Check for updates

Article

PERCEPTION

Perception
1–14
The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0301006618809919
journals.sagepub.com/home/pec
SAGE

Avatars in Pain: Visible Harm Enhances Mind Perception in Humans and Robots

Aleksandra Swiderska
University of Warsaw, Poland

Dennis Küster
University of Bremen, Germany

Abstract

Previous research has shown that when people read vignettes about the infliction of harm upon an entity appearing to have no more than a liminal mind, their attributions of mind to that entity increased. Currently, we investigated if the presence of a facial wound enhanced the perception of mental capacities (experience and agency) in response to images of robotic and human-like avatars, compared with unharmed avatars. The results revealed that harmed versions of both robotic and human-like avatars were imbued with mind to a higher degree, irrespective of the baseline level of mind attributed to their unharmed counterparts. Perceptions of capacity for pain mediated attributions of experience, while both pain and empathy mediated attributions of abilities linked to agency. The findings suggest that harm, even when it appears to have been inflicted unintentionally, may augment mind perception for robotic as well as for nearly human entities, at least as long as it is perceived to elicit pain.

Keywords

mind perception, pain, empathy, harm, robots, anthropomorphism

Date Received: 30 January 2018; accepted: 5 October 2018

Humans are naturally inclined to attribute a variety of human characteristics to an array of nonhuman and nonliving entities, such as pets, gods, and natural forces, to name but a few (e.g., Barrett & Keil, 1996; Gosling, Kwan, & John, 2003). This process is called anthropomorphism. Reasoning about these entities *as if* they had human-like internal states that

Corresponding author:
Aleksandra Swiderska, Department of Psychology, University of Warsaw, ul. Stawki 5/7, 00-183 Warsaw, Poland.
Email: aleksandra.swiderska@psych.uw.edu.pl

underlie their observable behaviors is parallel to how we make inferences about fellow humans (e.g., Epley, Waytz, & Cacioppo, 2007). Such a mechanism serves to make the nonhuman entities more understandable, predictable, and generally easier to relate to, which in turn contributes to a sense of control over our surroundings (Waytz, Epley, & Cacioppo, 2010). One of the essential qualities imputed with far-reaching consequences to nonhumans is a human-like mind.

While mind perception has long been assumed to occur along a single dimension (i.e., entities possess a mind to varying extents; Gray, Gray, & Wegner, 2007), bottom-up approaches have demonstrated that people spontaneously perceive differences in mental capabilities in terms of at least two main dimensions: *Experience* and *agency* (Gray et al., 2007; see also Weisman, Dweck, & Markman, 2017). Experience entails being able to experience feelings and emotions, for instance pain, pleasure, or hunger; agency involves the possession of mental capabilities to understand oneself and others, communicate, and prepare and execute purposeful actions (Gray et al., 2007). This distinction has informed the study of a broad range of topics (Weisman et al., 2017), from objectification of women (Gray, Knobe, Sheskin, Bloom, & Feldman Barrett, 2011) and dehumanization (Haslam, 2006), to discomfort in response to robots with elevated levels of perceived experience (Gray & Wegner, 2012).

Perceived capacity for experience, and in particular the capacity to feel pain, are theorized to be linked to casting an entity in the role of a moral patient, that is, a potential receiver of moral actions (Gray, Waytz, & Young; 2012; Sytsma & Machery, 2012). Moral agency, in turn, refers to an ability to carry these actions out (Gray & Wegner, 2009). A dyad composed of a moral patient and a moral agent constitutes a fundamental template for understanding moral interactions (moral typecasting theory; Gray & Wegner, 2009). Previous research has demonstrated that people tend to identify moral dyads even when objectively there are none, for example, in the case of *harmless wrongs*, whereby seemingly innocuous actions still engender perceived victims (Gray, Schein, & Ward, 2014).

Perceptions of minds of interacting dyads are highly context dependent. In a series of experiments, Ward, Olsen, and Wegner (2013) showed that infliction of physical harm by an intentional agent resulted in increased attributions of mind to a victim whose level of mind was relatively low (a patient in a vegetative state), or nonexistent (a corpse). This effect has been termed the *harm-made mind* and explained as an outcome of the victim's participation in a moral interaction (Ward et al., 2013). As Ward et al. (2013) argued, perceivers exhibited an automatic disposition to conclude that, if someone takes part in a moral interaction, then that someone must have an independent mind. By the same token, if there is a moral agent, then there must automatically be a moral patient so that the dyad was complete (Gray et al., 2014), and therefore, according to the authors, "morality creates minds" (Gray, Young, & Waytz, 2012; Ward et al., 2013, p. 1437).

To date, few studies on the harm-made mind have explored mind perception for nonhuman entities, but evidence suggests that the effect applies to them as well. Ward et al. (2013) observed that when harm was inflicted upon a "highly complex" robot (p. 1442), it was attributed mental capacities to a greater degree. The robot's human likeness resided in its abilities to express emotions and respond to emotional displays of others, as detailed by the vignette used as a stimulus in the study. Another recent vignette experiment demonstrated that attributions of mind were facilitated in the context of a positive moral interaction, when participants imagined a situation of actively helping a robot (Tanibe, Hashimoto, & Karasawa, 2017).

