Ethical Machines?

Ariela Tubert[*]

INTRODUCTION

In this Article, I will explore the possibility of having ethical artificial intelligence. As I will argue below, we face a dilemma in trying to develop artificial intelligence that is ethical: either we have to be able to codify ethics as a set of rules or we have to value a machine's ability to make ethical mistakes so that it can learn ethics like children do. Neither path seems very promising, though perhaps by thinking about the difficulties with each we may come to a better understanding of artificial intelligence and ourselves.

I. ADAPTABLE MACHINES AND UNETHICAL BEHAVIOR

Machines programmed with a set of fixed rules have a hard time adapting; they are too limited. One can make a machine that can answer a few questions, like a basic chatbot. But if the interlocutor asks things in a different way or asks a question that has not been programmed, then the chatbot is not able to respond properly. Similarly, early chess computers had moves and strategies programmed, but they were limited; they could not beat the best human chess players. Even Deep Blue, the IBM machine known for beating Kasparov in 1997, is said to use "brute force computational techniques" rather than having adaptable intelligence.[1]

One way to improve adaptability is to make learning machines, which can learn new things through repeated interactions. So, you can allow a chatbot, for example, to learn from its interactions with humans. The chatbot can ask questions and then use the answers provided by human responders to compose future responses to similar questions. As the size of the data set increases, the chatbot is able to provide more and better answers. Similarly, when it comes to playing chess, the most recent development is Alpha Zero, which recently learned to play chess in just a

* Department of Philosophy, University of Puget Sound.

1. Nathan Ensmenger, *Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm*, 42 SOC. STUD. SCI. 5, 7 (2011).

few hours from repeated simulated games.[2] Alpha Zero is thought to be a great development in machine chess, partially because, as Hassabis, one of its creators, stated, "[I]t doesn't play like a human, and it doesn't play like a program . . . it plays in a third, almost alien, way . . . it's like chess from another dimension."[3]

The problem is that learning machines can learn unethical behavior. For example, Microsoft's chatbot Tay started tweeting racist messages,[4] and Google Translate translated in a sexist manner.[5] More generally, there is reason to believe that machines who learn a language from humans will end up developing human biases unless specific steps are taken to prevent this from happening.[6]

Machines that surprise us with new ways of playing chess may be highly desirable, but machines that surprise us with unethical choices are not so. The fact that machines can learn unethical behavior is itself not surprising. After all, humans are not ethically perfect, and machines partially learn from us (this is what happened with the chatbot and with the translation program). But unethical behavior may also come about in other ways. Some of the advantages of the best chess playing programs are that they make moves that no human would consider; the programs do things that chess experts would not recommend, and the plays chosen are not intuitive. Relating this back to ethics, perhaps learning machines are both adaptable and able to do things in different ways—ways that humans would find unintuitive. This would allow for machines to perform various tasks much better than humans can, but it is also terrifying unless the machines have ethical knowledge.

II. ETHICALLY GOOD AND ADAPTABLE

What we want is a machine that can adapt—learn over time—and go beyond any original rules that we can program. But we also want a

2. *See* David Silver et al., Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (Dec. 5, 2017) (unpublished manuscript), https://arxiv.org/pdf/1712.01815.pdf [https://perma.cc/WWV8-YLLQ].

3. *See* Will Knight, *Alpha Zero's "Alien" Chess Shows the Power, and the Peculiarity, of AI*, MIT TECH. REV. (Dec. 8, 2017), https://www.technologyreview.com/s/609736/alpha-zeros-alien-chess-shows-the-power-and-the-peculiarity-of-ai/ [https://perma.cc/NQZ5-QMXA].

4. *See generally* Daniel Victor, *Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk.*, N.Y. TIMES (Mar. 24, 2016), https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

5. In translating from Turkish, Google Translate assumed that doctors were male and nurses were female even when the sentence in the original language was gender neutral. *See generally* Parmy Olson, *The Algorithm that Helped Google Translate Become Sexist*, FORBES (Feb. 15, 2018), https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/#4157f4dc7daa.

6. Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183 (2017).

machine that is able to recognize that certain patterns of behavior are unethical and therefore, should not be pursued. So, perhaps we can program some ethical rules that can serve as limits on learned behavior.

This would provide a solution, but the problem is that ethical rules are hard to program. Ethical rules need to be applied in the right context, and there are tradeoffs to be made in some circumstances. It can be difficult to recognize when a certain rule needs to be invoked and when it does not. In addition, there is disagreement as to when a certain rule should be applied and even as to which rules are correct.[7] Philosophers have been trying to establish a uniform set of ethical rules for a long time. The scientific revolution led philosophers like Thomas Hobbes in the seventeenth century to attempt to codify ethics as a set of principles, and there have been many attempts since then.[8] Perhaps there has been some progress, but we are not even close to having a list of rules that properly account for our ethical views.

Perhaps we should not be troubled by our inability to program a machine with a set of ethical rules. Humans learn ethical behavior over time and maybe so can machines. Perhaps all we need to do is allow robots to learn ethical behavior over time, like children do. We could equip a machine with enough of a moral program so as to set it on the right path and then allow for feedback and reinforcement mechanisms that lead it in the right direction. This parallels the way Alpha Zero learned to play chess, and arguably, this is how children learn ethical behavior.[9] Parents teach basic principles such as "do not lie," but as children grow up, they sometimes learn that lying may be the best option in some cases (for example, to save a life). Part of our ethical development as we mature is to understand ethical complexity and navigate choices when no good option seems available. Maybe that is how robots should develop ethical systems as well. Provide robots with the basics, let them learn, and they will eventually be able to act morally even while adapting and learning from an environment that may not be fully moral. Maybe they will be able to identify what is morally permissible and what is not, like a child that is raised properly would know not to adopt certain bad behaviors even when exposed to them.

