Perceived Moral Agency of Non-Moral Entities: Implications and Future Research Directions for Social Robots

Joel Wester
joelw@cs.aau.dk
Aalborg University
Aalborg, Denmark

Eike Schneiders
eike.schneiders@nottingham.ac.uk
Nottingham University
Nottingham, United Kingdom

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

Humans tend to perceive human qualities in interactive systems. This particularly applies to social robots that utilise human attributes such as human body characteristics and natural language capabilities. Social robots with such characteristics are increasingly deployed in critical settings, such as health and well-being, where it is key to align robot behaviour with end-user expectations. Relatively little is known about how people perceive these social robots' moral agency. In this position paper, we stress the difference between moral agency and perceived moral agency, and argue that the latter is a timely concern. We discuss the implications of perceived moral agency and outline research directions to explore how humans make sense of social robots in critical settings through perceived moral agency.

KEYWORDS

Perceived Moral Agency, Non-Moral, Human-Robot, Social Robots, Critical Settings

ACM Reference Format:
Joel Wester, Eike Schneiders, and Niels van Berkel. 2023. Perceived Moral Agency of Non-Moral Entities: Implications and Future Research Directions for Social Robots. In *HRI '23: ACM/IEEE International Conference on Human-Robot Interaction (HRI) @ Workshop on Perspectives on Moral Agency in Human-Robot Interaction, March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 3 pages.

1 PERCEIVING SOCIAL ROBOTS

Prior work has provided overwhelming evidence for humans' tendency to anthropomorphise interactive systems, such as social robots, by perceiving human attributes [24, 32], (e.g., friendliness [1]). This phenomenon is often triggered by social robots' human-like attributes through embodied and non-embodied acts [25, 26]. Various explanatory models from cognitive science can contribute to our understanding of how and why humans make sense of human-like attributes in robots in human-nonhuman interactions, e.g., through perception, attention, memory, or information processing [8]. Kim et al. point to information processing of social robots as a largely unexplored field of research [19]. While cognitive science shows promise in understanding better how humans anthropomorphise social robots, a gap in knowledge regarding humans' use of different cognitive models to make sense of social robots in different contexts remains. Konok et al. illustrate the complexity of this knowledge gap by assessing which qualities humans desire in companion dogs and how this can be implemented in social robots [20]. Among the many categories and subcategories that participants found desirable in companion dogs were personality, smartness, and kindness. Qualities as such share an indirect nature, meaning that it is difficult to assess precisely what sub-categories contribute to different conceptualizations made by humans in attempting to understand their experience.

Firstly, research in HRI commonly assesses peoples' understanding of human attributes in robots, by assessing, e.g., perceived friendliness in social robots [1]. Despite the extensive research covering aspects of anthropomorphic attributes in social robots, the challenge remains to precisely assess which underlying social cues contribute to friendliness. Following this argument, social robots' friendliness cues can directly trigger humans' perception of specific robot behaviour—whilst indirectly triggering humans' cognitive processes (e.g., mental representations) of the social robot.

Secondly, as the behaviour of social robots grows more complex, it follows that humans will increasingly anthropomorphise social robots [17]. With an increase in perceived anthropomorphism, additional human attributes and behaviours will also grow in prominence—such as perceiving moral agency (e.g., believing that a robot has a sense of what is right or wrong). Assessing higher-level constructs, such as the aforementioned friendliness, can thus inform how humans anthropomorphise social robots and inform future steps in HRI research.

Thirdly, we suggest that such higher-level constructs increase humans' tendency to anthropomorphise robots, and consequently increase perceived moral agency, which we argue is key to understanding how humans make sense of social robots. Moreover, we argue that perceived moral agency—or the triggering of believing that someone or something has a sense of what is right or wrong—can inform the design of social robots in critical settings, such as social robots in health and well-being. We outline future directions for HRI and HCI researchers to understand better how humans make sense of social robots in critical settings through perceived moral agency.

2 OUTLINING MORAL AGENCY

Moral agency has long been a topic of investigation outside the domain of both HRI and HCI. Moral agency is a concept describing the capacity to act on what is right and wrong (and the capacity to be responsible for one's actions) [16]. Only recently has the HRI community started investigating moral agency's role in human-robot interaction. Therefore, the role of moral agency in human-robot interaction remains largely unexplored.

