MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Machine Ethics and Superintelligence

Carl Shulman, Henrik Jonsson, Nick Tarleton
MIRI Visiting Fellows

Abstract

The developing academic field of machine ethics seeks to make artificial agents safer as they become more pervasive throughout society. Motivated by planned next-generation robotic systems, machine ethics typically explores solutions for agents with autonomous capacities intermediate between those of current artificial agents and humans, with designs developed incrementally by and embedded in a society of human agents. These assumptions substantially simplify the problem of designing a desirable agent and reflect the near-term future well, but there are also cases in which they do not hold. In particular, they need not apply to artificial agents with human-level or greater capabilities. The potentially very large impacts of such agents suggest that advance analysis and research is valuable. We describe some of the additional challenges such scenarios pose for machine ethics.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009. "Machine Ethics and Superintelligence."
In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings*, edited by Carson Reynolds and Alvaro Cassinelli, 95–97.
AP-CAP 2009. `http://ia-cap.org/ap-cap09/proceedings.pdf`.

This version contains minor changes.

1. Introduction

The developing academic field of machine ethics seeks to make artificial agents safer as they become more pervasive throughout society. Motivated by planned next-generation robotic systems, machine ethics typically explores solutions for agents with autonomous capacities intermediate between those of current artificial agents and humans, with designs developed incrementally by and embedded in a society of human agents. These assumptions substantially simplify the problem of designing a desirable agent and reflect the near-term future well, but there are also cases in which they do not hold. In particular, they need not apply to artificial agents with human-level or greater capabilities. The potentially very large impacts of such agents suggest that advance analysis and research is valuable. We describe some of the additional challenges such scenarios pose for machine ethics.

2. Machine Ethics

The research area of machine ethics (also called roboethics) has recently emerged as a subfield of Artificial Intelligence focusing on the task of ensuring ethical behavior of artificial agents (commonly called AMAs, Artificial Moral Agents [Wallach, Allen, and Smit 2008]), drawing contributors from both computer science and philosophy. By focusing on the behavior of artificial agents the field is distinguished from earlier work in ethics as applied to technology, which concerned itself with the use of technology by humans and on rare occasions on the treatment of machines by humans (Anderson and Anderson 2007a).

Machine ethics researchers agree that any AMAs would be *implicit ethical agents*, capable of carrying out their intended purpose in a safe and responsible manner but not necessarily able to extend moral reasoning to novel situations (Weng, Chen, and Sun 2009). Opinions within the field part on the question whether it is desirable, or even possible, to construct AMAs that are *full ethical agents*, which like ethical human decision-makers would be capable of making explicit moral judgments and justifying them (Anderson and Anderson 2007a).

While Isaac Asimov's "Three Laws of Robotics" are widely recognized to be an insufficient basis for machine ethics (Anderson and Anderson 2007a; Weng, Chen, and Sun 2009), there is little agreement on what moral structure AMAs should possess instead. Suggestions range from applying evolutionary algorithms to populations of artificial agents to achieve the "survival of the most moral" (Wallach, Allen, and Smit 2008), neural network models of cognition (Guarini 2005) and various hybrid approaches (Anderson and Anderson 2007b) to value systems inspired by the Golden Rule (Wallach, Allen, and Smit 2008), virtue ethics (Wallach, Allen, and Smit 2008), Kant's Categorical Imperative (Wallach, Allen, and Smit 2008; Anderson and Anderson 2007a), utilitarianism (Anderson and Anderson 2007a), and many others.

3. Superintelligence

> A superintelligence is any intellect that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills.
>
> —Nick Bostrom (2003)

It has been proposed by various authors that the probability of a superintelligent software agent being created within the coming few decades is high enough to warrant consideration (Bostrom 2003; Moravec 1999; Hall 2007; Kurzweil 2005; Yudkowsky 2008). A superintelligent agent would have such enormous consequences that even if the probability of development is considered low, the expected impact justifies careful consideration well before the capabilities to produce such an agent are realized (Posner 2004; Rees 2003).

Because software programs can be freely copied so long as hardware is available to run them on, the creation of a single superintelligent machine could be shortly followed by the existence of billions or more (Hanson 1994), e.g. expanding the size of the artificial intelligence research community by orders of magnitude, with each advance improving the intelligence of the researchers in a positive feedback loop.

The potential rapid succession of increasingly intelligent machines designing their successors has been called an "intelligence explosion" (Good 1965), and could conceivably be initiated with a single machine. This suggests that safety concerns might suddenly become critical, so research would have to be done before it was clear when it would be required (Yudkowsky 2008).