The studies by Ward et al. (2013) and Tanibe et al. (2017) indicate that robots may be seen as relevant actors in complex sociomoral scenarios. Anthropomorphic machines, that is, robots that look human-like, are no longer limited to science fiction literature, as shown

by detailed prototypes such as Geminoid (Ishiguru) and Sophia (Hanson). Less capable robots with a certain extent of human-likeness have been mass produced, for example, Nao (Aldebaran) and Pepper (Softbank Robotics). Together, these developments have resulted in a much greater exposure of human-like robots to the public, to a point where there is little controversy about whether highly anthropomorphic robots will eventually become part of everyday life (Dautenhahn, 2007). In consequence, robots are of increasing interest for psychological research on mind perception. However, to the best of our knowledge, no studies on the harm-made mind exposed participants to an actual robot or a photograph of a robotic entity.

Overview of the Present Research

Anthropomorphism is evoked in the presence of even minimal social cues, and the more social cues an entity provides, the more human-like it appears (Gong, 2008). The strongest cue to anthropomorphizing is human-like appearance and especially a face, the most social stimulus of all (Fiske & Taylor, 2013; Looser & Wheatley, 2010). In contrast, the use of textual materials to study social processes can be criticized as it implicitly assumes that participating in social reality is comparable to reading stories (Parkinson & Manstead, 1993). In this study, we aimed to validate and extend research on the harm-made mind effect by translating vignette stories describing harmed patients into pictures. To visually represent physical harm, we added a moderately severe wound (resembling a burn) to the faces of a robotic and a photorealistic human avatar. We included robotic and human avatars to study the impact of harm on entities expected to differ in their perceived level of mind by creating a more robotic version of the original human face.

Our second aim was to better understand the processes underlying the harm-made mind. As demonstrated by Ward et al. (2013), mind attributions were mediated by the patients' perceived capacity to feel pain. We therefore wanted to test whether perceived pain would act as a mediator for mind attributions in response to visual evidence for pain, even when the facial expression of the harmed entity is kept neutral. Furthermore, harm decreased mind attribution to a fully conscious person (Ward et al., 2013, Experiment 5). However, it is yet unknown how conscious an entity needs to be to be regarded as *fully conscious*, and the available literature on the harm-made mind has left much ground unexplored between the extremes of liminal minds and *fully conscious* minds. We thus aimed to examine the impact of visible evidence of pain on mind perception for an entity that is only relatively high in consciousness (human-like avatar) and one low to moderate in consciousness (robotic avatar).

The perception of pain is one of the main elicitors of empathy (e.g., Gallese, 2003). Empathy refers to quite a broad range of responses toward others, all of which entail the inherent ability to share another's feelings or take another's perspective in a given situation (see Batson, 2008). This then leads to a better understanding of their mental states (and, by extension, their actions; Singer & Lamm, 2009). Conversely, a lack of empathy is believed to ensue when mental states, and humanness in general, are denied to others, that is, when they are dehumanized. In particular, the mechanistic form of dehumanization has been associated with a failure to empathize (Haslam, 2006), as well as with the experience of dehumanization (Bastian & Haslam, 2011). Mechanistic dehumanization entails perceiving others as cold, inert, and *automaton-like* (Haslam, 2006). As Haslam and others have argued, empathy is often proposed to be of key importance for overcoming dehumanization (Halpern & Weinstein, 2004; Haslam, 2006). Therefore, empathy toward a robotic entity should be linked to the extent to which that entity is granted (or denied) mental states.

In line with this reasoning, earlier research has found that people indeed reacted empathically to an abused robotic toy (Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013). We therefore aimed to test if the perceivers' empathic response may fulfill a similar mediating role as Ward et al. (2013) observed for perceived pain.

Method

Participants

Two hundred seventeen participants (135 women; $Mdn_{age} = 19$ years[1]) completed the study online,[2] via a U.S.-based academic platform devoted to psychological research (http://psych.hanover.edu/research/exponnet.html). The experiment was conducted in English. All subjects took part on a volunteer basis and without compensation.

An independent sample of 53 participants (40 women; $M_{age} = 28.09$ years, $SD = 12.60$) completed a brief follow-up study. All of them were volunteers who accessed the link to the study posted online on the same platform as the main experiment. The study was conducted in English.

Materials

We first created images of faces of a photorealistic human avatar and a robotic avatar. These two faces were then modified using Poser Pro 2014 (Smith Micro) to create the faces' harmed analogues by adding the facial texture of a moderate burn to their right side (Figure 1). To avoid the risk of creating highly *uncanny* robotic avatars due to a perceptual mismatch between facial features (Kätsyri, Förger, Mäkäräinen, & Takala, 2015), our robotization manipulation aimed at an overall robotic appearance (glossy skin, eyes, lips, ears, and hair). To further control for possible confounds such as perceived ethnicity (Bartneck et al., 2018), skin smoothness (Tsankova & Kappas, 2016), or facial appearance (Balas & Pacella, 2017), the same color map was used for the diffuse channel for rendering the robotic skin surface. Reflective materials with high gloss are generally perceived as less natural than non-reflective materials (Karana, 2012), and gleaming or metallic surfaces have been found to be typical for robots (Riek & Howard, 2014). Therefore, the specular and reflective channels were adapted to produce a glossy plastic look of the robotic avatar's skin. All images measured 960×720 pixels and were displayed individually on a white background.