Jackson et al. argued that when social robots' abilities get more complex, we naturally perceive them as having morals, similar to human-human interactions [17]. Following the growing complexity

HRI '23, March 13–16, 2023, Stockholm, Sweden
2023. ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

of robots, Jackson et al. argued that we need to provide robots with moral competence. This aligns with research focusing on using computational methods and crowdsourced knowledge to formalise morality in systems [2]. However, the problem of formalising morality is that morality is fundamentally subjective—meaning that what one person believes to be moral may significantly differ from what another person believes to be moral. In this paper, we argue that a larger focus should be placed on how humans *perceive morality* in non-moral entities, such as social robots.

Regarding the moral understanding of non-moral entities, Semler argues that *"if a non-sentient entity can provide sufficient observable output, we must infer that it understands."* [27, p. 913]. Considering social robots providing observable output, we can also theorise in the following way: *if a non-moral entity can provide sufficient observable output, we can perceive it as having moral agency.* Moreover, Semler distinguishes between moral agency and quasi-moral agency—suggesting that quasi-moral agency is a concept that does not require criteria crucial for moral agency in humans (e.g., being accountable for one's actions) [27].

Recent work by Banks suggests perceived moral agency as crucial for realising more meaningful human-system interactions by looking at how humans perceive moral agency in non-moral entities [4]. Perceived moral agency is a promising approach to investigate further 'observable output' provided by 'non-moral entities' whilst avoiding the formalisation of morality. This is further illustrated by a recent paper on placebo effects in AI support. Kosch et al. results suggest that when humans *believing* in receiving AI support increase their expectations of one's own task performance [21], thereby supporting our argument that human sense-making relies heavily on subjectivity.

3 IMPLICATIONS

As aforementioned, different explanatory frameworks in cognitive science can contribute to how humans make sense of human-robot interactions. These frameworks can further inform HRI research and design, particularly concerning moral cognition [9, 12, 14, 22]. The relation between moral and social aspects of interactions is complex since humans naturally and subjectively make sense of those interactions through their mental models [10] and as impacted by their contextual setting [29].

By understanding how morality guides human perception, we can obtain a better understanding of how humans make sense of robots in different environments. Revisiting the research described in our introduction, we outline that HRI research utilises fundamentally human psychological constructs, such as friendliness or warmth [1], goodwill [32], and humour [24], combined with human physiological constructs such as embodied acts [25].

With an increase in a robot's social abilities (e.g., warmer behaviour), perceived moral agency increases. This implies that humans' subjective understanding (i.e., perception) of moral agency in social robots is key to understanding how humans evaluate robots. However, we know little about what more complex social robots will enact in humans, specifically in more critical settings. Therefore, it is crucial to avoid formalising moral competence in social robots before fundamentally understanding how humans construct their perception of a robot's moral agency. In doing so, we can contribute to the development of more appropriate, desirable, and relevant social robots in various domains.

In this position paper, we described a brief narrative leading to the argument that perceived moral agency is crucial for humans making sense of social robots, specifically in critical settings. Following, we outlined how an increased focus on perceptions of moral agency is key to understanding more socially complex robots. Based on this argument, we propose future research directions where the perception of moral agency may play a key role.

4 FUTURE RESEARCH DIRECTIONS

Following the argument of avoiding moral formalisation by focusing on the perception of moral agency, an investigation into how and what effects perceived robot moral agency has on crucial factors (e.g., trust) in critical human-robot interactions (e.g., mental health or long-term support) is necessary.

In mental health, therapists must carefully strategise their actions and decisions to avoid causing patients any psychological harm [13]. One important factor in interactions as such is trust. Hall et al. argued that trust is key for positive treatment outcomes in therapist-patient relations [15]. Social robots used in mental health interventions have received little attention in the literature. Recently, Kabacińska et al. found that robots intervening with children positively affect relief and distress [18]. Moreover, Björling et al. showed that teens tended to engage in interactions with social robots, expressing how social robots can function as emotional support [6]. As Björling et al. highlight, designing teen-robot interactions is categorically different from other human-robot interactions [7]. Furthermore, Baecker et al. argue that mental health concerns around older adults are increasing and investigate how social robots can be introduced to support older adults' mental health (e.g., depression) in the domestic space [3].

