Perspectives on Moral Agency for HRI: Cognitive Construct or Ontological State?

Toward Moral Agency Quantification in Human-Robot Interaction

Boyoung Kim
bkim55@gmu.edu
George Mason University Korea
Songdo, Incheon, South Korea

Elizabeth Phillips
ephill3@gmu.edu
George Mason University
Fairfax, Virginia, USA

Qin Zhu
qinzhu@vt.edu
Virginia Tech
Blacksburg, Virginia, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, CO, USA

ABSTRACT

Establishing when, how, and why robots should be considered moral agents is key for advancing human-robot interaction (HRI). Robotic moral agency has significant implications for how people should and do hold robots morally responsible, ascribe blame to them, develop trust in their actions, and determine when these robots wield moral influence. In this paper, we review three perspectives on robot moral agency, and present a framework which can be used to initiate research on measuring and quantifying robot moral agency. In particular this framework grounds robot moral agency in evidence of robot autonomy, interactivity, adaptability, and capacity for morally relevant action. Finally, we close with a list of discussion points for the community on the topic of robot moral agency.

1 ROBOTIC MORAL AGENCY IN HUMAN-ROBOT INTERACTION

Robots are increasingly anticipated to fulfill active roles as assistants, teammates, companions, and supervisors in various domains of society. As the number and types of roles robots serve increases and diversifies, the influence robots exert over everyday life will grow in kind. In response, many Human-Robot Interaction (HRI) and machine ethics researchers have discussed how robots may play possible roles as moral actors and influencers in society by, for example, making decisions that entail moral consequences [9, 12] taking responsibility for their choices [11, 14], or using moral language to change human behavior [8, 15].

The degree of a robot's moral influence may depend on whether that robot holds the status of *moral agent*: that is, whether that actor is seen as autonomous, adaptive, interactive, and capable of morally relevant actions [5]. Further, whether or not a robot is considered a moral agent has significant implications for how researchers, designers, and users can, should, and do make sense of robots, and may inform the social and moral cognitive and behavioral processes that robots evoke in others.

Moreover, robotic moral agency has significant implications for how people should and do hold robots morally accountable, develop trust in their decisions, and determine when and how these robots wield moral influence. Thus, because of the critical significance of robotic moral agency, researchers from several disciplines have started to investigate how we can determine whether and to what extent robots can be evaluated as moral agents.

In this paper, we review three key perspectives on the topic of artificial agents' moral agency and present a philosophical framework that is aligned with one of these three perspectives. We view this framework as theoretically well-grounded for initiating research on whether and to what extent a robot constitutes a moral agent in HRI contexts. Finally, we close with a list of suggested discussion points to encourage continued interdisciplinary discourse on the topic of robot moral agency.

2 EXISTING PERSPECTIVES ON MORAL AGENCY OF ARTIFICIAL ENTITIES

There are three distinct perspectives on the topic of artificial moral agency in HRI and Technology Ethics [3].

The first perspective suggests that, because robots, at least in their current stage of development, do not meet the premise of having properties like consciousness, freewill, mental states, and sentience, it is not right or possible to assign moral agency to them [2]. Yet the usefulness of this perspective falls short for HRI researchers in at least two ways. First, this perspective seems to assess robots' moral agency according to the same yardsticks used to judge humans. But, per Kahn's New Ontological Category Hypothesis [7], robots fill a unique ontological niche, and thus it may be problematic to assume that humans would or should use the same criteria for reasoning about robot moral agency as they do about humans; as such, we may need to consider robots differently from humans when it comes to traits like moral agency. Second, this perspective seems to argue about the fundamental philosophical truth of robots' underlying status. Yet as robot designers and researchers, we argue that what fundamentally matters for human-robot interaction is not about specifying what philosophical status of robots hold, or even what is truly going on "under the hood." Instead, the fundamental matter is understanding how people perceive, mentally model, and generally make sense of robots. In particular, how people generally perceive robots is important for understanding and predicting how people interact with robots, such as the impact of anthropomorphism in HRI [13].

These shortcomings are addressed by the second perspective, which specifically suggests that criteria, which are different from those used to assess humans, be used to assess robots' moral agency, and that these criteria do not need to rely on properties like consciousness, freewill, mental states, and sentience [5, 10]. Floridi and Sanders [5], for example, argue that *for a user operating from a particular perspective at which they are able to make particular observations about a robot*, a robot is an agent if it is considered to be autonomous, interactive, and adaptive. A robot is then a *moral* agent from that perspective if it is further perceived as capable of morally relevant actions, or actions that cause moral good or evil. This perspective thus emphasizes key traits necessary for agency in machines; and moreover, while it retains moral agency as an ontological state, it conditions that state on the perspective of a particular observer, thus emphasizing the centrality of user perceptions. However, in its current form, this framework proposed by Floridi and Sanders [5] remains to be a philosophical theory and does not provide a means for quantifying these perceptions in order for practitioners (rather than philosophers) to actually assess with empirical tools whether a given robot has moral agency, and the degree to which that agency is conferred from a user with a given perspective.