Procedure and Design

In the main experiment, the participants' task was to evaluate the degree to which mental capacities (experience, agency, consciousness, and pain) could be attributed to the faces and the extent to which the presented avatars elicited empathic reactions. Every page of the survey consisted of the respective face displayed above a 7-point, Likert-type response scale (1 = *strongly disagree* to 7 = *strongly agree*). The survey was delivered via EFS Survey (Version 9.0, QuestBack AG, Germany). The experiment followed a 2 (Harm: harmed vs. control) × 2 (Robotization: human vs. robotic) between-subjects factorial design.

As, compared with Ward et al. (2013), our stimuli provided limited contextual cues, in the follow-up study participants were presented with two harmed faces (between subjects) and asked about their impressions regarding whom the wound was caused by (1 = *definitely self-inflicted*, 7 = *definitely inflicted by another*) and whether it was caused intentionally (1 = *definitely unintentionally*, 7 = *definitely intentionally*).

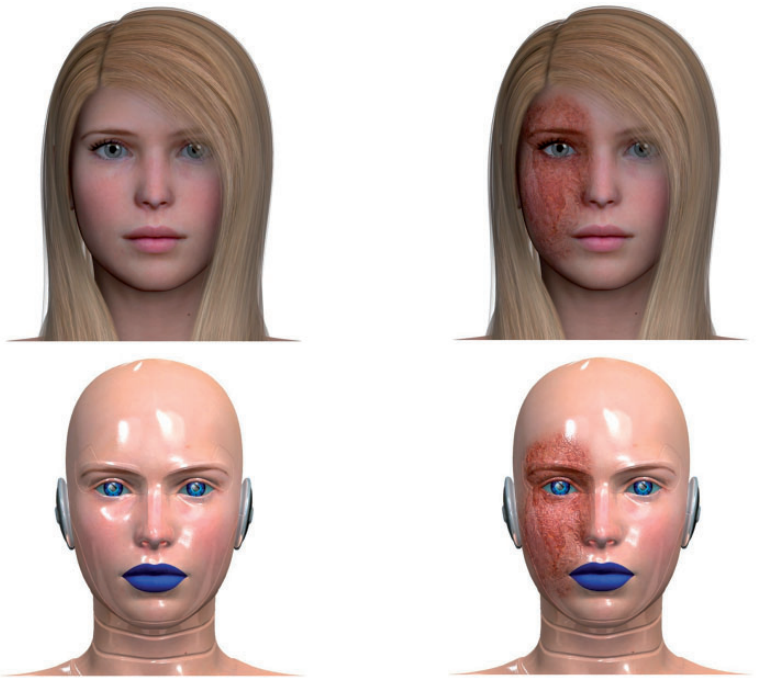

Figure 1. Unharmed (left) and harmed (right) versions of the photorealistic human avatar and the robotic avatar used as stimuli.

Dependent measures. Following Ward et al. (2013), in the main experiment, we assessed perceived *experience* (7 items; Cronbach's $\alpha = .96$), *agency* (7 items; $\alpha = .90$), and *consciousness* (2 items; $\alpha = .79$) as dimensions of mind perception. We also assessed the capacity to feel *pain* as a primary mediator of mind attributions as well as the degree to which the viewers reacted empathically to the entity or its perceived pain as another potential mediator in mediation analyses. For this purpose, we included a scale for participants' self-reported *empathy* toward the characters (7 items: perspective taking, understanding of feelings, compassion, giving comfort, sharing of feelings, being moved, and feeling empathy; $\alpha = .94$; adapted from Davis, 1980). The sequence of items of the included scales was fully randomized. Finally, we asked about the perceived *attractiveness* of the faces as a control variable.

Results

A multivariate analysis of variance with harm (unharmed, harmed) and robotization (human, robotic) as between-subjects factors was conducted on six dependent variables (pain, experience, agency, consciousness, empathy, and attractiveness). The multivariate main effects were significant for both harm, $F(6, 208) = 4.90$, $p < .001$, $\eta_p^2 = .12$, and robotization, $F(6, 208) = 39.21$, $p < .001$, $\eta_p^2 = .53$. The interaction between harm and robotization did not reach significance, $F(6, 208) = 0.82$, $p = .557$, $\eta_p^2 = .02$.

Subsequent univariate tests (with Bonferroni correction for multiple comparisons) showed significant main effects of harm for experience, $F(1, 213) = 19.55$, $p < .001$, $\eta_p^2 = .08$, agency, $F(1, 213) = 18.47$, $p < .001$, $\eta_p^2 = .08$, consciousness, $F(1, 213) = 19.98$, $p < .001$, $\eta_p^2 = .09$, capacity to feel pain, $F(1, 213) = 15.41$, $p < .001$, $\eta_p^2 = .07$, and empathy,

Table 1. Descriptive Statistics for the Harmed and Unharmed Versions of Human and Robotic Faces.

