The difference between artificial intelligence and artificial morality*

Minao KUKITA†

ICAE 2014, Hokkaido University, Sapporo
Oct. 31, 2014

Abstract

Some researchers insist that we should develop artificial systems that can make moral decisions and/or moral actions on their own in order to avoid unpredictable disasters caused by the complex interaction between humans and artificial autonomous agents. Most of the researchers seem to believe that, for this purpose, it is sufficient to make an artificial agent only apparently moral, without regard to whether they are fully/truly moral in the same sense that we, human beings, are moral. Just as a weak AI may be sufficient for a certain practical purpose, limited "functional morality" will do as well for a certain practical purpose. In this article, we want to critically consider whether or how far this analogy between artificial intelligence and artificial morality holds, pointing out that one of the critical difference is that, while to deploy artificial intelligence raises no question whatever, to deploy artificial morality can be morally blameworthy.

Keywords: machine ethics · moral agency · social robotics

1 Introduction

Today, an increasing number and variety of autonomous robots and software applications are working in many areas of our daily life, including communication, finance, transportation, healthcare, housework, entertainment, and so on. The fear is that their complicated interaction with each other and with humans might lead to unpredictable disasters. Thus, some researchers insist that we should develop artificial systems able to autonomously make a "good"

*This study is supported by JSPS Grants-in-Aid for Scientific Research (KAKENHI) Grant Number 25370033.

†Graduate School of Information Science, Nagoya University, Japan. E-mail: minao.kukita@is.nagoya-u.ac.jp

decision, i.e., decision that respects our values. In other words, we need "moral" machines.

For this purpose, according to researchers, it is sufficient to make an artificial agent only apparently moral, without regard to whether they are fully moral in the same sense that we, human beings, are moral.[1] Just as a weak AI may be sufficient for a certain practical purpose, limited "functional morality" will do as well for a certain practical purpose. And works have been done in line with the mainstream artificial intelligence.

In this article, we want to critically consider whether or how far this analogy between artificial intelligence and artificial morality holds. We want to point out that one of the critical difference is that, while to deploy artificial intelligence for some practical purposes raises no question in general, to deploy artificial morality can be morally blameworthy in some situations.

2 Attempts to make artificial moral agents

Last May, it is reported that researchers of Tufts University, Brown University and Rensselaer Polytechnic Institute launched a project supported by the U. S. Navy to develop autonomous robots that can make a "moral" decision on their own (http://www.kurzweilai.net/can-robots-be-trusted-to-know-right-from-wrong). Mathias Scheutz, a professor of computer science at Tufts University, says that moral competence can be thought of as "the ability to learn, reason with, act upon, and talk about the laws and societal conventions on which humans tend to agree" and that the question is thus "whether machines [...] can emulate and exercise these abilities." Selmer Bringsjord, head of the Cognitive Science Department at Rensselaer Polytechnic Institute, proposes to use established logics for ethical reasoning such as deontic modal logics, and newly invented logics specific for certain tasks the system has to address. Both Scheutz and Bringjord are conducting their researches in line with the mainstream tradition of AI researchers, which Donald Gillies [6] calls the "Turing tradition." Gillies characterises the Turing tradition by two features: use of logic and close attention to practical problems. What is important is to identify the rules on which human judgements and actions are based, and implement the rules into the machines in order to make it avail for some practical problems.

This is also the case with other proponents of artificial morality. For example, Susan Leigh Anderson [4], one of the leading figure in the "machine ethics" project, says that they assume that ethics can be made computable, and that their job is to make a program that works out correct answers to actual ethical dilemmas. This suggests the striking similarity between AI and machine ethics. In fact, Anderson and others have developed an ethical AI called MedEthEx, an ethical advisor concerning how caretakers should behave in certain ethical dilemma situations. MedEthEx is a variation of the machine-learning systems that can learn from instances of human experts' judgements and infer some general rules, according to which MedEthEx can make its own judgement when

[1] For example, see Wallach and Allen [13], chapter 5.

faced with new cases (Cf. Anderson and Anderson [1, 2], or Anderson, Anderson and Armen [3]).

