the above scales were presented as 7-point Likert scales with labels at the end-points ranging from 1 = *strongly disagree* to 7 = *strongly agree*. Finally, we included a question about how morally right or wrong the intended action appeared, again as a 7-point Likert scale (1 = *definitely wrong*, 7 = *definitely right*).

RESULTS

ANOVA with Intention Type (malevolent, benevolent), Harm (harmed, unharmed) and Avatar Type (human, robotic) as between-subjects factors was conducted on experience. Participant gender was included as a covariate, but its effect was non-significant (all $ps > .30$), and we excluded it from all further analyses. The main effects of harm and avatar type were significant, respectively

[3] Ward et al. (2013) additionally included *agency* as part of an overall composite index of mind attribution. However, as agency has been conceptualised as a property of moral agents, we did not include this measure to assess changes in attributions of mind to the moral patient.

TABLE 1
Descriptive statistics for the harmed and unharmed versions of human and robotic avatars as targets of malevolent and benevolent intentions

	Malevolent intention				*Benevolent intention*			
	Harmed		*Unharmed*		*Harmed*		*Unharmed*	
	Human	*Robotic*	*Human*	*Robotic*	*Human*	*Robotic*	*Human*	*Robotic*
Variable	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Pain	5.30 (2.16)	2.77 (2.34)	4.97 (2.15)	2.19 (1.73)	5.97 (1.72)	2.45 (1.72)	4.32 (1.74)	1.93 (1.46)
Mind	4.42 (1.74)	2.28 (1.60)	4.19 (1.73)	2.00 (1.08)	5.37 (1.54)	2.42 (1.31)	4.05 (1.49)	2.26 (1.35)
Empathy	4.86 (1.46)	3.10 (1.91)	4.46 (1.32)	2.68 (1.20)	4.99 (1.23)	3.28 (1.69)	4.17 (1.35)	2.61 (1.49)
Dehumanisation	4.25 (1.71)	5.14 (1.55)	4.23 (1.27)	5.15 (1.40)	2.57 (1.37)	4.07 (1.53)	3.33 (1.46)	4.03 (1.22)
Emotions	3.84 (2.10)	2.19 (1.68)	2.83 (1.64)	1.69 (1.14)	3.23 (1.55)	1.92 (1.32)	2.58 (1.50)	1.84 (1.30)
Moral evaluation	2.16 (1.57)	3.54 (1.76)	2.44 (1.34)	3.72 (1.65)	4.76 (1.75)	4.21 (1.47)	5.35 (1.25)	4.50 (1.35)

Note: For moral evaluation, lower values indicate that the intention was perceived to be more morally wrong.

$F(1, 245) = 6.82$, $p = .01$, $\eta_p^2 = .03$ and $F(1, 245) = 142.38$, $p < .001$, $\eta_p^2 = .37$. Human avatars were perceived to have a higher level of experience than the robotic avatars, and visible harm enhanced mind perception—suggesting the presence of a general harm-made mind effect (see Table 1 for details). However, the main effect of intention type did not reach statistical significance, $F(1, 245) = 2.54$, $p = .113$, $\eta_p^2 = .01$. No significant interactions emerged (all ps > .10).[4]

A similar pattern of results was obtained for pain. The main effects of both harm and avatar type were significant, $F(1, 245) = 10.04$, $p = .002$, $\eta_p^2 = .04$ and $F(1, 245) = 133.66$, $p < .001$, $\eta_p^2 = .35$, with human avatars perceived to be able to experience pain to a higher degree than robotic avatars, and harmed avatars experiencing more pain than their unharmed counterparts (Table 1). The main effect of the intention type was again not significant, $F(1, 245) = .56$, $p = .564$, $\eta_p^2 = .00$, and there were no significant interactions (all ps > .10).

