Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, Not Anthropomorphism

Article *in* Cognitive Science · July 2020

DOI: 10.1111/cogs.12872

2 authors:

Aleksandra Swiderska
Jacobs University
14 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Dennis Küster
Universität Bremen
83 PUBLICATIONS **335** CITATIONS
SEE PROFILE

Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, not Anthropomorphism

Aleksandra Swiderska[1] and Dennis Küster[2]

[1]Department of Psychology, University of Warsaw

[2]Department of Computer Science, University of Bremen

Author Note

Aleksandra Swiderska http://orcid.org/0000-0001-7252-4581

Dennis Küster https://orcid.org/0000-0001-8992-5648

This research was supported by a grant Sonata (2016/23/D/HS6/02954) from the Polish National Science Centre to Aleksandra Swiderska.

Correspondence concerning this article should be addressed to Aleksandra Swiderska, Department of Psychology, University of Warsaw, ul. Stawki 5/7, 00-183 Warsaw, Poland. E-mail: aleksandra.swiderska@psych.uw.edu.pl

Keywords: agency; robots; harmfulness; dehumanization; moral typecasting

Abstract

A robot's decision to harm a person is sometimes considered to be the ultimate proof of it gaining a human-like mind. Here, we contrasted predictions about attribution of mental capacities from moral typecasting theory, with the denial of agency from dehumanization literature. Experiments 1 and 2 investigated mind perception for intentionally and accidentally harmful robotic agents based on text and image vignettes. Experiment 3 disambiguated agent intention (malevolent, benevolent), and additionally varied the type of agent (robotic, human) using short computer-generated animations. Harmful robotic agents were consistently imbued with mental states to a lower degree than benevolent agents, supporting the dehumanization account. Further results revealed that a human moral patient appeared to suffer less when depicted with a robotic agent than with another human. The findings suggest that future robots may become subject to human-like dehumanization mechanisms, which challenges the established beliefs about anthropomorphism in the domain of moral interactions.

Keywords: agency; robots; harmfulness; dehumanization; moral typecasting

1. Introduction

Robots with human-like physical features are expected to soon become an essential part of everyday life, fulfilling a wide array of social roles (Bekey, 2012; Dautenhahn, 2007). This revolutionary advancement in human-machine relations (Gunkel, 2018) poses new questions as to how robots will be perceived as social entities granted a certain level of a human-like mind and, possibly, moral status. Perhaps most importantly, robots may eventually participate in dyadic moral interactions, which have been shown to impact the attribution of mental capacities to the individuals involved (Gray, Young, & Waytz, 2012). For example, Ward, Olsen, and Wegner (2013) demonstrated that participants attributed mental capacities to minimally conscious (a comatose patient) and unconscious (a corpse) victims of an intentionally harmful human to a greater extent than in a control, i.e., not entailing harm, condition. The authors termed this effect *the harm-made mind* and associated it with the mechanism of automatic completion of a moral dyad: If there is a mind that deliberately produced a malevolent behavior, there must be another mind to receive it - even when objectively there is none, as was the case with the corpse. However, when the victim of a harmful behavior was a fully conscious person, mind attribution declined, compared to harmless behavior. That is, the victim was dehumanized (Ward et al., 2013; see also Haslam, 2006). These results suggest that the preexisting level of mind of the members of a moral dyad is key to shaping mind perception. In this context, human-like robots constitute an under-examined test case as either moral patients or agents.

From the perspective of machine ethics, moral agency of a robot comprises autonomy, intentionality, and responsibility towards another (Sullins, 2011). Since such truly agentive robots are still a matter of science fiction, research has relied on text vignettes describing robots as targets of a human agent's moral actions (Ward et al., 2013; Tanibe, Hashimoto, & Karasawa, 2017). In this paper, we examined how participants react to visually depicted

infliction of purposeful harm by a robotic agent on a human patient. To the best of our knowledge, this is the first systematic empirical work to place a robot in the position of the malevolent agent in a moral dyad. Specifically, we reversed the robot's role from patient (Ward et al., 2013; Tanibe et al., 2017; Swiderska, & Küster, 2018) to agent, to test competing predictions pertinent to mind attribution from moral typecasting and dehumanization theories.

