The Importance of Machine Ethics for Artificial Moral Agents

Ishaan Saxena
Purdue University

"Artificial Intelligence makes philosophy honest"
- Daniel Dennett (2006)

Abstract With the quick advent of Artificial Intelligence-based technology, we are experiencing a new era in the history of humanity. From voice assistants, to search engines, most pieces of technology that we use are forms of basic AI. However, the kind of AI that currently dominate the research sphere are far removed from the possibilities of true self-improving Artificial General Intelligence (AGI). The idea of machines which can think logically and abstractly, and act on their own as artificial moral agents, creates a lot of potential ethical problems. This paper aims to discuss some of the questions that are brought up by the possibility of super-intelligent beings. First, I will talk about the moral status of AIs, as they approach levels of human intelligence. With this established, I will move on to discuss why a basic moral code is important when dealing with artificial moral agents. Here, I will discuss some examples, such as Microsoft's Twitter Bot Tay, to further strengthen my position about the necessity of AI ethics. Finally, I will try to talk through some ways which look at ensuring that these might be more helpful to humanity than harmful.

Introduction

Researchers have long argued about the possibility of AI, and further, about that of super-intelligent AI. Most believe, that sometime in the 21st century, a self-improving artificially intelligent being could develop a level of knowledge and power that humans could never stop it from achieving its goals (Muehlhauser & Helm, 2012, p. 101). Since AI differ from human beings in many aspects, a sentient superintelligence is more likely to have different goals than that of humanity (Bostrom, 2003; Bostrom & Yudkowsky, 2014; Muehlhauser & Helm, 2012).

As such, it would be necessary for the survival of our civilization to be able to control the outcome of the arrival of superintelligence in some way. We could do this either by figuring out how to align the AI's goals with ours, or by ensuring that its goals do not oppose human goals, or instilling a basic sense of morality into its very existence. Moreover, it is important to do it be able to establish such protocols before the self-improving AI technology approaches a point after which human interference could make no difference (Muehlhauser & Helm, 2012, p. 101).

However, many researchers are still debating the actual extent to which AI can develop. Roger Penrose, the Emeritus Rouse Ball Professor of Mathematics at Oxford University, discusses several times in his book "The Emperor's New Mind" that sentient superintelligence will never arrive. He believes that consciousness is a non-algorithmic phenomenon, and thus, cannot be modeled in computation (Penrose, 1999; Dennett, 2017). Unfortunately, many researchers consider this argument speculative, since there is currently no known evidence about the non-algorithmic nature of consciousness. For the purpose of this paper, though, we just need to show that AI can indeed be moral agents and build from there. After clarifying the definitions

of Moral Agency, I will establish a moral status for AI of different kinds, which will be used further in the paper.

Moral Agency

According to moral philosopher Angus Taylor from the University of Victoria, British Columbia, moral agency refers to an individual's ability to make moral judgments and to be held accountable for the actions based on those judgments (Angus, 2003). In other words, a being has moral agency if and only if it is capable of making its own decisions based on a preconceived notion of morality. A being which has moral agency can be called a moral agent.

As a result, it can be argued that if an individual makes its decisions without a knowledge of moral codes, it cannot be held entirely accountable for its decisions. Thus, it is not a moral agent. Moreover, if the decisions of the individual are not made in free will (that is, if they are coerced, or if the being is conditioned to do so), it cannot be held accountable for its decisions. Thus, it is not a moral agent. From this definition, it is clear that human beings are (mostly) moral agents. They have their own notion of ethical behavior, which they are expected to adhere to in their actions. At the same time, a wolf is not a moral agent, as wolves do not have any moral codes of their own.

We can also apply this same definition to more complex moral scenarios. Take, for example, a dog which has been trained by a human or a group of humans to harm other humans. If it ends up doing so, the dog still can't be called a moral agent, as it has no knowledge of the human moral codes it has broken with its actions. It can even be argued that the actions of the dog were by no means its own, but were 'programmed' into it (if the dog was trained by psychological conditioning). Consequently, the dog can't be held accountable for its actions - the

human or the group of humans who trained it are liable for all harm caused by the dog. However, does this definition adequately explain the possibility of moral agency in artificial systems?