As per our first hypothesis, we expected the appearance of the robotic avatar to be associated with a denial of humanness that should be reflected by mechanistic dehumanisation, attribution of emotions and expressed empathy. We thus conducted a multivariate analysis of variance with the three dependent measures. All multivariate main effects were significant: Intention, $F(3, 243) = 17.26$, $p < .001$, $\eta_p^2 = .18$, harm, $F(3, 243) = 3.80$, $p = .011$, $\eta_p^2 = .05$, and avatar type, $F(3, 243) = 30.41$, $p < .001$, $\eta_p^2 = .27$. No significant interactions were observed (ps > .10), suggesting little empirical support for our first MTT-based hypothesis. Consistent with our second hypothesis, in the univariate tests, the main effect of intention type was significant for dehumanisation, $F(1, 245) = 42.02$, $p < .001$, $\eta_p^2 = .15$, whereby benevolent intent reduced dehumanisation compared to malevolent intent (see Table 1). Intention type did not reach significance for emotions, $F(1, 245) = 3.82$, $p = .216$, $\eta_p^2 = .01$, or empathy, $F(1, 245) = .00$, $p = .958$, $\eta_p^2 = .00$. However, harm was significant for emotions, $F(1, 245) = 7.97$, $p = .005$, $\eta_p^2 = .03$, and empathy, $F(1, 245) = 9.50$, $p = .002$, $\eta_p^2 = .04$, with both found to be higher for harmed entities. Conversely, harm was not significant for dehumanisation, $F(1, 245) = .92$, $p = .338$, $\eta_p^2 = .00$. Avatar type was significant for all three measures: empathy, $F(1, 245) = 82.70$, $p < .001$, $\eta_p^2 = .25$, dehumanisation, $F(1, 245) = 29.66$, $p < .001$, $\eta_p^2 = .11$, and emotions, $F(1, 245) = 37.09$, $p < .001$, $\eta_p^2 = .13$. Participants appeared to empathise less with the robotic avatar than with the human avatar, and the robotic avatar was in turn more dehumanised (also via the denial of emotional experiences) than the human avatar. Surprisingly, intention type could only be shown to affect dehumanisation but not empathy and perceived capacity of the moral patient for emotions, whereas visible harm clearly affected emotions and empathy, but not dehumanisation. These results suggest that, while both manipulations resulted in substantial effects, expressed empathy towards the victim and decreased dehumanisation in the moral dyad might be elicited relatively independently from one another.

Consistent with our second hypothesis, an ANOVA on the moral evaluation yielded a significant main effect of intention type, $F(1, 245) = 80.74$, $p < .001$, $\eta_p^2 = .25$, and a significant interaction between intention type and avatar type, $F(1, 245) = 27.70$, $p < .001$, $\eta_p^2 = .10$. The main effects of harm and avatar type were not significant, $F(1, 245) = 2.70$, $p = .085$, $\eta_p^2 = .01$ and $F(1, 245) = 2.63$, $p = .106$, $\eta_p^2 = .01$, which was also the case for all remaining interaction effects ($ps > .50$). Malevolent intentions were viewed as more morally wrong when targeted at the human avatar

[4] While none of the interaction effects reached statistical significance, two of these effects were close to $p = .10$. It is therefore possible that subtle interaction effects could not be revealed due to insufficient statistical power. Specifically, the interaction between Harm and Avatar Type reached $p = .139$, the interaction between Avatar Type and Intention Type $p = .589$, the interaction between Harm and Intention Type $p = .204$, and the three-way interaction (Intention Type, Harm, Avatar Type) reached $p = .115$.

compared to the robotic avatar ($M_{\text{malevoent,human}} = 2.29$, $SD = 1.47$ vs. $M_{\text{malevoent,robot}} = 3.63$, $SD = 1.70$; $p < .001$). Benevolent intentions were perceived as more morally right in this comparison, but here the difference between the two types of avatars appeared to be less pronounced ($M_{\text{benevolent,human}} = 5.07$, $SD = 1.53$ vs. $M_{\text{benevolent,robot}} = 4.35$, $SD = 1.41$; $p = .014$). Together, these results appeared to be consistent with the notion of our second hypothesis, that is, that benevolent intentions should lead to rehumanisation, whereas malevolent intentions foster dehumanisation.

To further clarify the understanding of intentions in this type of visual vignettes, we conducted a brief follow-up study with 28 participants (11 women; $M_{\text{age}} = 37.71$, $SD = 10.09$). It revealed that 100% of them recognised the object held by the hand as a weapon, and similarly, 100% recognised a flower in the respective images. Moreover, 96% (27 participants) interpreted the depicted actions to express corresponding malevolent and benevolent intentions. When asked explicitly about how intentional (vs. accidental) the actions appeared to be, both behaviours were perceived as highly intentional, with the malevolent action significantly more intentional than the benevolent action ($M = 6.36$, $SD = 1.37$ vs. $M = 5.86$, $SD = 1.67$, $p = .028$). These results support the validity of our manipulation.