Wendel Wallach and Colin Allen, authors of *Moral Machines: Teaching Robots Right from Wrong*, are also notable proponents of artificial morality. They propose a hybrid approach in which a top-down (rule-based) mechanism and a bottom-up (learning- or evolution-based) mechanism are merged into one model. Specifically, they propose to employ Stan Franklin's learning intelligent distributed agent (LIDA) model, but furnish it with some moral-related information processing components. However, its overall structure is remains the same and not specific to moral-related activity. Wallach and Allen write:

> Within the LIDA model, moral decision making is a form of action selection similar to any other. From the perspective of action selection, a more human-like AMA does not need specially dedicated moral reasoning processes. Rather, the system needs only the normal set of deliberative mechanisms, applied to inputs having relevance to moral challenges. (Wallach and Allen [13], 173)

So their imagined artificial moral agents can be classified as a version of AI. Moreover, their emphasis on practical utility is also in accordance with the Turing tradition. Their aim, they say, is to build a system that produces judgements or actions that tolerably conform to our moral standards. So it does not matter whether the artificial moral agents are really moral or not, and we should let it suffice to have "functional morality" in the artificial agents.

What is, then, new about artificial morality? Is it just another branch of AI, or something essentially (or significantly) different from it? Does the difference lie only in the problems artificial morality try to solve? Does an artificial moral agent calculates, for example, whether it should shoot a particular person in the battle field, while AI system calculates whether it should capture a particular piece on the chessboard in the next move? Looking at what researchers have been saying and doing so far, they seem to think of artificial moral agents as just another kind of artificial intelligent systems that perform calculation, thus producing a desired solution to a practical problem. Moreover, as is evident in Anderson's opinion cited above, some seem even to identify morality with a certain ability of logical reasoning or calculating. They try to justify this reduction of morality by claiming that what they want to do is to make a practically useful machines, not fully moral artificial agents. Limited, apparent morality is enough for their purpose, they explain. However, I wish to claim that to deploy machines with such limited morality in actual moral dilemma situations may sometimes be morally blameworthy even if the behaviours of the machines are acceptable . And here lies the difference between artificial intelligence and artificial morality.

3 Differences

One thing concerning ethics that many people seem to agree on is that being ethical includes taking emotions of others into account — avoiding unnecessary damage to emotional well-beings of others. Wallach and Allen put much stress on emotional aspects of ethics. The ethical advisor of Anderson and others computes how much its judgement hurt the sense of autonomy of the patients it deals with. If emotions are essential to ethics, they adds to complexity and difficulty in designing artificial moral agents. This is partly because emotions are highly context-sensitive.The same response in the similar situations can result in different emotional effects. In fact, the sameness of the response itself can be the cause of bad effects. For example, imagine that you always repeat the same remark whenever your spouse prepares your favourite dish. The remark may please him or her once, but will eventually come to irritate him or her. Or consider politeness. Politeness is a good thing in general, but as people get closer to each other, the same polite manner can be a cause of frustration. Social relationship is dynamic. An action can change the relationship so that what is appropriate behaviour will also change thereafter. It will be extremely hard to calculate how this change take place.

Moreover, in our moral practice, in addition to what is done, it sometimes matters who does it. A particular action taken by an agent which are accepted could be rejected as unethical if it had been taken by another agent. Take for example punishment. Punishment is not something anyone can do. Even if a person is condemned to death, only executers are allowed to kill him or her.[2] The same holds true of education, preaching, etc. So letting a machine do some morally significant task can be morally suspect. If an artificial moral agent is to obtain general artificial morality, or to be a "strong" artificial moral agent, it will have to be able to calculate when and where it should assign itself some morally significant role. Otherwise — if it remains a "weak" artificial morality — we will have to decide when and where we should or may let an artificial moral agent in. In short, whether to use an artificial moral agent in a given situation is itself a moral decision making.

This point is connected with the Sherry Turkle's objection to social robotics ([11, 12]). She blames companion robots for impairing the authenticity of our social relationship by appearing to care for the users, and thereby eliciting their emotional responses. For Turkle, only authentic sentient entities can enter the caring relationship with humans.

Michio Okada, a roboticist, and Koutarou Matsumoto, a psychologist, also raise a question about social robots. They write in the beginning of their book titled *Sorrow of the Robot: Ecology of Human-Robot Communication* as follows:

> I was walking in the park. Then an elderly woman who stood alone caught my attention. Wondering if she was watching cherry blossoms, I got closer to her, when I found a tiny robot in her arms that resembled a stuffed toy. She was watching cherry blossoms with

[2] I have elsewhere argued against killer robots, including the execution robot [7].

