Perspectives on Moral Agency in Human-Robot Interaction

Boyoung Kim
bkim55@gmu.edu
George Mason University
Fairfax, Virginia, USA

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

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

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

ABSTRACT

Establishing when, how, and why robots should be considered moral agents is key for advancing human-robot interaction. For instance, whether 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 whether their agency in turn triggers social and moral cognitive and behavioral processes in humans. Robotic moral agency also has significant implications for how people should and do hold robots morally accountable, ascribe blame to them, develop trust in their actions, and determine when these robots wield moral influence. In this workshop on Perspectives on Moral Agency in Human-Robot Interaction, we plan to bring together participants who are interested in or have studied the topics concerning a robot's moral agency and its impact on human behavior. We intend to provide a platform for holding interdisciplinary discussions about (1) which elements should be considered to determine the moral agency of a robot, (2) how these elements can be measured, (3) how they can be realized computationally and applied to the robotic system, and (4) what societal impact is anticipated when moral agency is assigned to a robot. We encourage participants from diverse research fields, such as computer science, psychology, cognitive science, and philosophy, as well as participants from social groups marginalized in terms of gender, ethnicity, and culture.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Robot; Moral agency; Moral agent; Human-Robot Interaction; Assessment

ACM Reference Format:

Boyoung Kim, Elizabeth Phillips, Tom Williams, and Qin Zhu. 2023. Perspectives on Moral Agency in Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3568294.3579966

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3579966

1 BACKGROUND

Robots were once characterized only as passively carrying out dull, dirty, or dangerous tasks assigned by humans. Nowadays, however, 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 people's everyday life will grow in kind.

Certain domains that target vulnerable populations, such as in older adult care, childcare, and education, robots' actions can have clear and immediate impacts, which can either be negative or positive, on their human interactants. These possibilities raise several ethical questions, especially concerning moral accountability. For example, if a robot spills hot water on a child at a daycare center, would it be reasonable to hold the robot accountable for the child's injury and punish the robot? Otherwise, regardless of whether it is rational or not, would people be prone to viewing the robot as being responsible for its action? To answer these questions, it is necessary to discuss the role of robots as moral actors.

In response to this demand, many Human-Robot Interaction (HRI) and machine ethics researchers have discussed how robots may play possible roles as moral actors in society by, for example, making decisions that entail moral consequences [5, 7] and taking responsibility for their choices [6, 8]. However, describing precisely when and how robots will exert moral influence is challenging. To illustrate, some research showed evidence that robots can influence humans in both morally positive and morally negative ways [1, 2, 9], but other research where robots were explicitly designed to exert moral influence on people failed to do so at all [4].

Assuming that robots occupy a new ontological status, fundamentally distinct from humans [3], it would be important to specify the moral agency of these novel agents that determines their moral influence on people. Robots can be built via collaborative efforts among researchers across diverse disciplines, including but not limited to computer science, robotics engineering, psychology, cognitive science, and philosophy. Thus, to understand moral agency of robots that may enable them to have moral influence on people, researchers from diverse backgrounds should exchange their perspectives and have discussions about robots as moral agents.

In this workshop, therefore, we seek to bring together participants who are interested in or have studied topics concerning a

robot's moral agency and its impact on human behavior. We will gather researchers with diverse perspectives to discuss (1) which elements should be considered to determine the moral agency of a robot, (2) how these elements can be measured, (3) how they can be realized computationally and applied to the robotic system, and (4) what societal impact is expected when moral agency is assigned to a robot. We encourage participants from diverse backgrounds defined by research fields, such as computer science, psychology, cognitive science, information and communication science, machine ethics, and philosophy. We also welcome innovative perspectives enabled by diversity and uniqueness in participants' identity (e.g., gender, racial, ethnic, and cultural diversity).

2 TARGET AUDIENCE

Our target audience will be comprised of researchers from moral psychology, moral philosophy, machine ethics, and other related areas who have interests in the moral agency of robots. This is an active and growing research area as suggested by 2,200 papers that appear in Google Scholar with search words like moral agency and robot for the period between 2019 and 2022.

