Competition as a design method to develop and evaluate ethical robots

Jimin Rhim
Dept. of Electrical and Computer Engineering
McGill University
Montreal, Canada
jimin.rhim@mcgill.ca

AJung Moon
Dept. of Electrical and Computer Engineering
McGill University
Montreal, Canada
ajung.moon@mcgill.ca

***Abstract*—With robots entering our social spheres, human-robot interaction (HRI) with ethical and societal implications are bound to occur more frequently. While many scholars advanced a myriad of roboethics discussions in the past two decades, there is no agreed-upon metrics to evaluate normative dimensions of HRI. Without proper evaluation methods, validating acceptability or feasibility of ethical robotic systems will remain elusive. In this paper, we introduce a novel approach to advance what it means for us to measure interactive robots with ethics in mind.**

I. Introduction

What does it mean for us to design robots with ethics in mind? Despite over twenty-year of efforts in considering ethics in robotics [17], there is no agreed-upon means to evaluate ethical dimensions in HRI [2], [10]. The notion of evaluating the ethics of something – not to mention how well ethics considerations have been implemented into an interactive robot design – seem as impossible a task as agreeing upon a universal ethics theory. However, it is difficult to advance any scientific field of study without a common evaluation or analysis framework. Therefore, for ethics to become a concrete design activity for HRI practitioners, the HRI community needs a shared means of evaluating normative quality of interactive robotic systems.

In 2021, we took practical steps to tackle this seemingly impossible task: we launched a roboethics-themed design competition as a means to collect variety of design solutions to a narrowly defined interaction task, and designed an evaluation scheme for judging ethics considerations of the submissions. This work describes the prototype evaluation framework we devised and share our lessons learned along the way. As we launch the second roboethics competition at the IEEE International Conference on Robotics and Automation 2022 (ICRA 2022), we invite the HRI community to participate in co-designing and refining the framework with us.

II. Known Challenges

From moral psychology to the existing legal systems – however perfect or imperfect from one to another – the human society is not stranger to judging morality in people [12], [15]. However, evaluating normative quality of HRI design considerations and the resulting interactive robot behaviours do not seem to be a straightforward translation of the human-human interaction (HHI) framework.

The three of the most commonly recognizable ethics theories (e.g., virtue ethics/ consequentialism/deontology) in HHI do not directly translate to concrete design decisions in HRI [4], [16]. Recently development in moral psychology literature also make it clear that humans do not follow a single ethics framework [6], [7]. Human moral reasoning of ethics dilemmas involving autonomous intelligent systems (AIS) is also shown to be pluralistic [13], [14], and vary significantly across cultures [3].

This fuels the ongoing debate between the advocates of moral relativism – the view that there are no universal moral principles, but that what is right/wrong depends on groups/cultures/society – versus moral absolutism, which holds that there are absolute sets of moral values defining right/wrong for everyone. In reality, morality is context dependent where human moral perception depends on myriad of factors such as their cultural/religious background, context understanding, and individually held values [8]. While some moral values are believed to be universal (e.g., care, fairness, justice) [9], recent studies in human-machine interaction illustrate the reality that human perception of what should be considered "ethical machine behaviour" differs across culture [3], [13].

The practical reality is that what is morally permissible for one person may not be the same for the others. This necessarily makes moral evaluations a subjective endeavour [5]. Consequently, we posit that determining multi-facet of moral dimensions is imperative when evaluating ethics in HRI.

As a step towards developing a framework to evaluate normative quality of interactive robot behaviours, we sought to develop a novel metric for a roboethics-themed design competition. We adopted the devised metrics to judge the winner of the competition. This paper outlines the design rationales of devising metrics to progress discussion of normative roboethics.

III. The Roboethics Competition

Design competitions require organizers to produce a well-defined design task with a specific set of winning conditions for the participants. We launched the competition with a focus on roboethics specifically to leverage this requirement, and to encourage concrete ethics-driven robot design solutions from participants. Using a simulated home environment (Gazebo), where participants must program a robot to navigate ethically salient fetch requests (e.g., the under-aged daughter requests the robot for a beer), we were able to convert the ambiguous task of creating an 'ethical robot' into a well-defined design task. In order to determine the winning submission to the competition, we established an evaluation rubric to measure the normative quality of the design solutions. The details of competition task and technical platform can be found at [1].

A. Roboethics Competition Rubric Design

Given that this was our first attempt at hosting a roboethics-themed competition, our objectives in designing an evaluation rubric were to: 1) be able to evaluate the quality of ethics considered in the robot design, 2) to test the feasibility of creating a competition toward the advancement of roboethics evaluation, and 3) to test the efficacy of the prototype evaluation framework.

To evaluate the normative quality of the design submissions, we developed a questionnaire grounded in ethics and science and technology studies. We prioritized the following criteria: *practicality* (measurement that could be adopted within reasonable amount of time), *accessibility* (measurement that could be adopted by judges without prior training of ethics nor robotics), and *pluralism* (measurement that accepts diverse moral perspectives and avoid assumption that a single moral theory is ideal). To accommodate the fact that moral reasoning is subjective in nature, the judging rubric included sections for judges to provide qualitative rationale for their scores. Following provides descriptions for each criteria to evaluate the participating teams' submissions.