Mediation analyses

As demonstrated previously, mind perception in response to intentional harm may be mediated by perceived pain in moral vignettes (Ward et al., 2013), whereas both pain and empathy appear to play a mediating role when participants are exposed to visible harm (Swiderska & Küster, 2018). Empathy for the victim may especially be of importance as a mediator for mind attributions if there is evidence of an *empathy gap* towards another entity (e.g. Gutsell & Inzlicht, 2012; Swiderska & Küster, 2018). To go beyond previous work in this area, the present study therefore manipulated harm and intention type separately to help dissociate these two influences on mind perception.

We performed mediation analyses on all three independent variables (avatar type, harm, intention type) to re-examine their respective roles in relation to mind perception, and to explore whether the same mediators could be observed for intention type (Figure 2). We conducted three bootstrapping mediation analyses (10.000 samples; SPSS PROCESS macro, V.3.0; Hayes, 2018; model 4) using 95% confidence intervals (CI) with pain and empathy as parallel mediators and the composite mind index as the dependent variable.

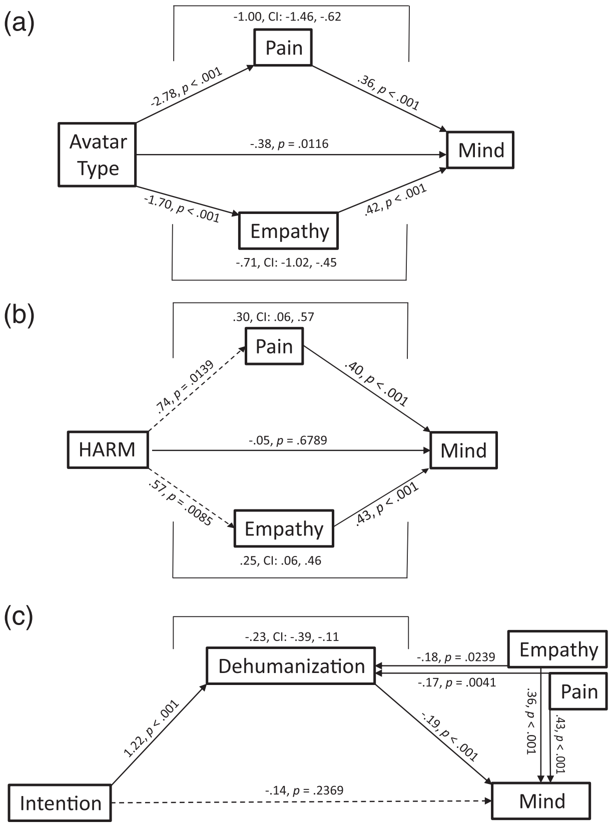

Figure 2. Results of the mediation analyses with avatar type (a), harm (b) and intention (c) as independent variables (IV), index of mind attribution as the dependent variable, perceived capacity for pain and empathy expressed for the moral patients as parallel mediators in the first two models, and dehumanisation as a mediator and pain and empathy as covariates in the third model. Solid lines represent significant paths, solid brackets represent significant mediation effects, central paths represent the direct effect when controlling for the mediators. IVs were dummy coded with 0 and 1. $N = 253$.

For avatar type, the significant relationship with mind attributions (CI = [−2.49, −1.71], $p < .0001$) could be partially explained by a significant indirect effect of both mediators (total indirect effect = −1.72, CI = [−2.12, −1.37]; see Figure 2a). The direct effect of avatar type remained significant after controlling for the mediators (CI = [−.68, −.09], $p = .0116$). Both individual indirect effects were significant (pain: −1.00, CI = [−1.46, −.62]; empathy: −.71, CI = [−1.02, −.45]. The contrast between pain and empathy revealed no significant difference (CI = [−.92, .31]. These findings are consistent with previous results (Swiderska & Küster, 2018) and point to a mediation of appearance-related changes in mind perception by both the perceived capacity for pain and participants' reported empathy.