Depending on the number of submissions, we will adjust the presentation type (e.g., switching from individual presentations to poster sessions) to best fit the time allotted while simultaneously allowing enough time for discussion among the group.

3 TOPICS OF INTEREST

- Theoretical definition of robot moral agency
- Antecedents of moral agency
- Measurement tool of robot moral agency
- Computational models for robot moral agency
- Implications of moral agency for HRI
- Downstream societal implications of robot moral agency

4 PLANS FOR DOCUMENTATION

We plan to archive the papers on a publicly accessible online website for future reference. Once their papers are accepted, authors will have a choice to opt-out from this plan.

5 FORMAT AND ACTIVITIES

We plan to hold a hybrid, half-day (4 hrs) workshop that consists of the following activities: (1) an invited academic keynote, (2) group discussions, and (3) lightning talks for extended abstracts and talks for short papers. We will adjust the time assigned to each talk according to the number of submitted and accepted papers.

We will use various forms of activities to facilitate brainstorming and discussions, such as breakout and group discussions. As we would like to encourage wide attendance and enhance accessibility, we will include the virtual attendance option.

In Table 1, we present a sample schedule for the workshop.

Submission Type We are interested in theoretical, experimental and computational work that has relevance to robot moral agency. We welcome papers on these topics that adhere to the following format guidelines:

- **Extended abstracts:** up to 2 pages excluding references
- **Short papers:** 3-4 pages excluding references

Table 1: Tentative Schedule

Time	Activity
9:00 am	Opening remarks (15 min)
9:15 am	A keynote speaker's talk (30 min)
9:45 am	Q and A (15 min)
10:00 am	Break (15 min)
10:15 am	Lightening talk (45 min)
11:00 am	Small group discussion (45 min)
11:45 pm	Group discussion (60 min)
12:45 pm	Closing remark (15 min)

6 ORGANIZING TEAM

This workshop will be organized by a team of researchers with backgrounds in psychology, philosophy, computer science, and engineering education. The diversity in this team's research background is expected to facilitate recruitment of participants from a variety of research backgrounds.

Boyoung Kim, bkim55@gmu.edu, is a postdoctoral researcher in the Department of Psychology and the Center for Security Policy Studies at George Mason University. Her research addresses the issues related to ethics and morality in the contexts of human-human interaction and human-robot interaction. Boyoung has co-organized the AI-HRI symposium as a part of the AAAI 2022 Fall symposium series.

Elizabeth Phillips, ephill3@gmu.edu, is an assistant professor in the Department of Psychology, Human Factors and Applied Cognition Group at George Mason University, where she directs the Applied Psychology and Autonomous Systems (ALPHAS) Lab. She studies how we can design robotic and autonomous systems to be better partners, teammates, and companions for people in the near future. She has previously served as an organizing member of the Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI) workshop at 'HRI 2019 and as a co-organizer of the 2017, 2018, and 2019 Meet the Women in Robotics Workshop at the Robotics: Science and Systems Conference.

Tom Williams, twilliams@mines.edu, is an associate professor in the Department of Computer Science, at Colorado School of Mines, where he directs the MIRRORLab. He studies language-based human-robot interaction, including the moral competence of these robots. He has previously served as an organizing member of multiple HRI workshops, including the VAM-HRI series and the workshop on the "Dark Side of HRI". He is a member of the HRI steering committee and a program committee co-chair for HRI 2024.

Qin Zhu, qinzhu@vt.edu, is an associate professor in the Department of Engineering Education at Virginia Tech. By drawing on theories from Confucian ethics, his work examines how human teammates' experience interacting with robots affect their self-reflections and moral development. He is currently working on how people from different cultural and educational backgrounds perceive and prioritize AI ethical issues. He is serving on the Executive Board of the Society for Ethics across the Curriculum. He has experience organizing workshops on ethics education assessment and how to use assessment tools for pedagogical purposes.

REFERENCES

[1] Gordon Briggs and Matthias Scheutz. 2014. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics* 6, 3 (2014), 343–355.

[2] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 401–410.

[3] 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.

[4] 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.

[5] Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, et al. 2022. Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology* (2022).

[6] 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.

[7] 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. In *Robotics and well-being*. Springer, 111–133.

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

[9] 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.