4. Machine Ethics Challenges of Superintelligence

How would an artificial moral agent ideally express human values? Gary Drescher describes two classes of agents: *situation-action machines* with rules specifying actions to perform in response to particular stimuli, and *choice machines*, which possess utility functions over outcomes, and can select actions that maximize expected utility. Situation-action machines can produce sophisticated behavior, but because they only possess implicit goals they are rigid in behavior and cannot easily handle novel environments or situations. In contrast, a choice machine can easily select appropriate actions in unexpected circumstances based on explicit values and goals (Drescher 2006). Intermediate between these extremes, agents may have goals that are partially explicit and partially implicit.

Much discussion in machine ethics implicitly assumes that certain key features of the situations of artificial moral agents will be fixed and can be relied on to help constitute implicit goals for agents towards the situation-action machine end of the spectrum. However, these assumptions are much less likely to apply to superintelligent agents, and the removal of each causes new design challenges.

Assumption: AMAs are designed by humans and cannot alter their own architectures (Veruggio 2006).

Design challenge: AI goal systems must be reflectively consistent, i.e. the agent must not wish to rewrite its goal system to some other (Yudkowsky 2008; Omohundro 2007).

Assumption: Humans will be more powerful than AMAs, so it is possible to use game-theoretic considerations to induce cooperation.

Design challenge: AMAs must prefer the same outcomes as humanity for their own sake, as they plausibly could enforce their preferences coercively at an advanced stage.

Assumption: AMAs can be experimented with and tested in environments that closely resemble those in which they will be deployed (Wallach, Allen, and Smit 2008).

Design challenge: As artificial intelligences increase in sophistication, their knowledge, environment, capabilities, and incentives will change, so motivational systems that previously produced benign behavior may in later stages cause behavior that is catastrophic for human beings (Yudkowsky 2008).

Assumption: The design of AMAs will be an incremental and iterative process, with extensive opportunity for human supervision (Wallach, Allen, and Smit 2008).

Design challenge: The potential for an "intelligence explosion" could cause large apparent jumps in rates of increase in intelligence and leave the superintelligent agent free to enact whatever goal system is in place (Yudkowsky 2008).

We argue that to deal with these challenges, it would be necessary to create an architecture towards the choice machine end of the spectrum, with values that fully reflect those of humanity. This task is complicated further by our lack of introspective access to the causes of our moral intuitions (Greene 2002; Haidt 2001).

5. Conclusions

As the sophistication of artificial moral agents improves, it will become increasingly important to construct fully general decision procedures that do not rely on assumptions of special types of agents and situations to generate moral behavior. Since such development may require extensive research and it is not currently known when such procedures will be needed to guide the construction of very powerful agents, the field of machine ethics should begin to investigate the topic in greater depth.

References

Anderson, Michael, and Susan Leigh Anderson. 2007a. "The Status of Machine Ethics: A Report from the AAAI Symposium." *Minds and Machines* 17 (1): 1–10. doi:10.1007/s11023-007-9053-7.

Anderson, Susan Leigh, and Michael Anderson. 2007b. "The Consequences for Human Beings of Creating Ethical Robots." In *Human Implications of Human-Robot Interaction: Papers from the 2007 AAAI Workshop,* edited by Ted Metzler, 1–4. Technical Report, WS-07-07. AAAI Press, Menlo Park, CA. http://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf.

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Greene, Joshua D. 2002. "The Terrible, Horrible, No Good, Very Bad Truth about Morality and What to Do about It." PhD diss., Princeton University. http://scholar.harvard.edu/joshuagreene/files/dissertation_0.pdf.

Guarini, Marcello. 2005. "Particularism and Generalism: How AI Can Help Us to Better Understand Moral Cognition." In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium,* edited by Michael Anderson, Susan Leigh Anderson, and Chris Armen, 52–61. Technical Report, FS-05-06. AAAI Press, Menlo Park, CA. http://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-008.pdf.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834. doi:10.1037/0033-295X.108.4.814.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine.* Amherst, NY: Prometheus Books.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Moravec, Hans P. 1999. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8–9. http://intelligence.org/summit2007/overview/abstracts/#omohundro.

Posner, Richard A. 2004. *Catastrophe: Risk and Response.* New York: Oxford University Press.

Rees, Martin J. 2003. *Our Final Hour: A Scientist's Warning: How Terror, Error and Environmental Disaster Threaten Humankind's Future in this Century—on Earth and Beyond.* New York: Basic Books.

Veruggio, Gianmarco. 2006. "The EURON Roboethics Roadmap." In *2006 6th IEEE-RAS International Conference on Humanoid Robots,* 612–617. Piscataway, NJ: IEEE. doi:10.1109/ICHR.2006.321337.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties." In "Ethics and Artificial Agents." Special issue, *AI & Society* 22 (4): 565–582. doi:10.1007/s00146-007-0099-0.

Weng, Yueh-Hsuan, Chien-Hsun Chen, and Chuen-Tsai Sun. 2009. "Toward the Human–Robot Co-existence Society: On Safety Intelligence for Next Generation Robots." *International Journal of Social Robotics* 1 (4): 267–282. doi:10.1007/s12369-009-0019-1.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.