Automating Moral Reasoning (tutorial)

Marija Slavkovik[0000−0003−2548−8623]⋆

University of Bergen, Norway
marija.slavkovik@uib.no

Abstract. Machine ethics has, as its topic of research, the behaviour of machines towards humans and other machines. One aspect of that research problem is enabling machines to reason about right and wrong. The automation of moral reasoning is on one end the field of dreams and speculative fiction, but on the other it is a very real need to ensure that the artificial intelligence used to automate various tasks that require intelligence does not neglect the ethical and value impact this 'replacement' of man with machine has. This tutorial introduces the problem of making moral decisions and gives a general overview of how a computational agent can be constructed to make moral decisions.

Keywords: Machine ethics · artificial morality · computational agency

What is Machine Ethics? Artificial intelligence (AI) is concerned with the problem of using computation to automate tasks that require intelligence [3]. In a society, we all affect each other with our activities and decisions. Ethics (or moral philosophy) is concerned with understanding and recommending right and wrong behaviours and decisions [6]. The right decisions being characterised by taking into consideration not only ones own interest, but also the interest of others [7]. The more computationally automated tasks are used to complement or replace people's tasks, the more concerns we have to ensure that the resulting actions and choices are not only correct and rational, but also do not have a negative ethical impact on society.

One way to ensure that AI has a non-negative ethical impact on society is to consider that moral reasoning is itself a cognitive task that we can consider automating. Machine ethics, or artificial morality, is a sub-field in AI that is researching this approach. The problem of automating moral reasoning can be considered as a problem of moral philosophy, whereas one is interested in questions such as: should machines be enabled with ethical reasoning [5], which norms should machines follow [8], can machines ever be moral agents [4], etc. As a problem of computer science, machine ethics focuses on the question of how to automate moral reasoning [2,9].

Here we are concerned with the question of how to automate moral reasoning. Although this problem, and machine ethics in general, have been raised since

⋆ A longer version of this abstract can be found at https://drops.dagstuhl.de/opus/volltexte/2022/16004/

2006 [1], it is an extremely difficult problem that requires a lot of improvement in the state of the art in AI and moral philosophy. We discuss the basic approaches in machine ethics, the advantages and challenges of each. These lecture notes are structured as follows.

Tutorial Overview. In this tutorial, first we discuss what is decision making and how decision-making is distinguished from moral decision-making. Decisions are made by an agent. Next we discuss what computational agents are, what does it mean for a computational agent to be autonomous and what kind of moral agents can computational agents be. One way to automate moral reasoning is to follow a specific moral theory. We give a very quick overview of what is a moral theory and some of the more known moral theories from moral philosophy. Next, In we discuss two general approaches to building artificial moral agents, we discuss open research problems and challenges.

Tutorial Scope. In this tutorial we do discussed specific examples of artificial moral agents. This tutorial is not intended to be a systematic review of implemented machine ethics systems. A very practical reason for avoiding discussing implementations of artificial agents here is that these implementations vary vastly in the approaches they use and considerable background knowledge in various reasoning and learning methods would be necessary to understand the implementations. However, references are given to these specific systems and the interested reader can follow them and explore them for learning more.

References

1. Michael Anderson and Susan Leigh Anderson. The status of machine ethics: A report from the aaai symposium. *Minds Mach.*, 17(1):1–10, mar 2007. doi:10.1007/s11023-007-9053-7.
2. Michael Anderson and Susan Leigh Anderson, editors. *Machine Ethics*. Cambridge University Press, 2011.
3. Richard E. Bellman. *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company, 1978.
4. Bartosz Brożek and Bartosz Janik. Can artificial intelligences be moral agents? *New Ideas in Psychology*, 54:101–106, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0732118X17300739, doi:https://doi.org/10.1016/j.newideapsych.2018.12.002.
5. Amitai Etzioni and Oren Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21:403–418, 2017.
6. James Fieser. Ethics. In Michael Boylan, editor, *Internet Encyclopedia of Philosophy*. ISSN 2161-0002, 2021.
7. R. M. Hare. *Community and Communication*, pages 109–115. Macmillan Education UK, London, 1972. doi:10.1007/978-1-349-00955-8_9.
8. Bertram F. Malle, Paul Bello, and Matthias Scheutz. Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 21–27, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3306618.3314252.

9. Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, Inc., USA, 2008.