Applying the Capability Approach to the Ethical Design of Robots

Naomi T. Fitter
Haptics Group, GRASP Laboratory
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA, 19104, USA
nfitter@seas.upenn.edu

Philip M. Nichols
Legal Studies and Business Ethics
The Wharton School
University of Pennsylvania
Philadelphia, PA, 19104, USA
nicholsp@wharton.upenn.edu

ABSTRACT

Relatively few roboethicists have focused on a capability approach as the mainstay of their discussions of robot policy formation. We build up to a capability-focused ethical framework by considering the definition of morality and expert opinions on fairness theory. Such a thought experiment helps with the shaping of an appropriate future of robot technology, as well as the programmatic instillment of ethical policy in robots. An evaluation by a business ethicist and a roboticist working together leads to the conclusion that ethics philosophy can yield notable and cogent ideas for future applications of robotics.

1. BACKGROUND

Robots need ethics and policy; it is generally agreed that robots and their makers should not act completely randomly, although opinions on how they should act vary from rejecting all future robotics research to mimicking many different types of ethical models inspired by distinct philosophical and ethical works. Deciding on a consistent approach to roboethics is a challenging problem. Some scholars within robotics have used Bayesian utilitarianism, which fits well with many other robot problem-solving strategies [2]. The definition of any ethical framework like this, however, requires the researcher to make several critical decisions, such as choosing the weight of importance of the individual, future individuals, animals, etc. This often leads to a somewhat arbitrary result that may be hard to defend as moral. Early roboethicists recognized the complexity of this issue and considered creating a superintelligent computer or robot to try to solve this problem, since human intellect has so far failed to conquer it completely [2].

John Rawls asks a question that is more focused than a simple inquiry of morality's general principles. Rawls asks how society should fairly distribute the goods, services and other things of value that it creates. This question is of special relevance to robotics, which might be on the verge of astonishing contributions and generation of value. Rawls does not directly address robotics, but he does offer a general rule: the inherent respect that is due to each person mandates that distribution should be made in a way that is generally recognized as fair [5].

Amartya Sen offers a solution that is more directed and of more use to a field such as robotics. Sen, as Rawls, respects the autonomy of individuals but does not yield to the temptations of simply satisfying raw preferences. Instead, Sen suggests that the imperative of improving overall general welfare requires that goods and services and other things of value be distributed in the manner that most enhances individuals' capabilities [7]. Sen understands life as sets of doing and being which together constitute functionings. The quality of life, therefore, depends on the capability to function in many ways at high levels. This quality does not depend on actually functioning - an autonomous person may choose what she wants to do and be - and thus the mere distribution of goods or services itself does not enhance welfare. Instead, goods and services must be distributed in the way that enhances the capability to function and that gives each individual the widest choice of doings and beings [7].

2. APPLICATION

The application of the ideas of Rawls and Sen to robotics and computer science is not brand new, but previous efforts have either neglected parts of the fairness theory or have involved only ethicists and policy-makers without roboticists as part of the conversation. Although they have not yet applied the theory to physical robots, economists and computer scientists have attempted to apply some Rawlsian principles in game theory problems, including the development of virtual robotic agents [4]. Although these game theory frameworks exist, some Rawlsian principles were neglected in their creation. Among these are stability considerations and the accurate definition of goods, items that Rawls discussed in his works but which are more difficult to quantify [4]. Improving this game theory approach might require more synergistic collaboration between roboticists and ethicists, but it is critical that more interactions of this sort begin to happen if we want to make fully informed decisions on how to proceed in the field of robotics.

Similarly, the possible link between Sen's frameworks and human-robot interaction merits further exploration. Only one pair of researchers - Borenstein and Pearson - strongly associate roboethics with Sen's capability theory in their discussion of the socially assistive case of robotic caregivers [1]. Hansson also hinted at the connection in an overarching discussion of the ethics of enabling technology, but it appears more as a passing thought than a pillar of the exploration [3].

None of the researchers mentioned above were engineers or computer scientists. It seems like a worthwhile cause for computer scientists, engineers, expert ethicists, and others to work together to fully incorporate this theory into robotics progress.

