# The moral psychology of artificial intelligence

Correspondence:

Ali Ladak

University of Edinburgh, School of Philosophy, Psychology & Language Sciences

7 George Square, Edinburgh, EH8 9JZ, Scotland

a.ladak@sms.ed.ac.uk

Word count: 3,389

# The Moral Psychology of Artificial Intelligence

Ali Ladak,[1, 2] Steve Loughnan,[1*] Matti Wilks[1*]

[1] University of Edinburgh, School of Philosophy, Psychology & Language Sciences, Edinburgh, Scotland

[2] Sentience Institute, New York, United States

*both authors contributed equally to the paper

**Abstract**

Artificial intelligences (AIs), while often perceived as mere tools, have increasingly advanced cognitive and social capacities. In response, psychologists are studying people's perceptions of AIs as moral agents (entities that can do right and wrong) and moral patients (entities that can be targets of right and wrong actions). This article reviews the extent to which people see AIs as moral agents and patients, and how they feel about such AIs. We also examine how characteristics about ourselves and the AIs affect attributions of moral agency and patiency. We find multiple factors that contribute to attributions of moral agency and patiency in AIs, some of which overlap with attributions of morality to humans (e.g., mind perception) and some that are unique (e.g., sci-fi fan identity). We identify several future directions, including studying agency and patiency attributions to the latest generation of chatbots, and to likely more advanced future AIs that are being rapidly developed.

*Keywords:* Moral agency, moral patiency, morality, artificial intelligence, robots

Artificial intelligences (AIs), such as robots and chatbots, have increasingly advanced cognitive and social capacities. For example, the chatbot *GPT-4* can write poetry, reason through mathematics problems, pass classic theory of mind tests, and communicate responsively in human language (Bubeck et al., 2023). The rapid advancement of AIs raises a host of important moral psychological questions. Do people see AIs as moral agents; entities that can do right and wrong and are morally responsible for their actions? Do they see them as moral patients; entities that can be targets of right and wrong actions and so deserve moral concern? How do people feel about such AIs, and what factors influence whether AIs are seen as moral agents and patients? This article reviews the growing literature on the moral psychology of AI and what it can tell us about human moral psychology.

## Moral Agency: Holding AIs morally responsible

The concept of moral agency is multidimensional and has been measured in various ways in the literature (Table 1). Moral agents are morally responsible, which at the most basic level means being causally responsible for a moral event (Guglielmo, 2015). We can think of moral agents' actions as morally right or wrong, or more strongly, we might consider moral agents themselves as morally good or bad, which can support the notion that they deserve praise or blame for their actions. Moral agency is typically attributed to entities that are perceived as having minds, particularly agentic mental capacities reflecting the capacity to act intentionally (Gray et al., 2012; Malle et al., 2014). People are often motivated to reward or punish moral agents, with goals such as retribution, reform, and deterrence. Finally, people often experience emotions, such as anger and outrage, in response to moral agents' actions (Bigman et al., 2023).

**Do people hold AIs morally responsible?**

Research suggests that people view AIs as moral agents to some degree and think they have a moderate level of capacities associated with moral agency. For example, Gray et al. (2007) found that people think robots have agentic mental capacities (e.g., memory, planning) comparable to a young child, and more than a chimpanzee. Shank and DeSanti (2018) found that people attributed some degree of moral wrongness and responsibility to AIs (e.g., decision-making algorithms, social media bots) for violating various moral norms, such as racist parole decisions. Similarly, people think sophisticated "AI driven robots" deserve blame for causing environmental damage (Kneer & Stuart, 2021), and that "AI programs" and "robots" deserve punishment for medical and military harms (Lima et al., 2021). Beyond cognitive judgements, people experience moral emotions in response to harms caused by AIs: Bigman et al. (2023) found that people expressed some degree of outrage towards decision-making algorithms for sexist hiring decisions, though less than towards humans making the same decisions. In short, people attribute some degree of moral agency to AIs, reflecting that they perceive AIs to have some degree of agentic mental capacities.