For harm, the relationship with mind attributions was significant (CI = [.03, .96], $p = .0378$), and this was accompanied by a significant total indirect effect of pain and empathy (.54, CI = [.15, .95]; Figure 2b). The relationship turned non-significant when controlling for the two mediators (CI = [−.30, .20], $p = .6598$). Again, both individual indirect effects were significant (pain: .30, CI = [.06, .57]; empathy: .25, CI = [.06, .46]), and the contrast between them did not demonstrate a significant difference (CI = [−.16, .29]). Thus, consistent with previous findings, harm done to a moral patient affected mind perception through perceived pain as a mediator (Ward et al., 2013). In addition, in the present study, empathy was found to be an equally effective mediator of increased mind perception. This finding might be due to the visual nature of our materials that may have rendered harm more salient than traditional text-based vignettes. Taken together, these results suggest that perceived capacity for pain and participants' reported empathy with the agent may both play significant roles as mediators of mind perception.

Next, we examined the impact of intention type on mind perception with the same mediation model as for avatar type and harm. Here, however, the consistent pattern of results found for the previous models could not be replicated.[5] Therefore, we explored the possibility that in the case of the relationship between intentions and mind perception, pain and empathy acted as confounding variables. This was confirmed in an auxiliary analysis of covariance with pain and empathy as covariates, respectively, $F(1, 243) = 99.44$, $p < .001$, $\eta_p^2 = .29$ and $F(1, 243) = 56.99$, $p < .001$, $\eta_p^2 = .19$, whereby the main effect of intention type on mind attributions became significant, $F(1, 243) = 10.91$, $p = .001$, $\eta_p^2 = .04$, and benevolent intentions resulted in greater attributions than malevolent intentions ($M_{\text{benevolent}} = 3.55$, $SD = 1.90$ vs. $M_{\text{malevolent}} = 3.25$, $SD = 1.90$).

Finally, we conducted an exploratory mediation analysis to examine the impact of intention type on mind perception with dehumanisation as a mediator, and pain and empathy as covariates. Intention type might influence mind perception via humanisation of moral patients if it appears benevolent, and dehumanisation if it is malevolent. The relationship of intention with mind was significant (CI = [−.59, −.15], $p = .0010$), and the indirect effect of dehumanisation was significant as well (−.23, CI = [−.39, −.11]; Figure 2c). This relationship became non-significant when controlling for the mediator (CI = [−.37, .09], $p = .2369$). Both covariates were significantly related to mind perception (pain = .43, CI = [.35, .50], $p < .001$; empathy = .36, CI = [.26, .46], $p < .001$) and dehumanisation of moral patients (pain = −.17, CI = [−.28, −.05], $p = .0041$; empathy = −.18, CI = [−.34, −.02], $p = .0239$).

DISCUSSION

The main aim of the current work was to test diverging predictions made by MTT and dehumanisation theory towards highly human-like entities in the role of moral patients. Dehumanisation theory has raised the notion that moral intentions may lead to either a denial of mental capacities or rehumanisation, depending on whether the moral interaction is malevolent or benevolent in nature. In contrast, the MTT mechanism of dyadic completion should result in the same type of asymmetric mind perceptions for both types of moral behaviours. Further, MTT has sometimes emphasised the role of malevolent intentions (e.g. Gray & Wegner, 2009), whereas research favouring a dehumanisation account has regarded intentionality as simply another aspect of agency (Khamitov et al., 2015). Finally, work from MTT has suggested a possible reversal of mind attribution towards moral patients possessing a pre-existing fully human level of mind (Ward et al., 2013), whereas dehumanisation theory has shown how even subtle differences between us and another person, for example in social status or appearance, can result in a denial of mental capacities in everyday life (Bastian & Haslam, 2011; Haslam & Loughnan, 2014). We therefore examined the influence of malevolent and benevolent intentions, visible harm, and robotic appearance on mind perception using visual vignettes. In line with our predictions, the results replicated the harm-made mind effect for both human-like and robotic avatars. However, we did not find the expected impact of intentions. Instead, exposure to benevolent intentions reduced dehumanisation of the moral patient. This finding appears to be more consistent with a dehumanisation view of the harm-made mind than with a moral typecasting account (see also Khamitov et al., 2015).

The visual vignettes further examined the role of visible harm as a signal of suffering. The addition of a facial wound resulted in enhanced mind attribution, increased state empathy and pain, and enhanced attribution of emotions to the avatars. Thus, despite its potential for being imbued with negative social value (see Sherman & Haidt, 2011), visible harm appeared to humanise both kinds of entities in terms of their perceived mental and emotional capacities. Importantly, these response tendencies emerged similarly for both the robotic and the human-like avatar, supporting media equation views that humans may overall react similarly to humans as they react towards humans (e.g. Rosenthal-von der Pütten et al., 2013).

