Can robots understand values?
Artificial morality and ethical symbol grounding

Minao, Kukita
PD, Kyoto University

Abstract

The main purpose of this paper is to reflect on the question of whether an artificial system can be a moral agent.

Recent decades have seen a great advance of autonomous machines or software programs in a variety of scenes in our society. This situation has led engineers and philosophers to set out to work on incorporating ethical codes into their robots (cf. Anderson and Anderson [1]). This area of research is called `machine ethics' or `artificial morality.'

While it is a great challenge of technological interest and of practical significance, artificial morality is likely to have a considerable impact on philosophy and ethics as well. For one thing, we have to prepare ourselves to answer questions of the extent to which we can permit robots to make moral decisions or to engage in activities that might have serious moral consequences. For sanother, artificial morality raise the question of whether we can create artificial agents which do not only simulate moral behaviors but are really moral. This question will in turn raise another difficult question of how we should treat these artificial moral agents.

Although these concerns (especially the latter) might sound too futuristic, Floridi and Sanders [3] argues that at some levels of `abstraction,' we can and should regard non-human beings --- including animals and robots --- as moral agents (or moral patients). Their doctrine is based on the argument similar to what supported the Turing test almost a century ago. Therefore, it is susceptible to the same criticisms that have been addressed to the Turing test. For example, the symbol grounding problem (SGP), which Harnad [4] raised, is also relevant in the context of artificial morality.

AI researchers have proposed various approaches to the SGP (cf. Taddeo and Floridi [6]), and have developed systems which realize increasingly human-like intelligences. It is true that such development blurs the boundary between human and machine intelligence, but it also indicates some limitation to the attempt of creating true artificial intelligence. This leads to the view that intelligence are only realized by a larger system comprising both human beings and machines as its component, bringing about a new conception of human intelligence (cf. Clark [2]).

We will draw some lessons here to apply to artificial morality.

We will argue that it is unlikely, if not impossible, that an artificial agent will be a real moral agent in itself. Just as artificial intelligent systems today are designed to enhance human intelligent activity and not to be intelligent in themselves, so will artificial moral systems of the future. However, we also argue that artificial morality will be an important issue in ethics --- in fact, even more so than artificial intelligence has long been in the philosophy of mind.

References

[1] M. Anderson and S. Anderson (eds.). *Machine Morality*. Cambridge University Press, New York, 2011.

[2] A. Clark. Reasons, robots and the extended mind. *Mind and Language*, 16(2):121–145, 2001.

[3] L. Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machine*, 14:349–379, 2004.

[4] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[5] J. R. Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 1:417–424, 1980.

[6] M. Taddeo and L. Floridi. Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence*, 17(4):419–445, 2005.