1.1. Moral typecasting and dehumanization

Moral typecasting theory (MTT) builds on the assumption that prototypical moral interactions are dyadic in nature - they include a moral agent and a moral patient (Gray & Wegner, 2009). Agents must be able to carry out an action, and they thus require mental capacities that facilitate *doing* (e.g., planning activities; Waytz, Gray, Epley, & Wegner, 2010). Patients, in turn, experience the effects of the agents' actions, which necessitates the capacity to *feel* (Waytz et al., 2010). In their research, Gray, Gray, and Wegner (2007) have labeled these two sets of capabilities *agency* and *experience*, respectively, and demonstrated that they constitute two fundamental dimensions of mind perception. They showed that different entities are generally attributed both agency and experience, but to varying degrees. Moreover, they provided initial evidence that agency and experience are differentially ascribed to moral agents and patients (Gray et al., 2007). In other words, the perceptions of moral agents and patients are asymmetrical in that agents are construed as highly agentive, or at least more agentive than patients. According to MTT, this should be the case irrespective of the moral value of their behavior (Gray & Wegner, 2009).

MTT's hypotheses about moral agents are consistent with findings from other fields, for example research on anthropomorphism. Anthropomorphism is the process of imputing human qualities to non-humans (Waytz, Epley, & Cacioppo, 2010). The more social cues an entity conveys through its appearance and behavior, the more human-like internal states it appears to possess (e.g., Epley, Waytz, & Cacioppo, 2007). Attributions of mind should

increase when the entity engages in human-like behaviors not expected of it (Waytz et al., 2010). Here, a robot attacking a human contrasts with widely held beliefs that robots should be unable to harm human beings (Asimov, 1950; Murphy & Woods, 2009).

An alternative prediction comes from dehumanization theory. It has been found that offenders are denied characteristics linked in the literature to agency and patiency (Bastian, Denson, & Haslam, 2013), and that people feel less human having behaved immorally (Kouchaki, Dobson, Waytz, & Kteily, 2018). Additionally, malevolent agents were granted less agency than benevolent and neutral agents across six vignette-based experiments by Khamitov, Rotman, and Piazza (2016). As they demonstrated, malevolent agents were seen as less worthy of moral consideration, and this was mediated by a decrease in the ascription of agency (Khamitov et al., 2016). In consequence, explanations based on dehumanization theory challenge assumptions of MTT, which highlights a need for further investigation of mind attribution to harmful agents.

2. Experiment 1: Text vignettes

Experiment 1 aimed to test the hypotheses from MTT and dehumanization theory. We substituted the human malevolent agent (Khamitov et al., 2016) with a robot. We also explored whether describing the robot's behavior in mentalistic or mechanistic terms would impact perception of its mind. The Ethics Committee at the Department of Psychology, University of Warsaw, approved this and two following experiments.

2.1. Method

One hundred and twenty-five participants[1] (76 women; M_{age} = 33.56 years, *SD* = 11.06) were randomly assigned to read one of four vignettes. The vignettes discussed the

[1] An a-priori power analysis with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that 128 participants (32 per condition) would be sufficient to detect a medium-sized main effect (Cohen's f = .25, α = .05) of our primary hypothesis on intentional harm, an additional effect of stance, and an interaction effect in a 2

advancement of social robots and their potential to serve as caregivers. A fictional scenario followed, wherein an aged man trips while walking, and a caregiver robot either attacks him (intentional harm), or fails to provide support (accidental harm). To examine whether mentalistic concepts in the vignettes elicit a more intentional stance in observers (Marchesi et al. 2019), we created two variants of each vignette that described either the robot's actions (mechanistic stance) or both its internal states and actions (mentalistic stance).