Sen's ideas connect especially well to the development of ethical frameworks for human-robot interaction because of the autonomy yet also structure they may provide to robots in real-world environments. The idea of creating and distributing goods and services in a way that enhances the capability of each individual applies well to robotic technology and indicates clear pathways to guide the design of robot morphology and schemes for shared autonomy. Sen's theory bases its discussion of fairness on providing individuals with capabilities to enable them to function [7]. In this way, it is an especially relevant theory for the diverse populations of humans touched by social robots; it is as relevant for developing strategies for disabled individuals, the elderly, and those undergoing therapy, as well as the general population. The end argument of capability theory as it applies to robotics would be that technology areas like artificial intelligence, robotics, and cybernetic development should be supported to help enhance the capabilities of humans, in spite of some recent warnings from technology experts like Bill Gates, Stephen Hawking, and Elon Musk. A pursuit of robotics in this style may actually help resolve some of the issues raised in [6]. An endorser of capability theory would support the continued development of current robotic technologies toward the goal of enhancing people's abilities.

2.1 Morphology

Following ideas from Sen's approach, it stands to reason that as long as robots' morphological features in some way enhance the capabilities of humans who use them or surround them, they are ethical. Accordingly, the current cautious approach to such issues may inhibit the development of technologies that will enhance human capability in the long run. Sen would encourage us to explore daring and controversial robotics possibilities and interrogate only the results for the potential to augment the capabilities of individuals. A robot's *potential* to treat vulnerable populations, manipulate humans, or even swindle users should not be evaluated alone, but rather as part of a larger picture that includes the actual effect on capabilities of a human user or of the human designer. For example, it is acceptable if human users are distressed by the termination of a robotic therapy user study as long as they develop overall increased abilities by interacting with the robot. Similarly, in the case of robots working with vulnerable populations in private activities such as bathing or dressing, the prospective augmented freedom resulting from interaction with the robot probably outweighs the potential dangers of creating sensitive recordings during these activities. To be too cautious about intermediate steps and considerations may inhibit the achievement of a distribution result that is ultimately ethical.

2.2 Autonomy

The capability approach developed by Sen also provides insights on an appropriate approach to transitions between robot autonomy and temporary human control as robots are introduced in more and more human social environments. Sen would likely encourage robotics researches to boldly go forth and try to combine the robotic systems supported by capability theory with the software that can literally be created by building off of similar theory. To solidify a possible computational model of morality and fairness, deep learning could even come into play to fill in gaps in ethical reasoning that the human intellect is not sophisticated enough to solve, or to tell us if the task of quantifying morality is in fact impossible to solve. Provided there is some solution to this problem, the most ethical approach to handling autonomy of robots in human social environments (in Sen's view) might actually be to never allow a human to intervene and take on momentary control of the robot. In situations that formerly used Wizard of Oz experimental techniques, the passing of robot control to a human operator would be unnecessary and confusion about who or what is controlling the robot would be circumvented. This would require an exceptional amount of robot development and testing before a robot is released into an uncontrolled social environment, but would ensure the most ethical possible course of action based on the framework of morality discussed herein. Although it may not be ethical to transfer control to a human operator, it may be prudent to still include an emergency stop in modern robotic systems until we achieve a perfect computational model of morality (if one does indeed exist).

3. CONCLUSION

The overall result of this research could be unexpected insights in the field of philosophy using ultra-intelligent robots, as well as ethical robots that will not start a robot apocalypse, unless that turns out to actually be the ethically best choice. Influential morphological design, robots with significant potential for social sway, etc. are legitimate as long as they enhance the capabilities of the human. This philosophical framework for roboethics does leave some unanswered questions, especially in cases where robotic technology might temporarily enhance a person's capabilities and then be taken away, such as in the case of many socially assistive studies. Nevertheless, the framework laid out here does provide a strong argument for justifying many types of long-term robotic applications using ethical philosophy. It is possible that the design of robots to improve human capability may be the overall most fair approach to roboethics.

4. REFERENCES

[1] J. Borenstein and Y. Pearson. Robot caregivers: harbingers of expanded freedom for all? *Ethics and Information Technology*, 12(3):277–288, 2010.

[2] I. J. Good. Ethical machines. *Proc. Machine Intelligence Workshop*, 246, 1982.

[3] S. O. Hansson. The ethics of enabling technology. *Cambridge Quarterly of Healthcare Ethics*, 16(03):257–267, 2007.

[4] A. Laden. Games, fairness, and rawls's a theory of justice. *Philosophy & Public Affairs*, pages 189–222, 1991.

[5] J. Rawls. Justice as fairness: political not metaphysical. *Philosophy & Public Affairs*, pages 223–251, 1985.

[6] L. D. Riek and D. Howard. A code of ethics for the human-robot interaction profession. *Proc. We Robot*, 2014.

[7] A. Sen. Development as capability expansion. *Human development and the international development strategy for the 1990s*, 1, 1990.