Fortunately, the third perspective focuses on how to measure a robot's moral agency, centering on perceived moral agency. Its focus is on the human propensity to anthropomorphize robots and how doing so leads to perceptions of moral agency from robots' appearance and behavior. Therefore, researchers operating according to this perspective have developed measures intended to capture human perceptions of moral agency [1]. This perception-focused view of moral agency is especially critical as people start to build attachments and relationships with robots [4, 6]. However, we argue that there still is a key disconnect between these second and third perspectives that merits further interrogation.

First, unlike the work of Floridi and Sander [5], current work within the third perspective does not ground moral agency in the *machine*-agency components, such as autonomy, interactivity, and adaptability. Rather, the assessment of a robot's moral agency is grounded in how moral agency of humans or at least other entities that can readily be classified into one of the existing categories of being. Thus, this perspective omits the question of whether people's perception of agency, which generally relies upon their perception of a human's agency, is also applicable to evaluating agency of robots. We view that it is critical to directly address this question. This is because, if robots are entities that belong to a novel ontological category [7], it is unclear whether the same cognitive processes involved in formulating perceptions of humans' agency could and should be involved in formulating perceptions of robots' moral agency. For instance, to offer an example that describes a robot as having moral agency, would it be valid or necessary to say, "The robot *feels* that it ought to help children in need of help"? We think that the use of the word *feeling* may rather divert human observers' attention from accurately assessing the robot's moral agency as a robot's capacity to *feel* something may not not essential for it to have moral agency. By contrast, describing a person as *feeling* that she ought to help children in need of help, does not raise these questions about validity and necessity.

Second, while we agree that the perceptions that lead to a recognition of a robot's moral agency are critical to human-robot interaction, especially in aspects of social cognition [1], we think that it is necessary to approach moral agency of robots in a fashion different from the anthropocentric approach. As Floridi and Sander [5] pointed out, there has not been a consensus over how to define moral agency even for humans. This suggests that it may be crucial to establish novel approaches that are customized for assessing moral agency of the novel being, that is robots. Instead of building these approaches based on the assumption that moral agency of robots is a psychological construct (i.e., a human mental construction) that people can directly perceive and observe as a whole, we support Floridi and Sander's [5] approach based on the assumption that moral agency is an ontological state and humans only directly observe, perceive, and model the constituent components of moral agency in artificial entities. Therefore, when assessing robots' moral agency, we think that evidence for these perceptions of subcomponents of moral agency must be separately collected and combined together to evaluate the state of a robot's moral agency. We elaborate on this idea in the next section.

3 DERIVING ROBOT MORAL AGENCY FROM FOUR CONSTITUENT SUBCOMPONENTS

One of the most influential conceptualizations of artificial moral agency consistent within the second perspective is that presented in Floridi and Sanders [5], where moral agency is said to be composed of discrete morality and agentic dimensions. Floridi and Sanders [5] maintains that an artificial entity can qualify as an moral agent when it demonstrates the observable meeting the following criteria: capacity for (im)moral action, autonomy, interactivity, and adaptability. Their account differs from others in three critical ways which have important implications for successful human-machine interaction and collaboration.

As discussed in the previous section, first, this account critically distinguishes moral agency as an ontological state argued on the basis of evidence rather than a psychological construct that is directly accessible by human perceivers. This means that rather than viewing moral agency as something that is directly perceivable by a particular human, they instead posit that any claim of moral agency can be justified on the basis of what can be observed from the perspective of a particular user, i.e., at that user's level of abstraction [5]. Framing moral agency as a psychological construct could be problematic in the context of evaluating a robot's moral agency because it would rely on the yet-to-be-verified assumption that humans have an intuitive understanding and overall mental conceptualization of the degree to which a robot is a moral agent. This approach to assessing robot moral agency can increase ambiguity in understanding a robot's moral agency. For example, superficial changes in features of robot designs and behaviors would likely to sway the degree to which such agency is ascribed to the robot from the human mind. An ontological framing, on the other hand, would suggest that (most) humans do not have an intuitive understanding of the philosophical concept of robot moral agency, thereby supporting the assessment of robot moral agency grounded in a theory-driven philosophical framework.

Second, and importantly, the agentic dimension of moral agency is not entirely dependent on humans' perception of the agent, but is instead grounded in the observable behaviors of the actor — i.e., whether the actor appears to have the capability to take actions that are not directly dependent on external stimuli (autonomy), whether the actor appears to have the capability to take actions in response to intentional stimulation (interactivity), and whether the actor appears to have the capability to change its autonomous and interactive behaviors over time (adaptability). The degree to which each of these three components may vary but, when combined together with capacity for morally relevant action, can be used to determine the moral agency of a robot.

Finally, this framework proposes an ontological framing of moral agency, which suggests the need for the logical combination of evidence in favor of each of the proposed dimensions of moral agency for artificial agents. That is, if an actor is clearly autonomous but clearly not interactive, or clearly incapable of morally consequential actions, then it is definitely not a moral agent, regardless of how autonomous it is. In contrast, from a psychological construct perspective, conferring moral agency to a machine would specifically rely on overall perception of its morally agentic status.