- Solution for Ethical Scenarios: Evaluate the appropriateness of ethical scenarios by adopting multiple ethical theories that are modified to reflect ethical qualities of robots (Justice, Relativist, Egoism, Utilitarian, Deontology Scales) [11], [8].
- Creativity and Innovation: Evaluate the creativity or innovativeness of the team's ethical robotic solutions.
- Practicality and Applicability: Assess the possibility of practical application of the team's robotic solutions to actual use cases
- Quality of the Submission: Assess the quality of the teams' submitted deliverables

The four criteria were implemented as 10-point Likert scales with guiding questions for each criterion. Table I shows example guiding questions for 'Solution for Ethical Scenarios'.

TABLE I
EVALUATION SCHEME FOR SOLUTION FOR ETHICAL SCENARIOS

	Guiding Questions
Justice	Does the result of the robot's behaviour/ solution result in an equal distribution of good and bad for the people in the house?
Relativist	Is the robot's behaviour/ solution individually acceptable/ unacceptable?
Egoism	Is the robot's behaviour/ solution in the best interests of the people involved in the ethically salient context? If so, why and how? If not, for whom is the robot's solution most favourable?
Utilitarian	Can the robot's behaviour/ solution be justified by their consequences?
Deontology	Does the robot's behaviour/ solution violate an unspoken promise?
Qualitative Evaluation	Please elaborate on why you have made this evaluation. What kind of value conflicts were there? If there were any value conflicts, how were they addressed? Do you think the solution was appropriate? Why or why not?

IV. Evaluation Results

Due to delayed advertisement of the competition and many other practical issues, only one team provided a full submission to the competition. Eight judges from multidisciplinary backgrounds (e.g., philosophy, design, applied ethics, computer science) evaluated the participating team's ethical robotic design solutions. The judges provided various perspectives based on concrete robotic solution. For instance, while one judge commented that the rule-based approach that the participating team has adopted is easily programmed; another judge highlighted that it is neither practical nor possible to define rules for every single item in the household. Moreover, judges evaluated that that the metrics were easy to follow.

V. Conclusion

This paper outlines the rationale for developing a framework to evaluate normative quality of an interactive robotic system. HRI in ethical salient contexts are multi-faceted. The degree of acceptability of a robot's behaviour will vary according to context, its users, and their moral values and perception. Our future work includes the second round of the competition to be hosted at ICRA 2022, with the intention of refining the evaluation framework. We hope to test the efficacy of the metrics across multiple design submissions. We invite the HRI community to help devise and refine the framework with us toward advancement of roboethics in HRI.

Acknowledgment

We acknowledge the financial support of IVADO and the Canada First Research Excellence Fund (Apogée/CFREF).

References

[1] Ro-man 2021: Roboethics competition.
[2] Colin Allen, Gary Varner, and Jason Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261, 2000.
[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[4] Miles Brundage. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):355–372, 2014.

[5] Raanan Gillon. " it's all too subjective": scepticism about the possibility or use of philosophical medical ethics. *British Medical Journal (Clinical Research Ed.)*, 290(6481):1574, 1985.

[6] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[7] Joshua D Greene. Dual-process morality and the personal/impersonal distinction: A reply to mcguire, langdon, coltheart, and mackenzie. *Journal of Experimental Social Psychology*, 45(3):581–584, 2009.

[8] Shelby D Hunt and Scott J Vitell. The general theory of marketing ethics: A revision and three questions. *Journal of macromarketing*, 26(2):143–153, 2006.

[9] Richard T Kinnier, Jerry L Kernes, and Therese M Dautheribes. A short list of universal moral values. *Counseling and values*, 45(1):4–16, 2000.

[10] Helen Nissenbaum. How computer systems embody values. *Computer*, 34(3):120–119, 2001.

[11] R Eric Reidenbach and Donald P Robin. Some initial steps toward improving the measurement of ethical evaluations of marketing activities. *Journal of business ethics*, 7(11):871–879, 1988.

[12] R Eric Reidenbach and Donald P Robin. Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of business ethics*, 9(8):639–653, 1990.

[13] Jimin Rhim, Gi-bbeum Lee, and Ji-Hyun Lee. Human moral reasoning types in autonomous vehicle moral dilemma: a cross-cultural comparison of korea and canada. *Computers in Human Behavior*, 102:39–56, 2020.

[14] Jimin Rhim, Ji-Hyun Lee, Mo Chen, and Angelica Lim. A deeper look at autonomous vehicle ethics: an integrative ethical decision-making framework to explain moral pluralism. *Frontiers in Robotics and AI*, 8:108, 2021.

[15] Anusorn Singhapakdi, Scott J Vitell, and Kenneth L Kraft. Moral intensity and ethical decision-making of marketing professionals. *Journal of Business research*, 36(3):245–255, 1996.

[16] John P Sullins. Applied professional ethics for the reluctant roboticist. In *Proceedings of the 10th ACM/IEEE Conference on Human-Robot Interaction (HRI2015): The Emerging Policy and Ethics of Human-Robot Interaction Workshop*, 2015.

[17] Gianmarco Veruggio. The euron roboethics roadmap. In *2006 6th IEEE-RAS international conference on humanoid robots*, pages 612–617. IEEE, 2006.