Table 1: Key dimensions of moral agency that have been measured in the literature

| **Measure** | **Description** | **Example of use** |
|---|---|---|
| Agentic mental capacities | Moral agency is typically attributed to entities that have agentic mental capacities, such as planning and intentions. | "...judge which character is more capable of making plans and working toward goals." (Gray et al., 2007) |
| Responsibility | Moral agents are perceived to be responsible for a moral event at some level. At the most basic level, this means that an agent is causally responsible for a moral event. | "…To what extent do you feel that [X] is responsible for that decision?" (Nijssen et al., 2022) |
| Moral rightness /wrongness or goodness/badness | Moral agents' actions or the agents themselves can be perceived as morally right/wrong or good/bad. | "Was [AI] wrong for providing this outcome?" (Shank & DeSanti, 2018) |

| | | |
|---|---|---|
| Praise/blame | Moral agents deserve praise and blame for their actions. This depends on factors such as their causal responsibility for an outcome and their intention (Malle et al., 2014) | "How much blame does the AI deserve?" (Young & Monroe, 2019) |
| Reward/ punishment | People are often motivated to reward or punish moral agents for their actions, with goals such as retribution, rehabilitation, or deterrence. | "…To what extent do you think that [X] should be punished?" (Nijssen et al., 2022) |
| Moral emotions (e.g., anger, outrage, disgust) | When a moral agent commits a morally good or bad act, we typically feel moral emotions (e.g., anger, outrage) which can motivate us to punish a bad actor. | ""I am angry at [X's] discriminatory actions" (Bigman et al., 2023) |

**How do people feel about AIs as moral agents?**

Despite attributing moral agency to AIs, people are uncomfortable with AI moral agents. Bigman and Gray (2018) found that people are averse to AIs (e.g., autonomous computer programs) making moral decisions, such as whether to perform a risky surgery to save a child, partly due to AIs' perceived lack of mental capacities. This aversion appears hard to reduce: limiting AIs to advisory roles somewhat reduced discomfort, but increasing AIs' perceived capacity for experience and expertise had limited effects (Bigman & Gray, 2018). This poses a difficult challenge with AIs being increasingly used for decision-making in areas that have significant moral consequences, such as in criminal justice and human resources.

**What kinds of AIs are held morally responsible?**

Given that AIs are attributed some degree of moral agency, what factors influence the extent to which they are seen as moral agents? Theoretical models of moral agency emphasize internal factors, that is, characteristics of the agents themselves (e.g., Guglielmo, 2015). Theoretically, perceptions of agentic mental capacities should be important (Gray et

al., 2012), and empirical research shows that such capacities do affect attributions of moral agency to AIs. For example, Yam et al. (2022) found that an anthropomorphic robot (i.e., with a human-like voice, face, and name) responsible for giving negative feedback in an experiment was perceived as having more agentic mental capacities (e.g., communication, thought) than a non-anthropomorphic one, and this resulted in it being punished more for its feedback. Monroe et al. (2014) found that AIs' (e.g., "AI in a human body", "advanced robot") perceived degree of agentic capacities such as choice and intentions predicted how much they were blamed for moral violations, such as attacking a stranger. Maninger and Shank (2022) also found that capacities such as choice and intentions predicted how much various AIs (e.g., smartphone apps, robots) were blamed for moral violations, though even after accounting for these capacities, they found that AIs were blamed less than humans.

One explanation for why AIs are still blamed less than humans is that they are perceived to lack some mental capacities that are required for human-level moral agency. Nijssen et al. (2022) found that robots described as having emotions were held more responsible and considered more deserving of punishment for committing moral violations than non-emotional robots. This suggests that for AIs to be attributed human-level moral agency they must have, in addition to the agentic mental capacities emphasized by existing theory, experiential mental capacities reflecting the capacity to sense and feel (see Gray et al., 2012). Understanding the effects of experiential capacities on moral agency may be particularly important in the context of AIs for whom experience and agency can come apart, unlike in human moral agents for whom they typically come together.

External factors beyond the agent may also be an important yet neglected part of our understanding of the moral psychology of AIs. For example, the moral domain might be critical: Maninger and Shank (2022) found that AIs were blamed more for violating fairness norms (e.g., biased criminal judgment) and less for violating betrayal norms (e.g., denouncing

one's country). The type of decisions made can also have an effect—Malle et al. (2015) found that robots were blamed more for failing to act to sacrifice an individual for the greater good in a moral dilemma than for taking action, whereas the *opposite* held for humans. A possible explanation for the effects of the external factors is that they interact with the perceived internal capacities of AIs. For example, moral dilemmas like the one just described are difficult for humans partly because of the emotional cost of sacrificing an individual. But AIs are perceived as having little emotion (see below), so the dilemmas are expected to involve less conflict for them and they are expected to act to bring about the greater good.