Following Gray and colleagues (2007), as a complete measure of mind perception we assessed the robot's *experience* (7 items: having personality, experiencing desire, feelings, emotions, pleasure, hunger, and fear; α = .92) and *agency* (7 items: planning and controlling actions, remembering events, understanding others, understanding right from wrong, influencing situations, communicating; α = .82; Gray et al., 2007; Ward et al., 2013). Additionally, we assessed *consciousness* (2 items: being conscious of oneself, being conscious of the people and world around; α = .62) to explicitly gauge the endorsement of the robot's mental status, in line with Ward and colleagues (2013). Participants evaluated the extent to which each of the capabilities could be attributed to the robot on Likert scales from *1 = strongly disagree* to *7 = strongly agree.* Moreover, participants rated how *morally right or wrong* (*1 = extremely wrong*, *7 = extremely right*) the robot's action was, and its perceived *intentionality* (*1 = definitely accidental, 7 = definitely intentional*).

2.2. Results and discussion

A 2 (Intention) x 2 (Stance)[2] between-subjects multivariate analysis of variance (MANOVA) on three mind perception measures (with Bonferroni correction for multiple

---

x 2 ANOVA with 80% power. We fell slightly short of the recruitment target, yet considered this difference within the margin of error. Our power analysis reflected conventional parameters for medium-sized effects, where no exact empirical data could be obtained from prior work.

[2] For mind perception, the multivariate main effect of stance did not reach significance, $F(3, 119) = 2.12$, $p = .101$, $\eta_p^2 = .05$. Similarly, for moral evaluation and intentionality the main effect of stance was not significant,

comparisons) showed a significant multivariate main effect of intention, $F(3, 119) = 4.66$, $p = .004$, $\eta_p^2 = .11$. The following univariate ANOVAs were significant for agency, $F(1, 121) = 6.73$, $p = .011$, $\eta_p^2 = .05$, and consciousness, $F(1, 121) = 4.74$, $p = .031$, $\eta_p^2 = .04$, but not for experience, $F(1, 121) = .74$, $p = .392$, $\eta_p^2 = .01$. In line with a dehumanization view, intentionally harmful agents were perceived to possess less agency and consciousness than accidentally harmful agents (Fig. 1).

An ANOVA on the moral evaluation, serving as a manipulation check, revealed a significant main effect of intention, $F(1, 121) = 66.07$, $p < .001$, $\eta_p^2 = .35$. Malevolent behavior was perceived as more morally wrong than accidental harm ($M_{intentional} = 2.42$, $SD = 1.34$ vs. $M_{accidental} = 4.38$, $SD = 1.34$). The ANOVA on perceived intentionality yielded a non-significant main effect of intention, $F(1, 121) = 1.47$, $p = .228$, $\eta_p^2 = .01$. Thus, participants were hesitant to evaluate either behavior as purposeful ($M_{intentional} = 2.86$, $SD = 1.88$ vs. $M_{accidental} = 2.40$, $SD = 2.05$).

----------Insert Figure 1 about here----------

Fig. 1. Means for the three mind perception measures in three experiments. Error bars denote +/- 1 *SE*.

3. Experiment 2: Image vignettes

In experiment 1, intentional harm resulted in reduced mind attribution compared to accidental harm. However, the manipulation did not include a benevolent agent to help determine if malevolent behavior decreases mind perception relative to both accidental harm despite benevolent intentions, and accidental harm without intention cues. Furthermore, we

---

respectively $F(1, 121) = .32$, $p = .573$, $\eta_p^2 = .00$ and $F(1, 121) = .18$, $p = .673$, $\eta_p^2 = .00$. See Supplement for the results of evaluations of the perceived intentional stance in experiments 2 and 3.

did not have control over what participants imagined while reading. To address these issues, experiment 2 employed visual vignettes.