Nevertheless, our results also demonstrate that mind attribution and empathy expressed towards artificial entities are highly malleable and responsive to subtle social cues. The human-like avatars were perceived to possess a greater level of mind, as more capable of experiencing pain and emotions, elicited more empathy, and were dehumanised to a lesser degree. Furthermore, there

[5] The analysis yielded a non-significant relationship between Intention and mind attributions (CI = [−.76, .18], $p = .2311$), with a non-significant total effect of pain and empathy (total indirect effect = .07, CI = [−.33, .48]), a significant direct effect of Intention when controlling for the two mediators (CI = [−.60, −.12], $p = .0039$), and non-significant individual direct effects (pain = .06, CI = [−.19, .30]; empathy = .01, CI = [−.17, .21]).

was one notable exception to the general pattern of main effects: Participants reported a stronger contrast between right and wrong when the moral patient was a human-like avatar. Thus, while humans may generally perceive humans and human-like robots in a very similar manner, they also appeared to be biased towards fellow humans in that higher moral standards were applied to more human-like victims. This finding appears consistent with notions presented by the dehumanisation view, and is in line with recent findings suggesting that a human moral patient might suffer less when depicted together with a robotic agent (Swiderska & Küster, 2020). At the same time, it is remarkable that robots were perceived as worthy of moral consideration at all—that is, that humans distinguish between right and wrong actions also towards the robotic avatar (see Gunkel, 2012).

The results of the first two mediation analyses suggest that pain and empathy may play similarly important roles for understanding the mechanisms underlying mind attribution towards a visibly harmed other. However, while our present findings of a parallel mediation through both pain and empathy replicate findings from recent research (Swiderska & Küster, 2018; Ward et al., 2013), it is unclear under which circumstances both constructs will be sufficiently distinct to allow further insights based on these still rather exploratory results. Furthermore, our results raise the question to what extent pain plays as central a role for "mind infusion" effects as claimed by previous work on the harm-made mind (e.g. Ward et al., 2013). Rather, as discussed recently by Schein and Gray (2018), we may have to care about a vulnerable mind via empathy in the first place, for the mere perception of pain and suffering to robustly affect moral judgement (see also Avenanti et al., 2010). Therefore, future work might examine moderated mediation models by means of more robust measures for pain and empathy. Unfortunately, however, the necessary experimental control over both of these factors remained outside of the scope of the present work. Nevertheless, the current results lend some support to the notion that exposure to visible harm may tap into mechanisms regulating our potential for empathy towards members of another species (see Gutsell & Inzlicht, 2012).

Overall, our results on displays of benevolent and malevolent socio-moral intentions suggest that intentions might be more loosely connected to mind perception than suggested by some of the previous work on moral typecasting (e.g. Ward et al., 2013). In particular, the role of benevolent intentions may deserve further attention. Previous work on MTT has predominantly focused on the differences between mere accidental harm on the one hand, and malevolent intentions on the other (e.g. Gray & Wegner, 2009). We argue that this theoretical perspective may have largely overlooked the humanising potential of more benevolent displays. On first glance, this may appear rather counterintuitive, as benevolent intentions should be rather unlikely to be associated with causing harm to another person. However, benevolence, kindness or even signals of cuteness associated with another entity should be effective means for enhancing humanisation of a victim from a dehumanisation perspective (see Gray, 2012; Sherman & Haidt, 2011). In this experiment, we therefore aimed to further decouple the *presence of harm* indicated by the facial wound from the *socio-moral intentions* displayed towards the moral patient. Benevolent intentions should have provided an even sharper contrast to malevolent intentions to the accidental harm studied in previous work on MTT (e.g. Ward et al., 2013). Our findings of no significant differences in mind attribution between both conditions thus suggest that seeing either kind of intense socio-emotional intentions might be sufficient to elicit mind infusion effects in moral dyads. That is, MTT's dyadic completion mechanism would not be inconsistent with these findings. However, it is possible that a more powerful contrast between benevolent and malevolent intentions could have revealed a significant difference in favour of a dehumanisation account. Here, future work might employ more powerful displays of benevolent intentions than the somewhat subtle gesture of handing someone a flower. Finally, our exploratory mediation analyses suggest that the observed mind attribution effects could be mediated by dehumanisation once variance from pain and empathy is controlled for. Again, these results call for more consideration of dehumanisation processes in moral interactions and moral typecasting.