	Harmed				Unharmed			
	Human avatar		Robotic avatar		Human avatar		Robotic avatar	
Dependent variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pain	6.56	0.82	4.26	2.25	5.84	1.27	3.15	2.05
Experience	6.29	0.84	3.96	1.75	5.67	1.17	2.90	1.59
Agency	5.59	0.91	4.53	1.35	4.99	1.00	3.65	1.56
Consciousness	5.78	1.09	4.24	1.62	4.86	1.46	3.26	1.82
Empathy	5.72	0.93	4.39	1.64	4.92	1.07	3.46	1.67
Attractiveness	4.97	1.60	2.95	1.94	5.32	1.55	2.43	1.81

Note. M = mean; SD = standard deviation.

$F(1, 213) = 20.54$, $p < .001$, $\eta_p^2 = .09$. In line with our predictions, the presence of a facial wound increased attributions of all of the aforementioned mental capacities. It also elicited more empathy toward the avatars (Table 1). No significant differences were found for attractiveness, $F(1, 213) = 0.14$, $p = .714$, $\eta_p^2 = .00$. The lack of a significant difference for attractiveness between the harmed and unharmed versions of avatars suggests that harm increased attribution of mental capacities irrespective of facial attractiveness.

The main effect of robotization was significant for all of the dependent variables: experience, $F(1, 213) = 181.74$, $p < .001$, $\eta_p^2 = .46$, agency, $F(1, 213) = 48.29$, $p < .001$, $\eta_p^2 = .19$, consciousness, $F(1, 213) = 55.07$, $p < .001$, $\eta_p^2 = .21$, pain, $F(1, 213) = 114.50$, $p < .001$, $\eta_p^2 = .35$, empathy, $F(1, 213) = 53.31$, $p < .001$, $\eta_p^2 = .20$, as well as attractiveness, $F(1, 213) = 104.92$, $p < .001$, $\eta_p^2 = .33$. In sum, participants were overall more inclined to perceive the human avatar as possessing a mind, they felt more empathy toward it, and they found it to be more attractive than the robotic avatar (Table 1). Neither the higher baseline level of perceived mental capabilities, nor the difference in attractiveness appeared to modulate the harm-made mind effect. This suggests that, at least for the cases investigated in this study, the evident differences in attractiveness were likely not a driving factor of the results.

Furthermore, a multivariate analysis of variance with robotization (human and robotic) was conducted on two questions asked in the follow-up study, but it yielded no significant main effect, $F(2, 50) = .19$, $p = .825$, $\eta_p^2 = .01$. The final analyses were thus one-sample t tests conducted to evaluate whether the mean responses to questions about who inflicted the wound, and whether it was done intentionally or not, were different from the midpoint of the response scales (4 on 7-point scales). Regarding the first question, for the human avatar, the sample mean of 5.46 ($SD = 1.42$) was significantly higher than the midpoint, $t(25) = 5.25$, $p < .001$, 95% confidence interval (CI) [.89, 2.04]. Similar results were obtained for the robotic avatar, whereby the mean of 5.26 ($SD = 1.40$) also differed significantly from the scale's midpoint, $t(26) = 4.66$, $p < .001$, 95% CI [.70, 1.81]. These findings suggest that, in both cases, the facial wound was likely perceived to be inflicted by another actor and not by the avatars themselves. Considering intentionality, for the human avatar, the mean of 3.58 ($SD = 1.90$) was not significantly different from the midpoint, $t(25) = -1.14$, $p = .267$, 95% CI [−1.19, .34], but for the robotic avatar, the mean of 3.33 ($SD = 1.69$) was found to be significantly lower than the midpoint, $t(26) = -2.05$, $p = .05$, 95% CI [−1.33, .00]. This suggests that harm was generally perceived as inflicted unintentionally (both means below

4). Interestingly, the perceptions of intentionality seemed not to be a necessary prerequisite for the harm-made mind effect.

As we had to explicitly refer to the possibility of another entity's involvement as well as to the wound itself to ask the two questions, participants in the follow-up study might have had slightly more (implied) contextual information than the participants in the main study. They may thus inadvertently have paid more attention to the wound and to the (implied) moral agent. Despite this, the wound was still perceived as unintentional rather than intentional in the robotic avatar condition, and there was no significant effect of robotization. That is, even though participants in the follow-up study may have been biased toward perceiving slightly more harm, and toward perceiving this harm as more likely to be other-inflicted and intentional, harm toward the robotic avatar was still perceived as unintentional. The direction of these effects was therefore consistent with the conclusion that the harm-made mind effect in this study was not directly dependent on perceiving harm to have been intentionally caused. Likewise, it is possible that participants in the follow-up study may have perceived more other-causation than the participants in the main experiment. However, there again was no evidence of the effect of robotization. It thus appears that very little, if any, additional context was needed for the wound to be perceived as having been caused by another.