> the robot in her arms. "Beautiful...," said she gently to the robot. "Look, beautiful, aren't they?"
>
> You will often see an elderly person walking in a park with a dog or a cat in her or his arms. In this case, the dog or the cat was replaced by a robot. So I could have passed her by, thinking that the times are changing. However, at that sight, I had a complicated feeling that I could not express easily.
>
> A vague question arose: "What? Isn't anything wrong?" In addition, I felt something painful, and uncomfortable at that sight. (Okada and Matsumoto [9], "Prologue," i-ii, my translation)

Starting with this episode, the contributors to the volume explore the sources of this painful and uncomfortable feeling. Their accounts, as well as Turkle's, may sound too naive or sentimental to the professional ethicists. However, their intuitions are, I think, important if we take seriously the coexistence of humans and machines.

What is missed in the discussion of artificial morality by ethicists and philosophers is the feelings of the third party of an action. A moral action is not relevant only to the agents or patients involved, but to the community or society around them as well. An action can be morally blameworthy if it goes against the sense of morality of many members of the community. Or, at least, an action is morally questionable if acceptance of that action demands the significant change of the community's moral common sense or/and moral practices.

For example, the acceptance of organ donation from brain-dead patients demanded the change of our definition of death.[3] Therefore, organ donation from brain-death patients is an action that calls for serious and careful discussion among the members of the society. Likewise, the acceptance of drone attacks will change our conception of what it is to fight a war or to be a soldier, and therefore morally significant, too.[4]

We claim that the same thing will hold true of the acceptance of certain artificial moral agents including care robots, companion robots, or autonomous lethal weapons which act according to the law of war (cf. Arkin [5]).

4 Conclusion

Artificial morality is a field of research of both practical urgency and academic attraction. It is technically challenging and philosophically inspiring. So far the investigations into artificial morality have been conducted in the similar fashions to those into artificial intelligence. Specifically, the proponents of artificial morality focus on morally acceptable judgements and behaviours that can be implemented in artificial agents and can be put to practical uses. The problem

[3] Morioka [8] describes how the definitions of "death" changed in two Presidential commission reports in the 1980s, and how "breath" regained the central role for life.

[4] Riza [10] points out that the drone technology imposes more risk on non-combatants than soldiers and thereby changes dramatically how wars are fought.

is that deploying such artificial systems in a given situation can be immoral, especially when it may lead to a significant shift of the moral common sense of the community. In such cases, we need serious discussion before we replace humans by machines.

References

[1] M. Anderson and S. L. Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazines*, 28(4):15–26, 2007.

[2] M. Anderson and S. L. Anderson. Robot be good. *Scientific American*, October:54–59, 2010.

[3] M. Anderson, S. L. Anderson, and C. Armen. Toward machine ethics. *American Association for Artificial Intelligence*, 2004. http://aaaipress.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf.

[4] S. L. Anderson. Machine metaethics. In M. Anderson and S. L. Anderson, editors, *Machine Ethics*, pages 21–27. Cambridge University Press, 2012.

[5] R. C. Arkin. *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC, Baca Raton, 2009.

[6] D. Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, New York, 1999.

[7] M. Kukita. Another case against killer robots. *Proceedings of Robo Philosophy: Sciable Robots and the Future of Social Relationship*, forthcoming.

[8] M. Morioka. An essay on the concept of brain death in the view of life philosophy: Presidential commission reports and revival of "breath". *Tetsugakuronsou*, 41:13–23, 2014. In Japanese (the original title is "生命の哲学から見た脳死概念の一考察 — 大統領レポートと「息」の復権 —", 『哲学論叢』).

[9] M. Okada and K. Matsumoto, editors. *Sorrow of the Robot — Ecology of Human-Robot Communication*. Shinyousha, 2014. In Japanese (the original title is "ロボットの悲しみ — コミュニケーションをめぐる人とロボットの生態学").

[10] M. S. Riza. *Killing without Heart*. Potomac Books, Washington D. C., 2013.

[11] S. Turkle. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, 2012.

[12] S. Turkle. Authenticity in the age of digital companitons. In M. Anderson and S. L. Anderson, editors, *Machine Ethics*, pages 62–76. Cambridge University Press, 2012.

[13] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York, 2009.

How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Is there a need for a new set of moral rules? What happens to human interaction when it is mediated by technology?Â Sections 3-7 consider how these core elements apply to artificial intelligence and robotics with discussion of fully autonomous and human-machine rule-generating approaches; types of moral reasoning; the difference between â€˜human willâ€™ and â€˜machine willâ€™; and respecting human dignity. 2. Core elements of Kantian ethics.