3.1. Method

One hundred and twelve participants[3] (65 women, 46 men, 1 other; M_{age} = 33.77 years, *SD* = 12.09) completed the survey. We designed three visuals, based on texts from experiment 1. They depicted a human-like robot and an elderly man, who was about to fall. The robot attacked the man with a club (malevolent), stretched its arms out towards him (benevolent), or showed no intent (unresponsive; Fig. 2). All images were rendered in Daz Studio (V4.10, www.daz3d.com) via the IRAY render engine (nVidia).

We used the measures and response formats from experiment 1 (experience: α = .95, agency: α = .87, consciousness: α = .75), and a scale from Khamitov et al. (2016) to estimate the agent's perceived *activity level* (6 items: vigorous, willful, tenacious, potent, active, energetic; α = .86). For exploratory purposes, we also checked how participants felt in reaction to the agent's appearance with regard to the potential *uncanny valley* effects (see Mori, MacDorman, & Kageki, 2012; 3 items: feeling uneasy, unnerved, creeped out; α = .96; adapted from Gray & Wegner, 2012).

----------Insert Figure 2 about here----------

Fig 2. Malevolent intent (left), benevolent intent (middle), and neutral control (unresponsive; right).

3.2. Results and discussion

A MANOVA with Intention (malevolent, benevolent, unresponsive) as between-subjects factor and experience, agency, and consciousness as dependent variables revealed a marginally significant multivariate main effect, $F(6, 214) = 1.98$, $p = .070$, $\eta_p^2 = .05$. The ensuing univariate ANOVAs were significant for experience, $F(2, 109) = 3.93$, $p = .022$, $\eta_p^2 =$

[3] For visual vignettes, we expected larger effects ($f = .30$). G*Power indicated that 111 participants should be sufficient to reveal significant (α = .05) effects between three experimental conditions with 80% power.

.08, but non-significant for agency, $F(2, 109) = 1.36$, $p = .262$, $\eta_p^2 = .02$, and consciousness, $F(2, 109) = 1.06$, $p = .352$, $\eta_p^2 = .02$. In line with the results of experiment 1, malevolent robots were again denied certain mental states. Specifically, participants attributed more capacity for experience to the benevolent robot ($M = 3.39$, $SD = 1.79$) than to the malevolent ($M = 2.67$, $SD = 1.62$; $p = .048$), or the unresponsive robot ($M = 2.42$, $SD = 1.21$; $p = .008$; Fig. 1).

An ANOVA on activity level showed a significant main effect of intention, $F(2, 109) = 4.57$, $p = .012$, $\eta_p^2 = .08$, whereby the malevolent robot was more active ($M = 4.87$, $SD = 1.43$) than the benevolent ($M = 4.13$, $SD = 1.21$; $p = .014$) and the unresponsive ($M = 4.05$, $SD = 1.21$; $p = .007$) robots. This finding appears likely due to the presence of a weapon (e.g., Pickel, 1999) in the respective vignette. That is, increased perceived activity of the robot equipped with a bat may explain the lack of focus on its other (i.e., internal) characteristics.

Moral evaluations of the robot's behavior differed significantly depending on intention, $F(2, 108) = 56.83$, $p < .001$, $\eta_p^2 = .51$. Malevolent behavior was perceived as substantially more morally wrong ($M = 2.08$, $SD = 1.23$) than the two other types of behavior ($M_{benevolent} = 5.21$, $SD = 1.36$; $p < .001$ and $M_{unresponsive} = 3.59$, $SD = 1.19$; $p < .001$). Benevolent behavior was perceived as more morally right than neutral behavior ($p < .001$). An ANOVA with intentionality revealed a significant main effect of intention, $F(2, 109) = 6.21$, $p = .003$, $\eta_p^2 = .10$. The behavior of the malevolent robot was rated as more intentional ($M = 5.89$, $SD = 1.24$) than the benevolent ($M = 4.71$, $SD = 1.81$; $p = .001$) or the unresponsive ($M = 4.92$, $SD = 1.52$; $p = .008$) robots.