Mediation Analyses

The observation that the harm-made mind effect might not require an explicit involvement of an intentionally malevolent moral agent highlights the need for a closer examination of other potential processes underlying mind attribution toward a harmed entity. One component of these processes appears to be the perception of someone's capacity to feel pain, due to its importance in theories of moral patiency (Gray et al., 2012; Gunkel, 2012; Sytsma & Machery, 2012). The capacity to feel pain has been argued to play a key role in mind perception, for example, when others are perceived as objects (Gray et al., 2011). As further demonstrated by Ward et al. (2013), the perceived capacity to experience pain appears to robustly mediate the effects of harm on mind perception.

However, perceiving capacity to feel pain may not tell the complete story when looking at nonhuman entities. That is, there is reason to believe that humans tend to perceive a clear *gap* between themselves and such entities, and understanding this gap might in turn help explain mind perception. For example, Prguda and Neumann (2014) found that humans showed higher subjective ratings of empathy and physiological responses to phylogenetically similar species. Westbury and Neumann (2008) observed a gradation of empathy for nonhuman targets by phylogenetic relatedness. Furthermore, even among humans, studies have shown ethnic biases toward less empathy shown to members of other races (e.g., Gutsell & Inzlicht, 2012). This evidence suggests the presence of evolutionary biases in empathetic responding to other species that might underlie the extent to which we grant mental states, and ultimately, moral status to animals or robots (Prguda & Neumann, 2014; cf. Gunkel, 2012). In the context of the present work, these results suggest that the apparent phylogenetic distance between humans and robots should reduce empathic responses toward avatars with a robotic appearance. Indeed, in support of the *empathy gap* hypothesis, our statistical results on empathy showed the expected significant reduction of empathy toward the robotic avatar.

We thus decided to perform an exploratory mediation analyses to examine the possibility that pain or empathy might mediate the mind perception effects elicited by our experimental manipulation. Harm has previously been shown to be associated with perceived pain and mind attributions (e.g., Ward et al., 2013). Likewise, empathic concern and compassion are

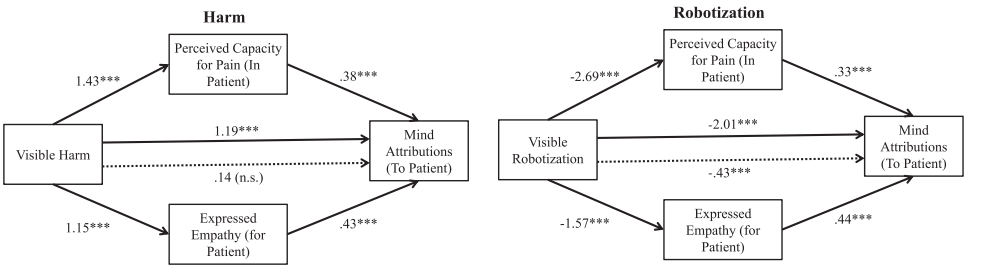

Figure 2. Parallel mediation analysis (Hayes, 2012, Model 4) of harm and robotization on mind attributions mediated by the perceived capacity for pain and empathy expressed for the patients. Solid lines along the center path of each model indicate the indirect relationship between the respective independent variable (harm and robotization) and mind attribution. The dashed lines represent the direct relationships when both mediators (pain and empathy) are accounted for. Harm and robotization were dummy coded with 0 and 1. Numbers reflect unstandardized regression weights. *$p < .05$; ***$p < .001$; $N = 217$.

generally believed to arise from appraisals of harm and suffering (e.g., Goetz, Keltner, & Simon-Thomas, 2010). For robotization, the *empathy gap* as well as mechanistic dehumanization (e.g., Haslam, 2006) suggest that robotic appearance should result in less-expressed empathy. Empathy for pain and suffering of dehumanized others has been argued to be reduced via automatic neural damping mechanisms (e.g., Murrow & Murrow, 2015). Finally, literature on objectification has generally suggested that focusing on someone as an object is associated with decreased perceptions of pain sensitivity (Heflick & Goldenberg, 2009; Loughnan et al., 2010; but see Gray et al., 2011). We therefore expected harm to facilitate perceived pain and empathy, and robotization to inhibit perceived pain and empathy. We further expected that both pain and empathy might act as mediators of the mind attribution effects.

Bootstrapping mediation analyses (10,000 samples; SPSS PROCESS macro, V.2.16.3; Hayes, 2013; Model 4) with pain and empathy as parallel mediators were conducted using 95% CIs. We tested if the avatars' perceived capacity for pain mediated the effects of harm and robotization on mind attribution. We also examined the role of expressed empathy, as an alternative mediator for both effects. The intercorrelated scales of experience, agency, and consciousness were collapsed into a single mind attribution index for these analyses (16 items; $\alpha = .96$), following the approach of Ward et al. (2013).