Furthermore, there are various factors that influence peoples' continuous upholding of their social relations. Simpson et al. argue that trust is crucial for establishing and maintaining such long-term human-human relationships [28]. In well-being, in contrast to mental health, social robots used in long-term interactions have received more attention. However, there is no established understanding of influential factors in long-term human-robot interactions. Designing to establish and maintain long-term human-robot relationships is, therefore, an open question for both the HRI and HCI community [5]. Researchers have investigated how social robots in long-term interactions can benefit human well-being in different ways (e.g., social companionship for older adults [31], social support for children [23], motivational support for diabetic children to keep a diary [30], or activity support for people with dementia [11]).

We suggest perceived moral agency to be a key factor in better understanding social robots in critical settings. In these contexts, alignment of robot behaviour with user expectations is key in building the necessary trust. We therefore call on the broader HRI and HCI research community to pursue a better understanding of people's perceptions of morality in non-moral entities.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation project 'Algorithmic Explainability for Everyday Citizens'.

REFERENCES

[1] Selen Akay, Berkay Arslan, Sabahat C. Bagci, and Junko Kanero. 2022. "My Robot Friend": Application of Intergroup Contact Theory in Human-Robot Interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 665–668. https://doi.org/10.1109/HRI53351.2022.9889570

[2] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M.J. Crockett, Jim A.C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. 2022. Computational ethics. *Trends in Cognitive Sciences* 26, 5 (2022), 388–405. https://doi.org/10.1016/j.tics.2022.02.009

[3] Annalena Nora Baecker, Denise Y. Geiskkovitch, Adriana Lorena González, and James E. Young. 2020. Emotional Support Domestic Robots for Healthy Older Adults: Conversational Prototypes to Help With Loneliness. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 122–124. https://doi.org/10.1145/3371382.3378279

[4] Jaime Banks. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371. https://doi.org/10.1016/j.chb.2018.08.028

[5] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[6] Elin A. Björling, Emma Rose, Andrew Davidson, Rachel Ren, and Dorothy Wong. 2020. Can We Keep Him Forever? Teens' Engagement and Desire for Emotional Connection with a Social Robot. *International Journal of Social Robotics* 12, 1 (01 Jan 2020), 65–77. https://doi.org/10.1007/s12369-019-00539-6

[7] Elin A. Björling, Kyle Thomas, Emma J. Rose, and Maya Cakmak. 2020. Exploring Teens as Robot Operators, Users and Witnesses in the Wild. *Frontiers in Robotics and AI* 7 (2020). https://doi.org/10.3389/frobt.2020.00005

[8] Jeffrey M. Bradshaw and J. Chris Forsythe. 2012. Cognitive science and sociocognitive theory for the HRI practitioner. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 495–496.

[9] Marco Brambilla and Colin Wayne Leach. 2014. On the Importance of Being Moral: The Distinctive Role of Morality in Social Judgment. *Social Cognition* 32, 4 (2014), 397–408. https://doi.org/10.1521/soco.2014.32.4.397

[10] John M. Carroll and Judith Reitman Olson. 1988. Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, Martin Helander (Ed.). North-Holland, Amsterdam, 45–65. https://doi.org/10.1016/B978-0-444-70536-5.50007-5

[11] Sara Casaccia, Gian Marco Revel, Lorenzo Scalise, Roberta Bevilacqua, Lorena Rossi, Robert A. Paauwe, Irek Karkowsky, Ilaria Ercoli, J. Artur Serrano, Sandra Suijkerbuijk, Dirk Lukkien, and Henk Herman Nap. 2019. Social Robot and Sensor Network in Support of Activity of Daily Living for People with Dementia. In *Dementia Lab 2019. Making Design Work: Engaging with Dementia in Context*, Rens Brankaert and Wijnand IJsselsteijn (Eds.). Springer International Publishing, Cham, 128–135.