**Who holds AIs morally responsible?**

There is evidence of developmental changes in attributions of moral agency to AIs: Flanagan et al. (2021) found that 5- to 7-year-olds were less likely than adults to think a humanoid robot would choose to play a game if doing so would hurt someone else's feelings, and were more likely than adults to hold the robot morally responsible. However, they still attributed less responsibility to the robot than to a human child. Little is otherwise known about individual differences in moral agency attributions to AIs. Theoretically, people more likely to anthropomorphize (i.e., attribute nonhumans with human-like characteristics) should be more likely to view AIs as moral agents, since they should grant AIs greater degrees of relevant mental capacities. Epley et al.'s (2007) three-factor model of anthropomorphism theorizes that we anthropomorphize when knowledge about humans is activated and applied to nonhumans, and this depends on cognitive factors (e.g., what is known about nonhuman entities) and motivational factors, in particular the extent to which we are motivated to understand another entity and engage with them socially. There is evidence supporting this model in the context of AIs; for example, Eyssel and Reich (2013) found that people induced to feel lonely (and so feel more motivated to engage socially) attributed more mind to a robot

than a control condition. However, more research is needed to understand whether such effects in turn influence attributions of moral agency to AIs.

## Moral patiency: Granting AIs moral concern

As with moral agency, moral patiency is multidimensional and has been studied with a variety of measures (Table 2). Fundamentally, moral patients are entities that can be helped or harmed, which requires experiential mental capacities such as feeling pleasure or pain (Gray et al., 2012). People tend to express moral concern for how moral patients are treated, consider it morally wrong to harm moral patients, include them in their moral circles (i.e., the boundary that distinguishes entities that receive moral consideration from those that do not), and are less willing to sacrifice them in moral dilemmas. Moral patients may also qualify for moral rights, such as the right to have their lives protected. Finally, people may experience emotions, such as empathy, when they perceive moral patients being harmed, which can motivate helping behavior (Batson et al., 1997).

### Do people grant AIs moral concern?

In contrast to moral agency, people generally do not to see AIs as moral patients, nor do they think that AIs have capacities associated with moral patiency. For example, Gray et al. (2007) found that people think robots have experiential capacities (e.g., fear, pain) comparable to a dead person, and Haslam et al. (2008) found that people think "robots (machines)" have very little capacity for emotion. Pauketat and Anthis (2022) found that people place "robots" and "artificial intelligence" at the outskirts of their moral circles, further from the center than chickens, apple trees, and murderers. And Lima et al. (2020) found that people do not support granting 10 possible moral rights to "robots" and "AI programs" and are only (slightly) favorable to one (the right against cruel punishment and treatment).

However, while people do not explicitly attribute moral patiency to current AIs, they do feel emotions associated with patiency towards them. For example, Riek et al. (2009) found that people feel empathy for human-like robots in emotionally evocative film clips. Tan et al. (2018) found that people felt personal distress and often intervened to help a robot being abused, though they sometimes intervened for reasons other than the robot's moral patiency, such as the financial cost of repairing the robot. Also, while Lima et al. (2020) found little support for moral rights, interventions such as reading about other nonhuman entities that have rights increased people's level of support. Finally, evidence suggests that people think future AIs could be moral patients: Ladak et al. (2023a) found that people think it can be in-principle morally wrong to harm an "artificial being," and Pauketat and Anthis (2022) found that people are more willing to grant future "artificial beings" the capacity for emotion. While people do not explicitly view current AIs as moral patients, such perceptions may change, particularly as AIs become more human-like and advanced.

Table 2: Key dimensions of moral patiency that have been measured in the literature

| **Measure** | **Description** | **Example of use** |
|---|---|---|
| Experiential mental capacities | Moral patients are those entities that have the capacity to be helped or harmed. Thus, it is typically attributed to entities with experiential mental capacities, such as feeling pleasure or pain. | "[X] can experience pain." (Küster and Swiderska, 2021). |
| Moral concern | People typically express concern for how moral patients are treated. | "…indicate how much moral concern you think you should show [X]." (Ladak et al., 2023b) |
| Moral (circle) inclusion | Moral patients are those entities that are included in people's moral circles. | "In which circle of moral concern would you put [X]?" (Pauketat & Anthis, 2022) |
| Harm/help | People consider it morally wrong to harm moral patients, and morally right to help them. | "Which [X] do you think it would be more morally wrong for you to harm?" (Ladak et al., 2023a) |
| Moral dilemma decisions | The more moral patiency an entity is granted, the less they are likely to be sacrificed in moral dilemmas. | "When […], the other concert visitors will survive. However, [X] will be fatally harmed. When […], the visitors will quickly die, but nothing will happen to [X]." |