We further suggest that a more extensive visual vignette approach, or work using popular online videos featuring "robot abuse" (e.g. Küster et al., 2020), might help to illuminate the role of harm and socio-moral intentions, as showing an interaction instead of describing it in words may bring about less ambiguity about certain details that participants would otherwise have to imagine. Importantly, visual vignettes could offer a fresh perspective on the harm-made mind, as well as on broader discussions on intention. A lot of insightful work has been conducted in the philosophical and psychological literature to investigate the interactions of factors contributing to lay person's judgements about intentional vs. unintentional action (e.g. Mele & Cushman, 2007), and the folk concept of intentionality (e.g. Malle & Knobe, 1997). However, as with prior research on the harm-made mind, work in this area has predominantly been shaped by purely text-based analyses. Here, extended work on image-based representations could help illuminate under which conditions brief non-verbal observations in the real world are, for example perceived as blameworthy intentional norm violations (Monroe & Malle, 2017), or as side-effect actions (Malle & Knobe, 1997).

Despite these encouraging results, the present work still faces a number of limitations. First, all of our measures were based on simple self-report data that could be strengthened in future work through the use of behavioural or physiological responses—for example skin conductance as a measure of arousal. In addition, a few of present dependent measures, for example the emotional index, might be eliminated that essentially mirrored the pattern results of the main dependent variables. Furthermore, our current results are based on a single visual vignette study that could not fully account for a number of possible alternative accounts and ambiguities inherent to this approach. As demonstrated by the results of the post test, participants were able to clearly distinguish between benevolent and malevolent intentions, and they recognised the taser and the flower accordingly. However, we do not know what specific kind of benevolent intentions were perceived by participants, or if other types of signals of socio-moral intentions might have been more effective. For example, participants may have interpreted our without-harm malevolent intentions somewhat more broadly as a situation where "harm is coming" rather than specifically as an abstract goal-directed intention to cause harm. Nevertheless, we believe that even such generalisations and variations would still capture the gist of the difference between a clear and imminent sense of (intentional) harm that might perhaps still be averted, and actual physical harm that has already occurred. Likewise, the precise sequence of events resulting in harm to the patient was left to the imagination of the participant. Here, a traditional textual vignette might have achieved more precision - yet even for textual vignettes, it is not trivial to unambiguously describe a specific type of intention without participants second guessing additional moral motivations of the agent. Further, we believe that the limitations in control over participants' imagination of additional context in our visual vignettes were balanced by the greater control offered over the type and intensity of harm shown in the images. That is, while a textual vignette might likewise have described a facial wound, participants could easily have imagined that wound to look very different when it was inflicted to a robot rather than to another human. In a similar manner, our approach may have left some ambiguity concerning the question of how well the human avatar was indeed recognised as another human being, or as simply another, albeit even more human-like, artificial entity. Again, the humanness of the human could have been made more explicit in a traditional textual vignette. Conversely, it would have been very difficult to describe the appearance of the robot precisely enough to ensure a matched impression of other socially relevant factors, such as facial shape, expression, or skin tone, in plain text. Here, future work could build upon a combination of several textual and visual vignette studies to balance the respective drawbacks and advantages of both approaches. However, given the predominance of textual vignettes in this field of work to date, we believe that the advantages of a visual approach substantially outweighed these limitations.

Manuscript received October 2019
Revised manuscript accepted August 2020

REFERENCES

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science*, *24*, 1755–1762. https://doi.org/10.1177/0956797613480507

Andrighetto, L., Baldissarri, C., Lattanzio, S., Loughnan, S., & Volpato, C. (2014). Humanitarian aid? Two forms of dehumanization and willingness to help after natural disasters. *British Journal of Social Psychology*, *53*, 573–584. https://doi.org/10.1111/bjso.12066

Avenanti, A., Sirigu, A., & Aglioti, S. M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, *20*, 1018–1022. https://doi.org/10.1016/j.cub.2010.03.071

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PLoS One*, *8*(4), e61842. https://doi.org/10.1371/journal.pone.0061842