[12] Jean Decety and Jason M. Cowell. 2014. The complex relation between morality and empathy. *Trends in Cognitive Sciences* 18, 7 (2014), 337–339. https://doi.org/10.1016/j.tics.2014.04.008

[13] Sona Dimidjian and Steven D Hollon. 2010. How would we know if psychotherapy were harmful? *American Psychologist* 65, 1 (2010), 21. https://doi.org/10.1037/a0017299

[14] Gerd Gigerenzer. 2010. Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality. *Topics in Cognitive Science* 2, 3 (2010), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x

[15] Amanda M. Hall, Paulo H. Ferreira, Christopher G. Maher, Jane Latimer, and Manuela L. Ferreira. 2010. The Influence of the Therapist-Patient Relationship on Treatment Outcome in Physical Rehabilitation: A Systematic Review. *Physical Therapy* 90, 8 (2010), 1099–1110. https://doi.org/10.2522/ptj.20090245

[16] Jaana Hallamaa and Taina Kalliokoski. 2020. How AI Systems Challenge the Conditions of Moral Agency?. In *Culture and Computing*, Matthias Rauterberg (Ed.). Springer International Publishing, Cham, 54–64. https://doi.org/10.1007/978-3-030-50267-6_5

[17] Ryan Blake Jackson and Tom Williams. 2019. On perceived social and moral agency in natural language capable robots. In *2019 HRI workshop on the dark side of human-robot interaction. Jackson, RB, and Williams*. 401–410.

[18] Katarzyna Kabacińska, Tony J. Prescott, and Julie M. Robillard. 2021. Socially Assistive Robots as Mental Health Interventions for Children: A Scoping Review. *International Journal of Social Robotics* 13, 5 (01 Aug 2021), 919–935. https://doi.org/10.1007/s12369-020-00679-0

[19] Mingyu Kim, Taesoo Kwon, and Kwanguk Kim. 2018. Can Human–Robot Interaction Promote the Same Depth of Social Information Processing as Human–Human Interaction? *International Journal of Social Robotics* 10, 1 (01 Jan 2018), 33–42. https://doi.org/10.1007/s12369-017-0428-5

[20] Veronika Konok, Beáta Korcsok, Ádám Miklósi, and Márta Gácsi. 2018. Should we love robots? – The most liked qualities of companion dogs and how they can be implemented in social robots. *Computers in Human Behavior* 80 (2018), 132–142. https://doi.org/10.1016/j.chb.2017.11.002

[21] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* (2022). https://doi.org/10.1145/3529225

[22] Justin F. Landy, Jared Piazza, and Geoffrey P. Goodwin. 2016. When It's Bad to Be Friendly and Smart: The Desirability of Sociability and Competence Depends on Morality. *Personality and Social Psychology Bulletin* 42, 9 (2016), 1272–1290. https://doi.org/10.1177/0146167216655984

[23] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Long-Term Interactions with Empathic Robots: Evaluating Perceived Support in Children. In *Social Robotics*, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 298–307.

[24] Nicole Mirnig, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2017. Elements of Humor: How Humans Perceive Verbal and Non-Verbal Aspects of Humorous Robot Behavior. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 211–212. https://doi.org/10.1145/3029798.3038337

[25] Raul Benites Paradeda, Mojgan Hashemian, Rafael Afonso Rodrigues, and Ana Paiva. 2016. How Facial Expressions and Small Talk May Influence Trust in a Robot. In *Social Robotics*, Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He (Eds.). Springer International Publishing, Cham, 169–178. https://doi.org/10.1007/978-3-319-47437-3_17

[26] P.L. Patrick Rau, Ye Li, and Dingjun Li. 2009. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior* 25, 2 (2009), 587–595. https://doi.org/10.1016/j.chb.2008.12.025 Including the Special Issue: State of the Art Research into Cognitive Load Theory.

[27] Jen Semler. 2022. Artificial Quasi Moral Agency. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 913. https://doi.org/10.1145/3514094.3539549

[28] Jeffry A. Simpson. 2007. Psychological Foundations of Trust. *Current Directions in Psychological Science* 16, 5 (01 Oct 2007), 264–268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

[29] Niels van Berkel, Benjamin Tag, Jorge Goncalves, and Simo Hosio. 2022. Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology* 41, 3 (2022), 502–518. https://doi.org/10.1080/0144929X.2020.1818828

[30] Esther J.G. van der Drift, Robbert-Jan Beun, Rosemarijn Looije, Olivier A. Blanson Henkemans, and Mark A. Neerincx. 2014. A Remote Social Robot to Motivate and Support Diabetic Children in Keeping a Diary. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 463–470. https://doi.org/10.1145/2559636.2559664

[31] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. 2012. Designing Relational Agents as Long Term Social Companions for Older Adults. In *Intelligent Virtual Agents*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 289–302.

[32] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2019. Effective Persuasion Strategies for Socially Assistive Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 277–285. https://doi.org/10.1109/HRI.2019.8673313