A Metaethical Reflection: The Ethics of Embedding Ethics into Robots*

Jason Borenstein, *Affiliate, IEEE*, Ronald C. Arkin, *Fellow, IEEE*, and Alan R. Wagner, *Senior Member, IEEE*

***Abstract*— This paper examines the metaethical dimensions of the computing community's efforts to program ethical decision-making abilities into robots. Arguments for and against that endeavor are outlined along with brief recommendations for the human-robot interaction realm.**

I. Introduction

Our research team is taking part in a multi-year project funded by the US National Science Foundation with the aim of programming humanoid robots to behave "ethically" when interacting with humans. A key focus of the project is examining whether, and when, it might be appropriate for robots to deceive a human if that person might benefit from the deception. The research endeavor has caused us to reflect on fundamental ethical questions about the enterprise of programming robots to behave ethically, a primary overarching one being: is it ethical to attempt to create ethical robots? The purpose of this paper is to outline some key considerations related to this matter and generate conversation within robotics communities.

As a starting point, what specifically does it mean to program ethics into a robot? Some potential goals include:

- Designing the technology so that it operates safely (e.g., an industrial robot in a manufacturing factory that avoids collisions with humans)
- Encoding the technology so that it adheres to formal laws or codes of ethics (e.g., an autonomous vehicle that strictly follows the speed limit)
- Enabling the technology to use moral reasoning so that it makes decisions in a manner similar to (or "better than") a human
- Enabling the technology to interact with humans in a manner consistent with the ethical norms of human-human interaction

There can be overlap among the above goals, but for our purposes here, we will largely concentrate attention on the third and fourth goals. We will primarily focus on physically-embodied social robots; in other words, types of robots that are being designed to serve as companions or friends for humans. However, the discussion may be relevant to other types of robots such as those used in healthcare environments or in the military.

*Research supported by the National Science Foundation as part of the Smart and Autonomous Systems Program under grants #1849068 and 1848974.

Dr. Jason Borenstein is with the School of Public Policy and Office of Graduate Education, Georgia Institute of Technology (email: borenstein@gatech.edu).

II. A Programming Strategy

Our ongoing human-robot interaction (HRI) project aims to create and test an architecture for a robot that allows the system to be adaptable and ethical in its decision-making in complex situations. The two main situations that our research project is focusing on are the (potential) use of deception while playing a boardgame with a child and when teaching an older adult how to sort pills into an organizer. To determine the circumstances under which deception might be appropriate, we used surveys to collect the views of laypersons and ethics experts in response to different versions of game playing and pill sorting scenarios. An overarching theme across the scenarios is whether it might be acceptable to use or allow deceptive behavior. We are relying on the ethical recommendations from survey participants in response to variations of human-human interaction scenarios as a foundation for programming robots. Yet critics might question whether it is necessary, or appropriate, to program some version of ethics into robots at all; something that we begin to explore in the next section.

III. The Potential Case For

Even if a robot is designed to perform "dull, dirty, or dangerous" tasks only, encoding some version of ethics may be necessary unless there is going to be constant human supervision. For example, mobile robots may be useful for public health (e.g., decontaminating an area). However, at a minimum, harm avoidance will be essential and encoding them with a version of ethics will help to more effectively accomplish this along with other goals.

Roboticists are driven to create "ethical" robots for many reasons. Given resource-limitations or other societal challenges, there have been many pushes to create robots to serve as caregivers. Arguably, robots could help, for example, with the frequent lack of individualized attention in nursing homes or with monitoring children, especially when parents need to work. In such circumstances, programming a robot merely to operate safely (in a narrow sense) may be insufficient because the robot would not be able to accomplish important, perhaps essential, goals. It may need to be recognize when a nursing home resident or child is in danger (rather than just avoiding collisions with the person), or it may

Dr. Ronald C. Arkin is with the School of Interactive Computing, Georgia Institute of Technology (email: arkin@cc.gatech.edu).

Dr. Alan R. Wagner is with the Rock Ethics Institute, Pennsylvania State University (email: azw78@psu.edu).

need to seek ways to mitigate a person's loneliness.

Another potential justification is that the process of developing ethical robots might enable us, as humans, to more fully reflect on what it means to be ethical, and the cognitive systems that allow us to do so. An ethical robot could also nudge humans to perform behaviors that are for their own benefit such as prompting them to stop smoking or for the benefit of other people such as by encouraging charitable donations [1],[2]. However, the appropriateness of designing technology with the deliberate intent of modifying a human's behavior is highly contentious and perhaps ethically dubious.

IV. The Potential Case Against

The potential criticisms of the effort to develop ethical robots are manifold. Here, we will focus on three main categories of critiques: the potential lack of ethical consensus; challenges pertaining to the computing community's ability to encode ethics into robots; and how ethical robots might (detrimentally) impact human beings and human-human relationships.

To begin, a profound lack of consensus/agreement persists on what is ethically right even if the scope is limited to a specific group, culture, or context. This is reflected in that human beings are very far away from universally embracing any particular ethical theory (although despite of this, a "correct" theory could still exist). Also competing ethical theories and frameworks may lead to prioritizing different considerations and to different outcomes. For example, there are many ongoing attempts to design social robots so they can assist children with educational or other needs; yet adults can have sharply diverging parenting philosophies (e.g., stricter versus more lenient). What follows is there may not be one "right" way of having a robot interact with a child even in cases when trying to teach the child how to share or play a game. Moreover, rigidly or strictly applying ethical rules in this context would fail to acknowledge individual differences in children (or people in general). At times, being able to act ethically towards another person would require a detailed profile of that individual along with an assessment of their current emotional state.