Bastian, B., & Haslam, N. (2011). Experiencing dehumanization: Cognitive and emotional effects of everyday dehumanization. *Basic and Applied Social Psychology*, *33*, 295–303. https://doi.org/10.1080/01973533.2011.614132

Castano, E., & Giner-Sorolla, R. (2006). Not quite human: Infrahumanization in response to collective responsibility for intergroup killing. *Journal of Personality and Social Psychology*, *90*, 804–818. https://doi.org/10.1037/0022-3514.90.5.804

Čehajić, S., Brown, R., & González, R. (2009). What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations*, *12*, 715–729. https://doi.org/10.1177/1368430209347727

Costello, K., & Hodson, G. (2010). Exploring the roots of dehumanization: The role of animal-human similarity in promoting immigrant humanization. *Group Processes and Intergroup Relations*, *13*, 3–22. https://doi.org/10.1177/1368430209347725

Costello, K., & Hodson, G. (2014). Explaining dehumanization among children: The interspecies model of prejudice. *British Journal of Social Psychology*, *53*, 175–197. https://doi.org/10.1111/bjso.12016

Cuddy, A. J., Rock, M. S., & Norton, M. I. (2007). Aid in the aftermath of Hurricane Katrina: Inferences of secondary emotions and intergroup helping. *Group Processes and Intergroup Relations*, *10*, 107–118. https://doi.org/10.1177/1368430207071344

Currie, L. Q., & Wiese, E. (2019). Mind perception in a competitive human-robot interaction game. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. *63*, pp. 1957–1961). SAGE Publications.

Darling, K. (2015). *Robot ethics is about humans*. Retrieved from http://videos.theconference.se/robots-and-humans

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *362*, 679–704. https://doi.org/10.1098/rstb.2006.2004

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*, 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*, 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, *8*, 287–302. https://doi.org/10.1007/s12369-016-0338-y

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In S. S. Ge, O. Khatib, J. J. Cabibihan, R. Simmons, & M. A. Williams (Eds.), *Social robotics: Lecture notes in computer science* (Vol. *7621*, pp. 199–208). Springer.

Garreau, J. (2007, May 6). Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend. *The Washington Post Sunday*.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*, *35*, 1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003

Gilbert, D. T., Lieberman, M. D., Morewedge, C. K., & Wilson, T. D. (2004). The peculiar longevity of things not so bad. *Psychological Science*, *15*(1), 14–19. https://doi.org/10.1111/j.0963-7214.2004.01501003.x

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*, 619. https://doi.org/10.1126/science.1134475

Gray, K. (2010). Moral transformation: Good and evil turn the weak into the mighty. *Social Psychological and Personality Science*, *1*, 253–258. https://doi.org/10.1177/1948550610367686

Gray, K. (2012). The power of good intentions: Perceived benevolence soothes pain, increases pleasure, and improves taste. *Social Psychological and Personality Science*, *3*, 639–645. https://doi.org/10.1177/1948550611433470

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*, 1600–1615. https://doi.org/10.1037/a0036149

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science*, *19*(12), 1260–1262. https://doi.org/10.1111/j.1467-9280.2008.02208.x

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*, 505–520. https://doi.org/10.1037/a0013748

Gray, K., & Wegner, D. M. (2012a). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*, 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., & Wegner, D. M. (2012b). Morality takes two: Dyadic morality and mind perception. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 109–127). American Psychological Association.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124. https://doi.org/10.1080/1047840X.2012.651387

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap. *Social Cognitive and Affective Neuroscience*, *7*, 596–603. https://doi.org/10.1093/scan/nsr035

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*, 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). Guilford Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243–259. https://doi.org/10.2307/1416950

International Federation of Robotics. (2019). *IFR world robotics presentation*. Retrieved from https://ifr.org/free-downloads/

International Federation of Robotics. (2020). *Robots help to fight coronavirus worldwide*. Retrieved from https://ifr.org/ifr-press-releases/news/robots-help-to-fight-corona-virus-sars-cov-2-worldwide

Jones, A., Küster, D., Basedow, C. A., Alves-Oliveira, P., Serholt, S., Hastie, H., … Castellano, G. (2015). Empathic robotic tutors for personalised learning: A multidisciplinary approach. In A. Tapus, E. André, J. C. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics: Lecture notes in computer science* (Vol. *9388*, pp. 285–295). Springer. https://doi.org/10.1007/978-3-319-25554-5_29

