A Robot Ethics Dilemma: To Side with Folk Morality or the Experts?

Our research team has been investigating methods for enabling robots to behave ethically while interacting with human beings. Our approach relies on two main sources of data for determining what counts as "ethical" behavior. The first are the views of average adults, which we refer to "folk morality", and the second are the views of ethics experts. Yet the enterprise of identifying what should ground a robot's decisions about ethical matters raises many fundamental metaethical questions. Here, we focus on one main metaethical question: would reason dedicate that it is more justifiable to base a robot's decisions on folk morality or the guidance of ethics experts? The goal of this presentation is to highlight some of the arguments for and against each respective point of view, and the implications such arguments might have for the endeavor to encode ethical decision-making processes into robots.

Programming Ethics into Robots

There are many ongoing efforts to program ethical decision-making capacities into robots. The strategies for attempting to do so are varied and multifaceted (e.g., Hendrycks et al. 2021; Kim et al. 2021; Tolmeijer et al. 2021). The particular strategy our team is exploring is to have robots make decisions that approximate what either an average adult or ethics expert would recommend as being appropriate. To inform this approach, our team first surveyed average adults and then ethics experts. Each of the two surveys asked participants to evaluate the appropriateness of a series of actions in connection with two scenarios; the first scenario involves playing a boardgame with a child and the second involves teaching an older adult a pill sorting task. An overarching theme across both surveys is to ascertain the ethical appropriateness, and limits, of using deception as a means of motivating someone to perform an act that is intended to benefit them. The research approach revisits longstanding metaethical questions and debates that warrant exploration, especially within the context of human-robot interaction.

Folk Morality

Among the challenges posed by relying on "folk morality" as a foundation for robot behavior is the fallacy of common practice. In other words, even though "most" people might perform a behavior, and perhaps even provide overt verbal or other indications of the behavior's ethical appropriateness, the behavior is not necessarily ethical. For example, merely because many or most drivers go beyond the legal speed limit on highways, it may not logically follow that the behavior is ethical to perform. As philosophers and others have noted for centuries, there can be a disconnect between what most people do or say and what is ethical (e.g., Rachels and Rachels 2019). The challenge here for roboticists is how to parse between what "society" deems to be ethical versus what is merely a common practice. An age-old problem persists regarding how to identify what folk morality is. Underlying this are intractable issues such as whether core ethical values are actually shared across various groups or cultures and whether morality shifts over time.

Expert Morality

Another potential pathway for informing the ethical decisions of robot is to rely on ethics experts. This option of course comes with its own series of difficult challenges. At least since the time of Socrates, philosophical questions have persisted about whether someone can possess expertise in ethics and if so, how to identify an ethics expert reliably. For the purposes of our research project, the proxy we used to make this determination is for the individual to be a philosophy professor who teaches ethics at an academic institution. Admittedly this is an imperfect means for identifying someone who is an expert in ethics. Moreover, even if consensus could be reached that someone is an expert in ethics, it does not necessarily follow that they have expertise on each particular matter at issue, including in the realm of human-robot interaction (or even related to human-human interaction for that matter). Moreover, it is

clear that ethics experts strongly and profoundly disagree on theoretical matters and the application of ethical concepts and principles to practice; a pathway for resolving such disputes is elusive and has been the subject of philosophical debate for centuries.

How Does the Ethical Robot Enterprise Proceed From Here?

After reviewing the case for and against the two aforementioned sources of ethical guidance, the situation for robot ethics might seem fairly bleak. Furthermore, overlapping with these complexities is that a significant number of people, in at least some circumstances, may want to be treated in a way that does not fully conform with strict ethical norms or obligations (for example, preferring comfort to honesty). Thus, should robots always be tasked with behaving ethically or could other goals take priority? Yet amidst these challenges, we propose at least two steps forward.

First, if overlaps between what folk and expert morality dictate can be found, that could increase the probability that an act is ethically appropriate to perform. The most challenging situations arise when no consensus appears to exist. Viewpoints on ethically correct behavior may differ depending on an individual's group membership characteristics. For example, political leanings (Haidt 2012) might influence the evaluation of ethical problems. However, if an ethical problem is so challenging that everyone cannot agree on the "right" course of action, then arguably the robot could pursue a course of action that at least some people evaluating the robot's behavior would deem as being reasonable. Of course, this option is more likely to be defensible if the robot performs a relatively low stakes, low risk behavior.

Second, programming robots may be more manageable if the endeavor is constrained to what is ethically appropriate in narrowly and strictly defined situations such as whether it is okay to deceive another player while playing a boardgame. Well-defined situations, such as those that occur in games, could serve as a training ground for managing more complex interactions involving conflict, intense emotional responses, or deception.

References

Haidt, Jonathan. 2012. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Vintage.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, Jacob Steinhardt. Aligning AI With Shared Human Values. International Conference on Learning Representations (ICLR) 2021. https://arxiv.org/pdf/2008.02275.pdf.

Kim, Boyoung, 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 (HRI '21 Companion). Association for Computing Machinery, New York, NY, USA, 10-18.

Rachels, James and Stuart Rachels. 2019. Elements of Moral Philosophy (9th edition). McGraw Hill.

Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. ACM Computer Surveys, vol. 53, no. 6.

Acknowledgments

This paper is based upon work supported by (removed for review purposes).