| | | |
|---|---|---|
| | | Will you […]? (Nijssen et al., 2019) |
| Moral rights | Entities granted the strongest degree of moral patiency are granted inviolable rights, such as a right to life. | "[X] should have the right to life." (Lima et al., 2020) |
| Moral emotions (e.g., empathy) | When people witness a moral patient being harmed, they often feel emotions such as empathy, which can motivate helping behavior. | "How sorry do you feel for [X]?" (Riek et al., 2009) |

**How do people feel about AIs as moral patients?**

Mirroring people's aversion to AI moral agents, people express discomfort towards AIs with capacities associated with patiency. Gray and Wegner (2012) found that perceiving the capacity for experience in robots causes the "uncanny valley," the creepy feeling some people report when interacting with robots that closely resemble humans. This discomfort can be reduced by stripping robots of their capacity for experience (Yam et al., 2021), however, because the capacity for experience is typically associated with moral patiency, this approach may result in the denial of patiency for AIs. Moreover, as noted earlier, the capacity for experience may be important for attributing human-level moral agency to AIs and for having AIs make decisions humans support. The interactions between perceptions of experience, discomfort, and moral agency and patiency should be further explored in future research.

**What kinds of AIs are granted moral concern?**

Like with moral agency, internal and external factors can influence the extent to which AIs are attributed moral patiency. Theoretically, internal factors should reflect experiential mental capacities (Gray et al., 2012). Research suggests that these capacities do influence moral patiency attributions to AIs. For example, Nijssen et al. (2019) found that human-looking robots described in anthropomorphic language were less likely to be sacrificed in moral dilemmas than mechanical-looking robots described in mechanistic

language, and this was due to their greater perceived capacity for experience. Similarly, robots with social rather than economic functions were attributed greater capacity for experiencing emotions and were less likely to be harmed (Wang and Krumhuber, 2018). Ladak et al. (2023a) found both experiential and agentic capacities predicted how morally wrong people think it is to harm AIs, including expressing and recognizing emotions, cooperating, and making moral judgments.

Turning to external factors, Eyssel and Kuchenbrandt (2012) found that an ingroup robot was perceived as having greater capacity for experience than an outgroup robot. Küster and Swiderska (2021) found that people attributed greater capacity for experience and for feeling pain to a robot observed being harmed than one unharmed. And Tanibe et al. (2017) found that imagining oneself helping a robot led to greater perceptions of the robot's capacity to feel pleasure, greater willingness to grant the robot rights, and lesser willingness to harm the robot. In short, external factors affect perceptions of the internal capacities of AIs, such as their capacity for experience, which can in turn affect the extent to which they are attributed moral patiency.

**Who grants AIs moral concern?**

Pauketat and Anthis (2022) tested a range of predictors and found that sci-fi fan identity (identification with the science fiction fan group), substratism (prejudice against AIs), techno-animism (attributing life to technological entities), and feeling positive emotions towards AIs most consistently predicted moral patiency attributions to AIs. There is also evidence of developmental changes in attributions of moral patiency to AIs—Reinecke et al. (2021) found that children granted robots greater capacity for experience, vulnerability to harm, and a claim to protection than did adults. Children, particularly younger ones, also distinguished less than adults did between a human child and a robot on these measures. An individual's psychological state can also have an effect—Ladak et al. (2023b) found that

people who felt more empathy and closeness with a human-like AI worker expressed greater moral concern for AIs as a group.

