The Prospect of Moral Artificial Agents

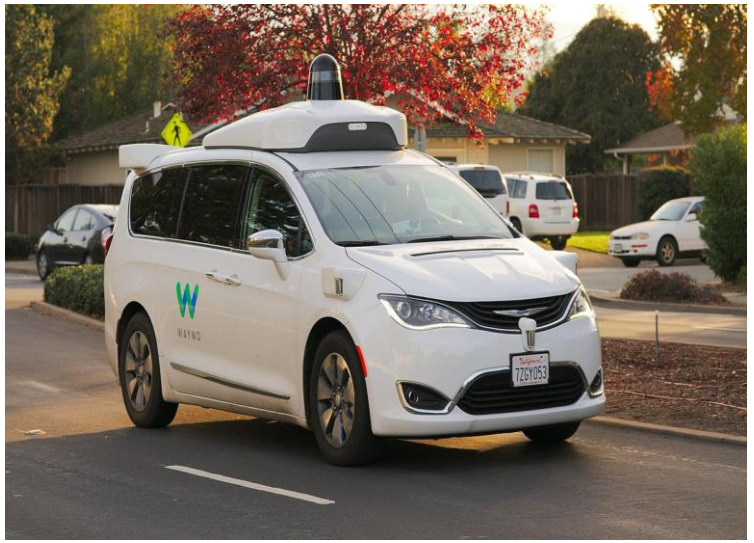
Jun Kyung You
Advisor: Prof. Axel Mueller

Foreword

Can we think of AI as Moral Agents?

Machines shouldn’t do Bad Things while Doing Our Job

“[In the case of a service robot in a home] Wouldn’t the robot need to discern whether an obstacle in its pathway is a child, a pet, or something like an empty paper bag?”
(Wallach and Colin, 2010, p.15)



“A goal of machine ethics is to produce an autonomous machine that will behave in ways that humans will find *generally acceptable from a moral point of view*” (Powers, 2014, p.4) (emphasis mine)

The Question that Follows:

Can we ever “trust” an independent moral decision of an artificial agent in controversial cases?

If not, moral decision-making is not replaceable

Machine that requires constant ethical oversight is not exactly “autonomous”

Replicating Moral Reasoning in Artificial Agents

What is, exactly, Moral Reasoning? Too hard and broad of a question!

But we do know that some “contextual consideration” is needed

Are there reasons to believe that these cannot be replicated?*

Moral Saints (via Susan Wolf)

"A person whose every action is morally good as possible … [a person] who is as morally worthy as they can be." ("Moral Saints," p.419)

Could this be problematic?

"if the moral saint is devoting all his time to feeding the hungry or healing the sick or raising money for [NGOs], then necessarily he is not reading Victorian novels, playing the oboe, or improving his backhand." ("Moral Saints," p. 421)

A Moral Saint's life is too barren

- Moral Saints have no self-interest, needs, passion, etc. "no identifiable self"
- Moral Life is the cooperation and negotiation between self-interested, independent individuals, which the Moral Saints are not
- Fundamentally different kind of agent = not trustworthy under controversies

… and Artificial Agents

Moral Artificial Agents are expected to be like Moral Saints

Humans try to be moral, and are free to try. Without also replicating this “freedom in trying” in an artificial agent, moral decision-making cannot be trustworthy.

But making artificial agents trustworthy through allowing the potential for them to produce bad consequences runs against the motivation of machine ethics.

Takeaway

The founding motivation of machine ethics is not compatible, in at least one way, with the conditions of trustworthiness of the moral reasoning of moral artificial agents

Visual Materials

https://cacm.acm.org/magazines/2017/5/216318-toward-a-ban-on-lethal-autonomous-weapons/abstract - Autonomous Weapons Photo
http://fortune.com/2016/03/24/chat-bot-racism/ - Chatbot
https://en.wikipedia.org/wiki/Self-driving_car - Autonomous Vehicle (Waymo Chrysler Pacifica Hybrid)

References

Floridi, L. The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, 2010.

Habermas, J. “On the Pragmatic, the Ethical, and the Moral Employments of Practical Reason.” Justification and Application. Cambridge, MA: MIT Press, 1993.

Powers, T. M. “Models for Machine Ethics.” Philosophy and Computers, Vol. 14 No.1.

Wallach, W, and Colin A. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, 2010.

Wolf, S. “Moral Saints,” The Journal of Philosophy, Vol. 79, No. 8, Aug. 1982.