The link between harm and mind attributions was completely mediated by perceived pain and expressed empathy (Figure 2). The mediation model revealed a significant relationship between harm and mind attributions (CI = [.79, 1.60], $p < .0001$), and a significant indirect effect of the mediators (total indirect effect = 1.06, CI = [.68, 1.43]). This relationship was no longer significant when accounting for pain and empathy (CI = [−.05, .33], $p = .14$). The individual indirect effects of pain and empathy were both significant (pain = .55, CI = [.34, .82]; empathy = .50, CI = [.29, .76]). The contrast between both mediators revealed no significant differences in strength between them (CI = [−.36, .23]).

An equivalent bootstrapping mediation analysis of robotization on mind attributions (10,000 samples) showed a partial mediation by pain and empathy. The significant relationship between robotization and mind attributions (CI = [−2.35, −1.67], $p < .0001$) was substantially reduced by the combined indirect effect of both mediators (total indirect effect = −1.58, CI = [−1.92, −1.27]). The remaining direct effect of robotization, after controlling for the mediators, still remained significant (CI = [−.65, −.21], $p < .001$). Again, the

individual indirect effects of pain and empathy were both significant (pain = −.89, CI = [−1.17, −.67]; empathy = −69, CI = [−.96, −.46]). The contrast between both mediators showed no significant difference (CI = [−.19, .58]). Apparently, robotization reduced the perceived mind of the avatars, and this effect was found to be (partially) mediated by the lower perceived capacity for pain, as well as by less empathy for the more robotic-looking entity. These results suggest that pain and empathy may play complementary roles for mind attribution and that the underlying mechanisms may be similar for harm and robotization, even if both effects go into opposite directions.

The mediation analyses based on a composite mind index followed the approach by Ward et al. (2013). However, it is possible that pain and empathy could play different roles if overall mind attribution is further decomposed into its more fine-grained underlying dimensions of agency and experience (Gray et al., 2007, 2011). Robots are generally perceived as very low in experience but somewhat higher in perceived agency (Gray et al., 2007). This suggests that robotization might affect experience more than agency, a finding that appears to be supported by our present results. Gray and Wegner (2012) found that increases in perceived experience, but not agency, may result in increased feelings of discomfort and uncanniness in response to the robots. Finally, pain might be especially linked to attributing the capacity for experience to the body of another entity (Gray et al., 2011), suggesting that the relationship between pain and experience might be stronger and more direct than the relationship between pain and agency. To probe this possibility, we conducted a set of additional exploratory bootstrapping mediation analyses (10,000 samples) on the agency and experience subscales.

For harm, both mediators showed significant indirect effects on experience (pain = .84, CI = [.50, 1.21]; empathy = .44, CI = [.25, .69]) as well as agency (pain = .31, CI = [.16, .53]; empathy = .53, CI = [.30, .84]). The contrast for experience showed a significantly stronger indirect effect of pain (contrast = .39, CI = [.07, .79]), whereas for agency, there was no significant difference between both mediators (contrast = −.22, CI = [−.58, .06]). For robotization, both mediators again showed significant indirect effects on experience (pain = −.1.27, CI = [−1.61, −.97]; empathy = −.60, CI = [−.85, −.39]), as well as on agency (pain = −.59, CI = [−.88, −.33]; empathy = −.74, CI = [−1.08, −.48]). Contrasts showed a significantly stronger negative indirect effect of pain on experience (contrast = −.67, CI = [−1.10, −25]), yet no significant differences between both mediators for agency (contrast = .16, CI = [−28, .65]). These additional results suggest that perceived capacity for pain might indeed be a better mediator than empathy for the dimension of experience. However, pain and empathy might be equally relevant for understanding attributions of agency.

Discussion

We investigated the impact of visible harm and robotic appearance on perceptions of artificial entities' minds. In accordance with our predictions, harmed entities were attributed mental capacities to a greater degree than unharmed controls, and the human avatar was attributed mental states to a greater extent than the robotic avatar. Surprisingly, we found no significant interaction between harm and robotization, and no indication that the harm-made mind effect diminished for avatars of medium to high perceived consciousness. This indicates that harm was associated with increased mind attributions irrespective of significant differences between human and robotic avatars in baseline perceived consciousness.

The harm-made mind effect has been originally explained via the mechanism of dyadic completion, whereby people automatically grant mental capacities to a victimized entity and

in this way cast it in the role of a moral patient when given a malevolent, intentional agent (Ward et al., 2013). Nevertheless, not depicting an agent, as in the current experiment, does not mean that participants did not perceive an agent to be implicitly involved. Indeed, the results of our follow-up study showed that the facial wound was viewed as having been inflicted by another actor. Therefore, once participants see an entity that looks like it has been harmed, they appear able to readily fill in much of the missing contextual information as well as imbue the harmed body with a mind. The moral patient might thus still be completed in a manner similar to that described previously (Ward et al., 2013), provided the available information is sufficient.[3] However, the present results could also raise some questions with regard to the theory.