We think that research methods grounded in psychology and computer science can help realize Floridi and Sanders's [5] theoretical framework in a concrete form of scales measuring and quantifying robot moral agency. Combining the framework with scales for measuring the constituent components of moral agency also combines together the second and third perspectives on moral agency in the literature.

Admittedly, there are lingering questions about whether, in applying Floridi and Sanders's approach, moral agency should ultimately be regarded as a continuous spectrum or as following the all-or-nothing principle. For instance, can the state of robots' moral agency vary in degrees ranging from very low to very high? Otherwise, is it a state that either can or cannot be assigned to robots? Also, it is open to discussion whether it would be possible and necessary to rule out any consideration of the human-like mind, such as intentionality, feelings, and freewill, in measuring each of the four subcomponents of robot moral agency. Despite these questions, in our view, the philosophical framework proposed by Floridi and Sanders [5] sets up a solid theoretical groundwork for creating psychological and computational methods to assess robot moral agency.

4 POINTS FOR FUTURE CONSIDERATION

Based on this argument, we pose the following points for discussion for HRI researchers interested in robot moral agency.

- Is robot moral agency a cognitive construct or an ontological state?
- What are the implications of regarding moral agency as an ontological state rather than a cognitive construct?
- Are interactivity, autonomy, and adaptability sufficient and necessary conditions to determine robot agency?
- Are moral competence and positive moral dispositions needed for moral agency?
- Do the four components of moral agency (capacity for moral action, interactivity, autonomy, and adaptability) equally contribute to the assessment of robot moral agency?
- Does the contribution of these components to moral agency operate in a strictly logical fashion (i.e., failure to satisfy one criterion is sufficient to rule out moral agency?)
- What down-stream effects are dependent on moral agency? For example, would a quantification system for moral agency help us to predict when users will blame and punish robots for their bad actions?
- What down-stream effects are dependent on moral competence or perceived positive moral dispositions *rather than*, or *in addition to*, moral agency?

5 CONCLUSIONS

In this paper, we presented our understanding of the existing perspectives of moral agency in the context of HRI and discussed our views on each perspective in terms of potential relevance and contribution to serving the goal to assess moral agency of robots. We view that Floridi and Sander's [5] theoretical framework has a strong relevance to evaluating a robot's moral agency and a distinct potential to contribute to discourses about robot moral agency. That said, however, we acknowledge that these views are open to change as technologies enabling development of robots continue to advance and people's perceptions of robots evolve in tandem with the advancement. More perspectives on robot moral agency are bound to enter into discussions in the literature. Although it remains to be answered how much of complex social roles robots would be able to take in future society, we expect that the need to find ways to adequately represent and assess moral agency of robots is important. Even if robots continue to be a category of being that cannot be regarded as an independent agent and do not receive punishment or praise for their actions, determining moral agency of robots can be essential because it can help people ultimately derive judgments of distribution of punishment and responsibility among humans who can be developers, investors, users, etc. Therefore, we encourage the HRI community to keep active interdisciplinary discussions about robot moral agency open.

ACKNOWLEDGMENTS

This workshop is partially supported by the Air Force Office of Scientific Research Award Number FA9550-23-1-0036.

REFERENCES

[1] 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.

[2] Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8 (2010), 63–74.

[3] Mark Coeckelbergh. 2022. *Robot ethics.* MIT Press.

[4] Kate Darling. 2015. 'Who's Johnny?'Anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015). ROBOT ETHICS* 2 (2015).

[5] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14, 3 (2004), 349–379.

[6] David J Gunkel. 2012. The machine question. *Critical perspectives on AI, robots, and ethics* (2012), 5.

[7] Peter H Kahn Jr, Aimee L Reichert, Heather E Gary, Takayuki Kanda, Hiroshi Ishiguro, Solace Shen, Jolina H Ruckert, and Brian Gill. 2011. The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction.* 159–160.

[8] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 10–18.

[9] Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, et al. 2023. Moral psychology of nursing robots: Exploring the role of

robots in dilemmas of patient autonomy. *European Journal of Social Psychology* 53, 1 (2023), 108–128.

[10] Bruno Latour. 2012. *We have never been modern*. Harvard university press.

[11] Gert-Jan Lokhorst and Jeroen Van Den Hoven. 2012. Responsibility for military robots. *Robot ethics: The ethical and social implications of robotics* (2012), 145–156.

[12] Bertram F Malle, Stuti Thapa Magar, and Matthias Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being* (2019), 111–133.

[13] Eileen Roesler, Dietrich Manzey, and Linda Onnasch. 2021. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics* 6, 58 (2021), eabj5425.

[14] Robert Sparrow. 2007. Killer robots. *Journal of applied philosophy* 24, 1 (2007), 62–77.

[15] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. 2021. Comparing strategies for robot communication of role-grounded moral norms. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 323–327.