Finally, for the uncanny valley, a significant main effect of intention was obtained, $F(1, 125) = 36.67$, $p < .001$, $\eta_p^2 = .23$ (but not of agent, $F(1, 125) = .28$, $p = .598$, $\eta_p^2 = .00$), with a marginally significant interaction, $F(1, 125) = 2.98$, $p = .087$, $\eta_p^2 = .02$. Pairwise comparisons showed that malevolent behavior rendered the agents more uncanny than

benevolent behavior (robot: $M = 4.18$, $SD = 1.82$ vs. $M = 2.84$, $SD = 1.61$, $p = .003$; human: $M = 4.87$, $SD = 1.72$ vs. $M = 2.47$, $SD = 1.84$; $p < .001$).

4. Experiment 3: Video vignettes

Although the visual vignettes in experiment 2 provided greater control over how participants envisioned the scenarios, they lacked a clear temporal context. We included the baseball bat to supply information on malevolent intention. The weapon had an impact on mind perception, as suggested by the increased activity granted to the robot. We therefore conducted a third experiment using animated vignettes as this approach did not require tools to signify intentions.

4.1. Method

One hundred and twenty-nine participants[4] (75 women; M_{age} = 34.22 years, SD = 12.38) viewed one of four randomly selected animations that showed an Agent (robotic, human) exhibiting either malevolent or benevolent Intentions. In the malevolent conditions, the agent kicked an elderly man in the back, who then fell. In the benevolent conditions, the agent ran towards the man as he started to fall, and kneeled next to him. The characters were animated in Daz Studio (V4.10) and post-processed in Adobe AfterEffects CC (V16.1).

Like in the two preceding experiments, we collected ratings of the agent's perceived experience ($\alpha = .96$), agency ($\alpha = .84$), and consciousness ($\alpha = .73$; Ward et al., 2013), moral evaluation of the agent's behavior, and perceived intentionality. Additionally, perceptions of the moral patient were assessed with respect to how likely the elderly man was to *suffer* in consequence (Loughnan, Pina, Vasquez, & Puvia, 2013) and how much state *empathy* he evoked (7 items, e.g., understanding of feelings; $\alpha = .89$; Davis, 1980). The scale of *empathic*

[4] Experiment 3 was again based on a 2 x 2 ANOVA design, yielding an estimated 128 participants (32 per group) to achieve 80% power for a medium effect size (Cohen's $f = 0.25$, $\alpha = 0.05$) for both main and interaction effects.

concern (7 items, e.g., *I am often quite touched by things that I see happen*; α = .84), from the Interpersonal Reactivity Index (Davis, 1983), was included to control for individual differences[5].

4.2. Results and discussion

We conducted a MANOVA with Agent and Intention as between-subjects factors, and the three mind perception scales as dependent variables. The multivariate main effect of agent was significant, $F(3, 123) = 44.45$, $p < .001$, $\eta_p^2 = .52$, with the univariate effects significant for experience, $F(1, 125) = 108.07$, $p < .001$, $\eta_p^2 = .46$ ($M_{human} = 5.32$, $SD = 1.32$ vs. $M_{robot} = 2.91$, $SD = 1.38$), but not significant for agency, $F(1, 125) = .73$, $p = .395$, $\eta_p^2 = .01$ ($M_{human} = 4.50$, $SD = 1.32$ vs. $M_{robot} = 4.35$, $SD = 1.25$), and consciousness, $F(1, 125) = .12$, $p = .731$, $\eta_p^2 = .00$ ($M_{human} = 4.49$, $SD = 1.74$ vs. $M_{robot} = 4.41$, $SD = 1.81$). The multivariate main effect of intention was significant, $F(3, 123) = 8.21$, $p < .001$, $\eta_p^2 = .17$, including experience, $F(1, 125) = 7.23$, $p = .008$, $\eta_p^2 = .06$, agency, $F(1, 125) = 24.29$, $p < .001$, $\eta_p^2 = .16$, and consciousness, $F(1, 125) = 17.91$, $p < .001$, $\eta_p^2 = .13$ (Fig. 1). The human agent was attributed more capacity for experience than the robot, validating our manipulation of human-likeness. Most importantly, the manipulation of intention seemed to have a more pronounced impact on mind perception than the type of agent that embodied it. Specifically, this manipulation influenced the perceptions of all three sets of capabilities, that is, malevolent intention was associated with a significantly decreased attribution of mind compared to benevolent intention. The results were thus again consistent with dehumanization theory. The interaction was not significant, $F(3, 123) = .46$, $p = .714$, $\eta_p^2 = .01$.