If, out of the incomplete moral dyad (i.e., two bodies, one mind, and harm), the mere presence of just one body and harm is already sufficient, then the original dyadic completion account might not be the most parsimonious explanation. The present results suggest that salient harm might sometimes be more important than moral intent. Nonetheless, it should also be noted that our work deviates from prior research (e.g., Khamitov, Rotman, & Piazza, 2015; Ward et al., 2013) by presenting visual *snap shots* rather than more abstract textual vignettes. Such direct depiction of harm might draw less immediate attention to the moral context, as compared with text that might seem strikingly incomplete without said context. It might therefore be easier for perceivers to fill-in, or ignore, incomplete elements, as this might be more consistent with how people see the world. More research using more varied visual vignettes would be required to examine the minimal conditions of visually elicited harm-made minds.

The present work suggests that an explicit presence of a malevolent agent is not a requirement for the harm-made mind effect to occur. The results further hint that the (implicit) moral agent may not necessarily need to act intentionally—as in the follow-up study, intentionality of harm was perceived as, at best, ambiguous. Still, the harm manipulation in our study resulted in increased mind attributions. This may further challenge the assumption of moral typecasting theory that intentionality is an essential determinant of the harm-made mind effect (Ward et al., 2013). In sum, our findings appear to be more consistent with recent research that has demonstrated a denial of agency to harmful agents (Khamitov et al., 2015), that is, observations opposite to what would be predicted if intentionality of harm were a requirement of the harm-made mind.

Overall, work on automatic dyadic completion (Gray et al., 2014), findings of denial of agency to malevolent agents (Khamitov et al., 2015), and the findings of this study suggest that unambiguously intentional harm inflicted by a visible moral agent may not be required for the harm-made mind effect to occur. Rather, a broader model might be needed that allows for a shift toward more patient-oriented ethics and a more patient-oriented definition of moral patiency. This would be in line with recent criticisms of traditional agent-oriented approaches in the philosophy of moral patiency as applied to machines. Gunkel (2012) argued that the question of moral patiency of machines might parallel the patient-oriented question of animal rights philosophy by asking *Can they suffer?* rather than on the moral character of the agent. We believe that an approach featuring visual vignettes might help take a step toward such a clearer patient-oriented focus, as this approach may be able to avoid the necessity of presenting, often highly construed, moral scenarios that are typically found in text-based vignettes in this field.

The results of the mediation analyses suggest that perceived pain might mediate the harm-made mind even in the absence of feelings of empathy for the harmed entity. This has implications for future research, in particular with regard to the question of machine patiency (Gunkel, 2012), and in general for the question of under what conditions seeing

evidence of pain leads to attribution of mental states to someone or something that we would otherwise ignore or even dislike. Social norms, at least in western societies, emphasize that we should show empathy in response to a harmed human being. In consequence, complex moral vignette scenarios that describe fully or liminally conscious humans (e.g., Ward et al., 2013) would first need to control for the effects of norms, before the respective roles of pain and empathy can be further dissociated. Here, examining physical harm done to different types of robotic entities might present an opportunity for future work on the minimal conditions for attributions of mind and moral patiency to machines. However, it should be noted that the mediation analyses in the present results were partially exploratory. While we based most of our initial predictions about the relationships between the variables on the literature, we were not aware of any previous mediation analyses that have investigated the interplay of both factors in mind perception. Furthermore, the concepts of pain perception and empathy often appear to be closely entangled. We were therefore unable to make more precise predictions about the relationships between pain and empathy, and although we decided to test parallel mediation models, it seems plausible that future work might support, for example, a sequential mediation model.

Our experiment was limited to the use of still images. While it extended previous vignette studies, a research design investigating the harm-made mind hypothesis by presenting animations of moral dyads might show even more powerful effects on mind perception. However, a focus on the negative consequences for the victim, such as in this study, might be a key ingredient toward countering victimization, that is, the denial of mental states to a victim perceived to be responsible for the victimizing situation (e.g., an objectified rape victim; Loughnan, Pina, Vasquez, & Puvia, 2013). A further limitation is related to the type of wound representing harm. While we found harm-made mind effects of similar magnitude for both human and roboticized avatars, it is possible that a more robot-typical damage might not elicit the same kind of pain perception. In consequence, the effects of less human-like wound might be less pronounced.

This study did not aim to systematically examine the broad range of visual features that determine robotic appearance. It was concerned with only a subset of visual features (glossiness and plastic appearance of the skin) to create a robotic avatar with the same shape, gender, and ethnicity as the human avatar. While we believe that our design succeeded in this manipulation, the single exemplar of a robotic avatar used in the present work cannot answer the question of which visual features may have been most relevant for the perceptions of the overall robotic appearance and the associated mind attributions. However, our results appear to be in line with the emerging finding that certain social stereotypes and biases, for example, toward *robots of color* (Bartneck et al., 2018), may sometimes mimic findings of first impressions of other humans. The robots shown in such studies (e.g., Nao, SoftBank Robotics) do not always need to be highly human-like, which facilitates manipulation of *simple* material properties (e.g., glossiness or color) across a number of exemplars. Therefore, these particular visual features could be manipulated on the basis of two-dimensional images of extant robots to ensure that the robots are indeed perceived as robots. On the other hand, for the study of mind perception as such, fine-grained control over more complex features associated with visual appearance may be more important. For example, it is possible that a less anthropomorphic-looking robotic avatar might be regarded as more of an object, and that mind infusion effects might cease beyond a certain threshold where the robot no longer looks sufficiently human-like. Features such as the perceived age, gender, or ethnicity of robots might thus require a higher degree of anthropomorphic appearance to elicit comparable effects. Future work might expand upon this study and publicly available three-dimensional resources, to create sufficiently

anthropomorphic virtual robots to study mind perception as well as a broad range of other topics in visual social cognition.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant Sonata (2016/23/D/HS6/02954) from Polish National Science Centre to Aleksandra Swiderska.