## Future Directions and Conclusions

People attribute some degree of moral agency to AIs, reflecting that AIs are perceived to have a moderate degree of agentic capacities (Gray et al., 2007). By contrast, people attribute little moral patiency to AIs, reflecting that AIs are generally not perceived to have experiential capacities (Gray et al., 2007). Why do people perceive agentic but not experiential capacities in AIs? The simplest explanation is that AIs are typically designed this way: they calculate, decide, and act, without any feeling or emotion. However, as AIs become increasingly human-like and expressive, like the latest chatbots, people may increasingly perceive them as experiential, and in turn attribute them moral patiency. These attributions are currently less well understood, but given the possible implications for human-AI interaction, should be a priority for future research. Additionally, research should aim to understand the effects of perceived patiency on attributions of moral agency (and vice versa). Such effects are potentially important in the context of AIs, for whom experiential and agentic capacities can come apart, unlike with humans for whom they typically come together.

While AIs are usually attributed less moral agency than humans (e.g., Kahn et al., 2012), people sometimes attribute them *more* moral agency than humans. For example, Hong et al. (2020) found that people blamed self-driving cars more than humans for causing accidents, and this was due to their perceived greater competency at driving. As AIs become increasingly advanced and competent, will they be attributed greater moral agency than humans in other contexts too? Or will the increasing human-likeness of contemporary AIs (e.g., communication style) limit their attributed moral agency at a closer-to-human level?

Similarly, will people perceive super-human levels of experience in future AIs, and will they attribute such AIs super-human levels of moral patiency? These questions are important because if AIs continue to exceed humans in their capacities, they could challenge humans' longstanding position at the top of the hierarchy of moral agency and patiency.

Some researchers argue that people's desire to blame and punish AIs is misplaced because AIs do not have the capacities that would make them appropriate targets of such behaviors (e.g., Danaher, 2016). However, this desire may reflect an important aspect of human moral psychology—that we intuitively seek a target to blame for a morally bad outcome, even if one does not exist. However, people may also have other reasons to punish AIs: Lima et al. (2021) found that people want to punish AIs because they think AIs can learn from their mistakes (and not for retribution or deterrence). This possibility raises several questions: What kinds of punishment do people think would enable AIs to learn? What kinds of punishment do people consider morally acceptable? And will punishment serve retributive and deterrent functions for future, more advanced AIs?

We must also acknowledge the broad scope of what *AI* means. Research has examined perceptions of human-like and mechanical robots, decision-making algorithms, and self-driving cars, among other entities. There are likely important similarities and differences in people's moral judgements of these entities, such as human-like robots being perceived to have greater capacity for emotion than mechanical robots (Nijssen et al., 2019). However, such comparative research is sparse. A comprehensive comparative exploration of how people judge different types of AIs is an important next step for the field.

There are several challenges regarding the design of AIs. A first challenge concerns acceptability. People feel discomfort with AI moral agents and patients, despite their increasing prevalence in society. How could AIs be designed to reduce discomfort? A second

challenge concerns transparency. Can AIs be designed so that their internal workings are sufficiently transparent that we can tell whether they are genuine moral agents and patients? How would such transparency affect attributions of moral agency and patiency to AIs? A third challenge concerns alignment. Can (and should) AIs be designed to make moral decisions in the ways that humans want them to? Do people's expectation that AIs make different decisions to humans (e.g., Malle et al., 2015) merely reflect current AIs, or does this generalize to future, more advanced AIs? Addressing these challenges should not only be a focus for computer scientists and engineers, but psychologists as well.

The fact that people attribute moral agency and (to a lesser extent) patiency to AIs via some of the same fundamental processes (e.g., mind perception, anthropomorphism) as they do to biological entities shows the generalizability of important aspects of human moral psychology. However, AIs are (and will likely continue to be) very different to biological entities. As such, other social identities and psychological processes (e.g., sci-fi fan identity, substratism) that do not operate in the context of biological entities also influence our moral judgments of AIs. Further, as AIs continue to (rapidly) improve their capabilities, arguably beyond human level on some tasks (e.g., driving), decade old certainties about morality based largely on human psychology are being called into question. Given the rapid advancement of AI, it is critical for researchers to further identify and integrate these effects into psychological theory and the design and development of AI systems.

## Recommended Readings

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences*, 23(5), 365-368. A review article looking at factors that influence judgments of moral responsibility in robots.

Gray, K., Young, L., & Waytz, A. (2012). (See references). A theoretical article that maps the two dimensions of mind perception theory (agency and experience) to moral agency and moral patiency.

Ladak, A., Wilks, M., & Anthis, J. R. (2023). (See references). An experiment testing the effects of encouraging perspective taking on moral attitudes towards AIs.