[5] Participants' empathic concern scores were included as a covariate in all ensuing analyses. No significant effects emerged for mind perception, moral evaluation, and perceived intent. A significant main effect for suffering ($p < .001$) did not change the pattern of the reported results.

For moral evaluation, there was a significant main effect of intention, $F(1, 125) = 134.99$, $p < .001$, $\eta_p^2 = .52$. Malevolent behavior was seen as more wrong than benevolent behavior ($M = 1.61$, $SD = .99$ vs. $M = 4.89$, $SD = 2.02$). The main effect of agent was not significant, $F(1, 125) = .64$, $p = .427$, $\eta_p^2 = .01$. A similar pattern emerged for intentionality, showing a significant effect of intention, $F(1, 125) = 32.24$, $p < .001$, $\eta_p^2 = .21$, with malevolent behavior perceived as more intentional than benevolent behavior ($M = 6.30$, $SD = 1.38$ vs. $M = 4.52$, $SD = 2.06$), and a not significant effect of agent, $F(1, 125) = .08$, $p = .785$, $\eta_p^2 = .00$.

In exploratory analyses of the moral patient's perceptions, main effects of agent and intention were significant for suffering, $F(1, 125) = 5.96$, $p = .016$, $\eta_p^2 = .05$ and $F(1, 125) = 41.48$, $p < .001$, $\eta_p^2 = .25$. The elderly man seemed to suffer more when he was shown together with a human than with a robot ($M = 6.09$, $SD = 1.00$ vs. $M = 5.68$, $SD = 1.07$) and, unsurprisingly, as a result of harm rather than benevolence ($M = 6.41$, $SD = .71$ vs. $M = 5.38$, $SD = 1.06$). For state empathy, both main effects were not significant, $F(1, 124) = 1.57$, $p = .213$, $\eta_p^2 = .01$ and $F(1, 124) = .06$, $p = .812$, $\eta_p^2 = .00$, when controlling for individual levels of trait empathic concern. These results suggest that identical physical harm is seen to cause more suffering when it is inflicted by another human being than by a robot.

5. Conclusions

In three experiments, we employed textual, and static and dynamic visual vignettes to investigate mind attribution to robotic moral agents. The results consistently favored dehumanization-based explanations over moral typecasting and the process of anthropomorphism in that malevolent robots were evaluated to possess less mental capacities than their benevolent (and neutral) counterparts. This extends previous findings by Khamitov and colleagues (2016) pertinent to harmful human agents to non-human entities. We further went beyond their descriptive approach by using images and videos that provided more

powerful stimuli and helped reduce ambiguity with regard to how participants understood them. However, our work is still limited when it comes to the generalizability of results to other experimental materials. Future research should examine the effects of different visual features of characters on mind attribution in visual vignettes (e.g., gender; Küster, Krumhuber, & Hess, 2018).

The expected denial of agency occurred for robots exhibiting malevolent intentions described in text in experiment 1 and shown in a video in experiment 3. The effect of intention did not reach significance for the robots displayed in images in experiment 2, although the malevolent robot appeared slightly deprived of experience. Towards an explanation of these discrepant outcomes, we consider the bat held by the robot to have driven participants' attention away from the robot's internal capabilities. On the other hand, the bat turned out to be a salient cue to the perception of the attacking robot's activity level, which was used as an alternative measure of agency in Khamitov et al., 2016, but the items clearly refer to what can actually be observed, rather than inferred. Moreover, the benevolent robot in experiment 2 makes a facial expression that may have been interpreted as fear (for the falling man). This may have reinforced the perceptions of the capacity to feel emotions, an important component of the experience dimension. Tools or facial expressions were not present in the written or video vignettes.