Even if agreement could be reached about what is ethical, significant skepticism persists about the computing community's ability to emulate the various facets of human (moral) reasoning successfully. To some, perhaps significant, degree the endeavor would be like running an experiment on human beings. Much trial and error will be needed. And the more advanced a robot's "reasoning" approach is, the more unpredictable the robot's behavior might become. What is observed in the lab might not match what the robot does when interacting with diverse sets of humans and the conditions are more dynamic (e.g., Microsoft's Tay chatbot [3]).

Also, societal values change, and any ethical robot may need to adapt to evolving values and ethical norms, both over time and with respect to the humans around them. Thus, embedding ethics within a robot assumes some form of continual learning on the part of the robot, or at least the flexibility to adjust to changing values and ethical norms. Moreover, fundamental assumptions about the design enterprise warrant examination. For instance, is it necessary for humans and robots to arrive at the same ethical conclusion or behave in the same exact manner? There may be compelling reasons for humans to prioritize family over others but should a robot operate in a similar way? How does context affect moral decisions? Should it?

And of course, a broad range of concerns will emerge regarding how ethical robots might alter human well-being. Human thinking and action, especially in young children, may come to be shaped by and imitate the robots around them. Some humans may come to forgo their own ethical analyses and defer to robots for weighty decisions, or at the least, allow an ethical robotic advisor to be part of their lives. A significant reason for concern is that humans may overtrust robots including in simulated emergency situations [4]. In addition, assuming an ethical robot can be created, should it be allowed to usurp control from humans, and if so, under which circumstances? Should it prevent a person from committing self-harm or restraining a person from harming others? Who (or what) should have the ultimate authority to determine what is a "better" or "more ethical" decision?

The erosion of human-human relationships may follow as well; at least some humans may prefer the company of robots, especially if robots are perceived as being "more ethical". The adoption of ethical robots could also reinforce stereotypes or other problematic beliefs on a large-scale. For example, problematic forms of bias have already been frequently detected within AI technology, including in the criminal justice system [5] and in facial recognition [6].

Furthermore, the enterprise of programming robots may be too consumed with developing a technical fix; it may be symbolic of a diminished commitment to improving human-human relationships [7],[8]. Technical fixes, in the best case, may temporarily mitigate a lingering problem [9] (e.g., an AI system that monitors racist or sexist language on social media). Yet truly solving the underlying problem may require a meaningful change in human attitudes and behaviors.

V. Conclusion and Recommendations

The robotics community is in the process of determining whether it is feasible to create an ethical robot *but whether it should*, more importantly, must also be considered. In that context, our research team is seeking to outline some of the key considerations regarding the metaethics of the robot ethics enterprise. Arguably, it may be justifiable to program robots to adhere to formal laws or codes (e.g., from IEEE [10], clinical manuals [11], or international protocols [12]). Yet at the present time, there are too many technical and non-technical concerns to deploy an "ethical" robot into the world that tries to reason like a human.

To mitigate at least some of the aforementioned ethical concerns, such as bias embedded in a robot's design, a more diverse range of people, such as those with disabilities and other historically underrepresented groups, should be directly

and consistently involved in design, deployment, and use decisions about social robots. More extensive partnerships between roboticists and citizen groups, along with other stakeholders, should be considered.

REFERENCES

[1] J. Borenstein and R. C. Arkin, "Nudging for Good: Robots and the Ethical Appropriateness of Nurturing Empathy and Charitable Behavior," *AI & Society: Journal of Knowledge, Culture and Communication*, vol. 32, No. 4, pp. 499-507, Nov. 2016, doi:10.1007/s00146-016-0684-1.

[2] J. Borenstein and R. C. Arkin, "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being," *Science and Engineering Ethics*, vol. 22, no. 1, pp. 31-46, February 2016.

[3] S. Kleeman, "Here Are the Microsoft Twitter Bot's Craziest Racist Rants," *Gizmodo*, March 24, 2016, https://gizmodo.com/here-are-the-microsoft-twitter-bot-s-craziest-racist-ra-1766820160.

[4] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of Robots in Emergency Evacuation Scenarios." In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 101-108, March 2016.

[5] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[6] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research*, vol. 81, pp. 1-15, 2018.

[7] R. Sparrow, "The March of the Robot Dogs," *Ethics and Information Technology*, vol. 4, pp. 305–318, 2002. https://doi.org/10.1023/A:1021386708994.

[8] N. Sharkey and A. Sharkey, "The Crying Shame of Robot Nannies: An Ethical Appraisal." *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, vol. 11, no. 2, pp. 161–190, 2010.

[9] A. M. Weinberg, "Can Technology Replace Social Engineering?" *Bulletin of the Atomic Scientists*, vol. 22, no. 10, pp. 4-8, 1966.

[10] IEEE, *IEEE 7010-2020: Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*, 2020, https://standards.ieee.org/ieee/7010/7718/.

[11] J. Shim, R. C. Arkin, and M. Pettinati, "An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationship: Implementation and Evaluation," *Proceedings of ICRA 2017*, Singapore, May 2017.

[12] R.C. Arkin, *Governing Lethal Behavior in Autonomous Systems, Chapman and Hall Imprint*, Taylor and Francis Group, Spring 2009.