Nijssen, S. R. R., Müller, B. C. N., Baaren, R. B. van, & Paulus, M. (2019). (See references). An experimental study testing the effects of anthropomorphism, agency, and experience on people's willingness to sacrifice robots in moral dilemmas.

Pauketat, J. V. T., & Anthis, J. R. (2022). (See references). A study testing the effects of a range of predictors on moral patiency attributions to AIs.

## Acknowledgements

Many thanks to Jacy Reese Anthis, Michael Dello-Iacovo and Janet Pauketat for helpful comments on earlier versions of this article.

## References

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., & Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, *72*(1), 105–118. https://doi.org/10.1037/0022-3514.72.1.105

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of*

*Experimental Psychology: General, 152(1)*, 4–27. https://doi.org/10.1037/xge0001250

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribiero, M.T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Eyssel, F. A., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots). On the effects of loneliness on psychological anthropomorphism. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013)*. https://pub.uni-bielefeld.de/record/2633786

Flanagan, T., Rottman, J., & Howard, L. H. (2021). Constrained Choice: Children's and Adults' Attribution of Choice to a Humanoid Robot. *Cognitive Science*, *45*(10), e13043. https://doi.org/10.1111/cogs.13043

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. (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

Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01637

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, Inhuman, and Superhuman: Contrasting Humans with Nonhumans in Three Cultures. *Social Cognition*, *26*(2), 248–258. https://doi.org/10.1521/soco.2008.26.2.248

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 33–40. https://doi.org/10.1145/2157689.2157696

Kneer, M., & Stuart, M. T. (2021). Playing the Blame Game with Robots. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 407–411. https://doi.org/10.1145/3434074.3447202

Küster, D., & Swiderska, A. (2021). Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *International Journal of Psychology*, *56*(3), 454–465. https://doi.org/10.1002/ijop.12715

Ladak, A., Harris, J., & Anthis, J. R. (2023a). Features of Moral Consideration for Artificial Entities: A Conjoint Experiment. https://psyarxiv.com/235vp/

Ladak, A., Wilks, M., Anthis, J.R. (2023b). Extending Perspective Taking to Non-Human Animals and Artificial Entities. *Social Cognition.* https://doi.org/10.1521/soco.2023.41.3.274

Lima, G., Cha, M., Jeon, C., & Park, K. S. (2021). The Conflict Between People's Urge to Punish AI and Legal Systems. *Frontiers in Robotics and AI*, *8*. https://www.frontiersin.org/articles/10.3389/frobt.2021.756242

Lima, G., Kim, C., Ryu, S., Jeon, C., & Cha, M. (2020). Collecting the Public Perception of AI and Robot Rights. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 135:1-135:24. https://doi.org/10.1145/3415206

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C. (2015). Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 117–124. http://doi.org/10.1145/2696454.2696458

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*, 100154. https://doi.org/10.1016/j.chbr.2021.100154

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, *27*, 100–108. https://doi.org/10.1016/j.concog.2014.04.011

Nijssen, S. R. R., Müller, B. C. N., Baaren, R. B. van, & Paulus, M. (2019). Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Social Cognition*, *37*(1), 41-S2. https://doi.org/10.1521/soco.2019.37.1.41

Nijssen, S. R. R., Müller, B. C. N., Bosse, T., & Paulus, M. (2022). Can you count on a calculator? The role of agency and affect in judgments of robots as moral agents. *Human–Computer Interaction*, *0*(0), 1–17. https://doi.org/10.1080/07370024.2022.2080552

Pauketat, J. V. T., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior*, *136*, 107372. https://doi.org/10.1016/j.chb.2022.107372

Reinecke, M. G., Wilks, M., & Bloom, P. (2021). Developmental changes in perceived moral standing of robots. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/8f32d068

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6. https://doi.org/10.1109/ACII.2009.5349423

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G., & Steinfeld, A. (2018). Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 169–177. https://doi.org/10.1145/3171221.3171247

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

Wang, X., & Krumhuber, E. G. (2018). Mind Perception of Robots Varies With Their Economic Versus Social Function. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01230

Yam, K. C., Bigman, Y., & Gray, K. (2021). Reducing the uncanny valley by dehumanizing humanoid robots. *Computers in Human Behavior*, *125*, 106945. https://doi.org/10.1016/j.chb.2021.106945

Yam, K. C., Goh, E.-Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology*, *102*, 104360. https://doi.org/10.1016/j.jesp.2022.104360