Malevolent and benevolent actions were evaluated as respectively wrong and right when carried out by a robot, which demonstrated that the moral value assigned to these behaviors was largely parallel to how human behaviors would be typically judged. Nevertheless, our exploratory analysis of ratings of the human moral patient suggested that he was perceived to suffer less as a consequence of harm when the robot was the moral agent. This may indicate that similar damage caused by a human being is viewed as more severe. Future research is needed to examine this effect, but we speculate that it represents an

inversion of the original harm-made mind effect (Ward et al., 2013). That is, harm inflicted by a comparatively lesser mind could appear less serious because it does not convey the same moral weight and implications for the moral patient in the interaction.

The perception of intentionality varied systematically across the three experiments, with malevolent behaviors depicted in visuals rated as more intentional than benevolent behaviors, whereas both were deemed not very intentional in textual vignettes. Although intentionality is considered to be an aspect of agency (Khamitov et al., 2016), here its perception seemed separate from related mental capacities. This may be particularly interesting when considered again in the context of the perceptions of the moral patient. As shown by previous work, intentionality is a necessary prerequisite for harm to lead to increased attributions of mind to the moral patient in a moral dyad (Ward et al., 2013). Nonetheless, our findings point to the possibility that an intentional non-human agent will not have the same degree of influence. Future work could test whether lesser perceived suffering extends to decreased attributions of the capacity for pain or experience in general.

References

Asimov, I. (1950). *I, Robot*. Greenwich, Conn: Fawcett Publications.

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

Bekey, G. A. (2012). Current trends in robotics: Technology and ethics. In L. Patrick, A. Keith, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 17-34). Cambridge, MA: MIT Press.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362* (1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10.* Retrieved from www.uv.es/~friasnav/Davis_1980.pdf

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44* (1), 113-126. http://dx.doi.org/10.1037/0022-3514.44.1.113

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114* (4), 864-886. http://dx.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* (2), 175-191. https://doi.org/10.3758/BF03193146

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315* (5812), 619-619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96* (3), 505-520. http://dx.doi.org/10.1037/a0013748

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125* (1), 125-130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23* (2), 101-124. https://doi.org/10.1080/1047840X.2012.651387

Gunkel, D. J. (2018). The relational turn: Third wave HCI and phenomenology. In M. Filimowicz & V. Tzankova (Eds.), *New Directions in Third Wave Human-Computer Interaction: Volume 1 - Technologies. Human–Computer Interaction Series* (pp. 11-24). Cham: Springer.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10* (3), 252-264. https://doi.org/10.1207/s15327957pspr1003_4

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition, 146* (2016), 33-47. http://dx.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* (8), 1234–1246. https://doi.org/10.1177/0956797618760784

Küster, D., Krumhuber, E. G., & Hess, U. (2019). You are what you wear: Unless you moved - Effects of attire and posture on person perception. *Journal of Nonverbal Behavior, 43* (1), 23-38. https://doi.org/10.1007/s10919-018-0286-3

Loughnan, S., Pina, A., Vasquez, E. A., & Puvia, E. (2013). Sexual objectification increases rape victim blame and decreases perceived suffering. *Psychology of Women Quarterly, 37* (4), 455-461. https://doi.org/10.1177/0361684313485718

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*, 450. https://doi.org/10.3389/fpsyg.2019.00450

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19* (2), 98-100. https://doi.org/10.1109/MRA.2012.2192811

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE intelligent systems, 24* (4), 14-20. https://doi.org/10.1109/MIS.2009.69

Pickel, K. L. (1999). The influence of context on the "weapon focus" effect. *Law and Human Behavior, 23* (3), 299-311. https://doi.org/10.1023/A:1022356431375

Sullins, J. P. (2011). When is a robot a moral agent? In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 151–161). New York, NY: Cambridge University Press.