Notes

1. Participants' age was calculated based on the year of birth chosen from a drop-down menu, which went back only to the year 1970. Fourteen participants selected the last option available in the menu, *before 1970*. As we did not know their exact age, we treated this information as missing. One participant skipped the question, and their information was likewise treated as missing. With 15 responses missing, we report median age, which is unaffected by missing information. The estimated mean age of our sample was 22.32 years, $SD = 6.20$.
2. Of the 338 people who accessed the survey, 86 failed to finish it; additional 19 were excluded from any data processing due to being below 18 years of age; further 16 were excluded due to indicating they only wanted to browse the pages of the survey (instead of giving consent to participate).
3. As pointed out by one anonymous reviewer, it could be argued that completion of both the agent and the patient might devolve into circular reasoning, if the agent's existence is assumed to depend upon the presence of the patient—and the patient is inferred as a function of the harm inflicted by the agent. While this could be strictly true based on the (linear) requirements of the original theory (Ward et al., 2013), we believe that these results highlight a perceptual process that appears to be considerably more flexible and centered on the harm. In the real world, such harm usually has a cause, and our perceptual system appears remarkably well equipped to perceive cause and effect (and even both at once) even if the evidence is only circumstantial.

ORCID iD


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

References

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, *77*, 240–248.

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive Psychology*, *31*, 219–247.

Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., & Eyssel, F. (2018). Robots and racism. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 196–204). New York, NY: ACM.

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

Batson, C. D. (2008). These things called empathy. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–15). Cambridge, MA: MIT Press.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions: Biological Sciences*, *362*, 679–704.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*, 864–886.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). Thousand Oaks, CA: SAGE.

Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, *36*, 171–180.

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: An evolutionary analysis and empirical review. *Psychological Bulletin*, *136*, 351–374.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*, 1494–1509.

Gosling, S. D., Kwan, V. S., & John, O. P. (2003). A dog's got personality: A cross-species comparative approach to personality judgments in dogs and humans. *Journal of Personality and Social Psychology*, *85*, 1161–1169.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*, 619.

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.

Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology*, *101*, 1207–1220.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*, 206–215.

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.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*, 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124.

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

Halpern, J., & Weinstein, H. M. (2004). Rehumanizing the other: Empathy and reconciliation. *Human Rights Quarterly*, *26*, 561–583.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*, 252–264.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]*. Retrieved from http://www.afhayes.com/public/process2012.pdf

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Heflick, N. A., & Goldenberg, J. L. (2009). Objectifying Sarah Palin: Evidence that objectification causes women to be perceived as less competent and less fully human. *Journal of Experimental Social Psychology*, *45*, 598–601.

Karana, E. (2012). Characterization of 'natural' and 'high-quality' materials to improve perception of bio-plastics. *Journal of Cleaner Production*, *37*, 316–325.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*, 390. doi:10.3389/fpsyg.2015.00390

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.

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, *21*, 1854–1862.

Loughnan, S., Haslam, N., Murnane, T., Vaes, J., Reynolds, C., & Suitner, C. (2010). Objectification leads to depersonalization: The denial of mind and moral concern to objectified others. *European Journal of Social Psychology*, *40*, 709–717.

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*, 455–461.

Murrow, G. B., & Murrow, R. (2015). A hypothetical neurological association between dehumanization and human rights abuses. *Journal of Law and the Biosciences*, *2*, 336–364.

Parkinson, B., & Manstead, A. S. R. (1993). Making sense of emotion in stories and social life. *Cognition and Emotion*, *7*, 295–323.

Prguda, E., & Neumann, D. L. (2014). Inter-human and animal-directed empathy: A test for evolutionary biases in empathetic responding. *Behavioral Processes*, *108*, 80–86.

Riek, L., & Howard, D. (2014). A code of ethics for the human-robot interaction profession. *Proceedings of We Robot*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757805

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.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, *1156*, 81–96.

Sytsma, J., & Machery, E. (2012). The two sources of moral standing. *Review of Philosophy and Psychology*, *3*, 303–324.

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS One*, *12*, e0180952.

Tsankova, E., & Kappas, A. (2016). Facial skin smoothness as an indicator of perceived trustworthiness and related traits. *Perception*, *45*, 400–408. doi: 10.1177/0301006615616748

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.

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

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*, 11374–11379.

Westbury, H. R., & Neumann, D. L. (2008). Empathy-related responses to moving film stimuli depicting human and non-human animal targets in negative circumstances. *Biological Psychology*, *78*, 66–74.