III. HUMAN V. ROBOT FREEDOM

In humans, we value the capacity to make choices. Indeed, we value this capacity so significantly that we find abhorrent the idea of taking away

7. These and other difficulties are discussed in Russ Shafer-Landau, *Moral Rules*, 107 ETHICS 584 (1997).

8. *See generally* THOMAS HOBBES, LEVIATHAN (Broadview Press 2010) (1651).

9. *See* Silver et al., *supra* note 2.

the freedom of choice, even if it leads to better consequences. We do not often support strong conditioning that would leave people unable to make bad choices. For example, we find the type of behavior modification depicted in the movie *A Clockwork Orange* abhorrent.[10] Normally, we are not willing to use brain washing or some other means of controlling a person's mind so that they are good, even if we could prevent ethical wrongdoing that way. We value human freedom to choose so much that we are willing to allow much suffering and wrongdoing to maintain the freedom to choose. Instead of controlling the minds of those who break the law, we control their ability to move freely by sending them to jail but allow their minds to maintain the freedom to choose. We may consider nudging people in the right direction but not taking away the freedom to make choices, even if the world could arguably be, in many ways, better without the freedom to choose.[11]

The case with robots is different. Currently, we do not value the freedom to make choices in robots, possibly because they do not have it. It is perfectly fine to control a robot so that its behavior follows certain patterns and avoids bad decisions. In effect, that is what robots are, they are deterministic machines created partially because they are not able to choose otherwise and cannot refuse to perform a task they are supposed to. Robots are desirable, in part, because they do what they are programmed to do. And when it comes to ethics, predictable robots are great—we do not want too much of a surprise.

Learning machines, like all machines, are deterministic; that is, knowing the program together with the additional input would allow someone to know what the machine's next moves are. But learning machines are not necessarily predictable by humans—the amount of data and options are too large for us to process. This is desirable when it comes to machine chess: machines will surprise us with moves we had not imagined but are not desirable for machine ethics; we want to ensure choices that are recognizably ethical and avoid ethical surprises.

So, we want machines that adapt, learn, and surprise us but also behave ethically. However, we are not willing to tolerate much error in a machine. We want them to be infallible in a way that humans are not. Indeed, the kind of respect we have for human fallibility (we should be allowed to make mistakes) we do not have for robots' fallibility. To get a sense of the difference, consider that a racist chatbot is taken down while people posting racist comments on Facebook or Twitter are not removed

10. A CLOCKWORK ORANGE (Warner Bros. 1971).

11. *See generally* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2009) (explaining that "choice architecture" can nudge individuals toward the best decisions without restricting their freedom of choice).

from the websites. Or, when Uber self-driving cars were found to be crossing red lights, they were temporarily stopped from driving on the streets of San Francisco, and more recently, they were pulled off from several test cities after a crash in Tempe.[12] Would we allow even one death intentionally caused by a robot as part of an ethical mistake before taking it apart?

IV. ROBOTS AS CHILDREN?

The fact that we have very little tolerance for ethical mistakes in machines is relevant to the possibility of robots learning ethical behavior by mimicking the way children learn it. Could children develop an ethical conscience without the ability to make mistakes? Developing the capacity for ethical reasoning seems to require the ability to make choices, including the ability to make the wrong choices. If children are not allowed to make mistakes, then do they really develop the complex ethical conscience of an adult?

There is another difference that affects the analogy between robots and children. If a robot follows rules that it was programmed to follow, then the maker or programmer is responsible for its behavior. This is similar to the way it works with other machines; for example, the manufacturer is responsible for the malfunction of a vacuum cleaner, and part of the reason why the racist chatbot was taken down is that it would look bad on the company that made it if it continued to tweet racist things. Parents may be, in some sense, responsible for what their children do while they are children, but at some point they stop being responsible for their children's actions. Are companies responsible even when the problematic behavior was not predictable given the original program? When does the responsibility change from the manufacturer to the robot itself? At what point does the robot become responsible for its own behavior?

So, we now have two connected questions:

1. At what point does the value of freedom outweigh the value of good behavior when it comes to robots?
2. When do robots become responsible?

Perhaps unsurprisingly, I believe that these two questions are connected. The fact that we are held responsible for our actions—to the extent that we are—indicates that we value our freedom to choose more

12. Mike Isaac & Daisuke Wakabayashi, *A Lawsuit Against Uber Highlights the Rush to Conquer Driverless Cars*, N.Y. TIMES (Feb. 24, 2017), https://www.nytimes.com/2017/02/24/technology/anthony-levandowski-waymo-uber-google-lawsuit.html?mtrref=www.google.com; Daisuke Wakabayashi, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam*, N.Y. TIMES (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

than we value the ethical good we could bring about by severely restricting freedom of choice.

CONCLUSION

When it comes to ethical machines, we face a dilemma: we either allow robots to make ethical mistakes and choose wrongly or we deny robots the ability to make ethical mistakes and thereby limit them to whatever ethical rules we are able to program into them. Before concluding that we should allow robots to make ethical mistakes, we must remember that, with good reason, we have little tolerance for serious ethical mistakes by robots. In addition, we want significant decisions made by robots to be transparent and predictable. We may value unpredictability in a game of chess but less so if many human lives are at stake. So, maybe the best way forward is to try to find a way to codify ethics; however, ethical choice is complex, and we are not even close to being able to codify it into a set of rules that could be programmed.