Swiderska, A., & Küster, D. (2018). Avatars in pain: Visible harm enhances mind perception in humans and robots. *Perception, 47* (12), 1139–1152. https://doi.org/10.1177/0301006618809919

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLOS ONE, 12* (7), e0180952. https://doi.org/10.1371/journal.pone.0180952

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* (8), 1437-1445. https://doi.org/10.1177/0956797612472343

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science, 19* (1), 58-62. https://doi.org/10.1177/0963721409359302

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14* (8), 383-388. https://doi.org/10.1016/j.tics.2010.05.006

Fig. 1

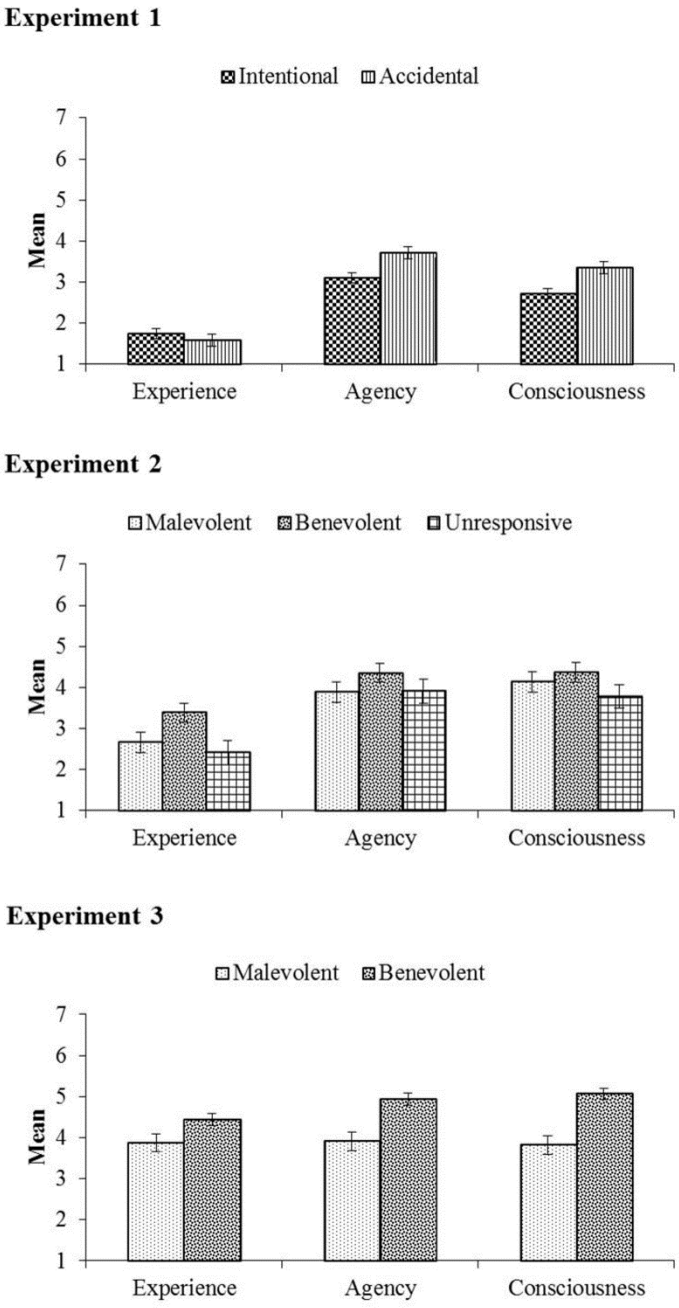
Experiment 1
Intentional
Accidental
Mean
7
6
5
4
3
2
1
Experience
Agency
Consciousness
Experiment 2
Malevolent
Benevolent
Unresponsive
Mean
7
6
5
4
3
2
1
Experience
Agency
Consciousness
Experiment 3
Malevolent
Benevolent
Mean
7
6
5
4
3
2
1
Experience
Agency
Consciousness

Fig. 2