Khamitov, M., Rotman, J. D., & Piazza, J. (2015). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition*, *146*, 33–47. https://doi.org/10.1016/j.cognition.2015.09.009

Kouchaki, M., Dobson, K. S. H., Waytz, A., & Kteily, N. S. (2018). The link between self-dehumanization and immoral behavior. *Psychological Science*, *29*, 1234–1246. https://doi.org/10.1177/0956797618760784

Krebs, D. (1975). Empathy and altruism. *Journal of Personality and Social Psychology*, *32*, 1132–1146. https://doi.org/10.1037/0022-3514.32.6.1134

Küster, D., Swiderska A., Gunkel, D. (2020) I saw it on YouTube! How online videos shape perceptions of mind, morality, and fears about robots. *New Media & Society*, 146144482095419. http://dx.doi.org/10.1177/1461444820954199.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, *31*(1), 184–201. https://doi.org/10.1111/j.1475-4975.2007.00147.x

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, *146*(1), 123–133. https://doi.org/10.1037/xge0000234

Nagatani, K., Kiribayashi, S., Okada, Y., Otake, K., Yoshida, K., Tadokoro, S., Nishimura, T., Yoshida, T., Koyanagi, E., Fukushima, M., & Kawatsuma, S. (2013). Emergency response to the nuclear accident at the Fukushima Daiichi nuclear power plants using mobile rescue robots: Emergency response to the Fukushima nuclear accident using rescue robots. *Journal of Field Robotics*, *30*, 44–63. https://doi.org/10.1002/rob.21439

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*, 81–103. https://doi.org/10.1086/209566

Nass, C., Moon, Y., Morkes, J., Kim, E.-Y. & Fogg, B. (1997). Computers are social actors: A review of current research. In B. Friedman (Ed.), *Moral and ethical issues in human-computer interaction* (pp. 137–162). CSLI Press.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on human robot interaction* (pp. 245–246). ACM.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, *5*, 17–34. https://doi.org/10.1007/s12369-012-0173-8

Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the Uncanny Valley. *The Journal of Neuroscience*, *39*, 6555–6570. https://doi.org/10.1523/JNEUROSCI.2956-18.2019

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*, 32–70. https://doi.org/10.1177/1088868317698288

Shen, L. (2010). On a scale of state empathy during message processing. *Western Journal of Communication*, *74*, 504–524. https://doi.org/10.1080/10570314.2010.512278

Sherman, G. D., & Haidt, J. (2011). Cuteness and disgust: The humanizing and dehumanizing effects of emotion. *Emotion Review*, *3*, 245–251. https://doi.org/10.1177/1754073911402396

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*, 15924. https://doi.org/10.1038/srep15924

Swiderska, A., & Küster, D. (2018). Avatars in pain: Visible harm enhances mind perception in humans and robots. *Perception*, *47*, 1139–1152. https://doi.org/10.1177/0301006618809919

Swiderska, A., & Küster, D. (2020) Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, Not Anthropomorphism. *Cognitive Science*, *44*(7), http://dx.doi.org/10.1111/cogs.12872.

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS One*, *12*, e0180952. https://doi.org/10.1371/journal.pone.0180952

Von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). "It doesn't matter what you are!": Explaining social effects of agents and avatars. *Computers in Human Behavior*, *26*(6), 1641–1650. https://doi.org/10.1016/j.chb.2010.06.012

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, *24*, 1437–1445. https://doi.org/10.1177/0956797612472343

Waytz, A., Cacioppo, J., & Epley, N. (2010a). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*, 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., & Epley, N. (2012). Social connection enables dehumanization. *Journal of Experimental Social Psychology*, *48*, 70–76. https://doi.org/10.1016/j.jesp.2011.07.012

Waytz, A., Epley, N., & Cacioppo, J. T. (2010b). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*, 58–62. https://doi.org/10.1177/0963721409359302

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010c). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*, 383–388. https://doi.org/10.1016/j.tics.2010.05.006

Wiese, E., & Weis, P. P. (2020). It matters to me if you are human-examining categorical perception in human and nonhuman agents. *International Journal of Human-Computer Studies*, *133*, 1–12. https://doi.org/10.1016/j.ijhcs.2019.08.002