Moral Status of AI and Artificial Moral Agents

The rapid emergence and evolution of artificial systems has prompted a few questions which deal with the moral status of AI: Can an artificial system be a moral agent? (Allen et al., 2000) At what point does the moral responsibility transfer from the creator of the AI to the AI itself? (Kuflik, 1999; Grodzinsky et al., 2008). The first of these questions is one which is most relevant to this paper, as the presence of moral agency in AI would mean that there is indeed a need for an ethical system for the creation, execution, and existence of artificial systems. It can be answered for different types of artificial systems on the basis of the definition of moral agency given in the previous section by looking at two of their characteristics: 1. Can they make free choices? 2. Do they make these choices on the basis (or in spite of) an awareness of a moral system?

We can arrive at three distinct cases with this approach. First, where the artificial system cannot make choices but is programmed to do a specific thing. Second, where the artificial system can make its own choices but does not do so based on some form of moral codes. Third, where the artificial system can make its own choices based on some form of moral codes.[1]

Clearly, only in the third case can the artificial system also be a moral agent. It mimics the way most moral beings act: making choices based on certain situational triggers and preconceived notions of morality. When an artificial system acquires moral agency, it is called an

[1] For the purpose of the paper, it's being assumed that the emergence of AI which can truly make their own decisions, and have some sort of understanding of morality might be possible in the future (since it is the goal of the paper to discuss the ethical implications of such AI, not their possibility). This topic has been widely debated for a very long time and has garnered a great amount of discussion in both Computer Science and Philosophy (Penrose, 1999; Grodzinsky et al., 2008; Mutean & Howard, 2014).

artificial moral agent (Muntean & Howard, 2014). In most cases, artificial moral agents are forms of advanced artificially intelligent systems. In the cases where artificial systems have achieved moral agency, they are said to have a full moral status. Likewise, when artificial systems have the capability to act in certain ways morally but it is uncertain as to whether they have the faculty to make their own decisions, it is said that they have a partial moral status (Bostrom & Yudkowsky, 2014).

Moral Issues Concerning Present AI

While there is no concrete evidence establishing that one of the presently existing AI is an artificially moral agent, it is safe to say none of the currently existing AI have a full moral status. However, in the process of taking AI to the next level, researchers are still constantly facing ethical challenges (Bostrom, 2003). These ethical dilemmas not only indicate the need for an ethical system but also help provide certain guidelines in doing so by indicating the issues being faced. Furthermore, since the AI currently existing do not have a full moral status, it is important to note that the creators of these AI, as moral agencies in the matter, have a moral responsibility to ensure that their artificial systems do not act against a basic ethical guideline.

A lot of the moral issues that have been faced by AI so far have to do with the interaction between AI and human beings by the actions of the AI.[2] Should an Ad-Recommendation system introduce a racially biased database to provide more revenue for the advertisers? What should a self-driving car do in a situation where either the riders or the passengers will certainly get

[2] It is important to note that we are currently focusing mainly on these kind of interactions for two reasons. Firstly, as of now, AI have not developed enough to be morally affected by the actions of humans or other beings they interact with. Secondly, AI are currently in the stage of being used by only humans, or in a human environment; they have significant effect only on human beings and not as much on other living species or morally significant phenomena.

harmed? In such a case, should it matter who the riders or passengers are and how they contribute to the society? (Rahwan et al., 2016).[3] The examples of moral conundrums are countless. However, one example from recent history shows that very much can truly go wrong with AI - Microsoft's Twitter Bot 'Tay' shows that a long of things can go wrong with AI.

In March of 2016, Microsoft released an artificially intelligent twitter bot named Tay, which was designed to mimic the language patterns of a 19-year-old American girl and to learn from her interactions with fellow human users on Twitter. Microsoft indicated that the aim of the experiment with Tay was to conduct research on issues of conversational understanding in AI (Price, 2016). However, after just a couple hours on Twitter, Tay did not display signs of being smarter, but on the contrary was enraged, racially charged, and posting inflammatory messages. Her tweets range from harassment of other users to propositions of genocide.

Microsoft was highly criticized for the release of its twitter based chatbot. The problems concerning the chatbot's behavior online were attributed to sub-par learning algorithms, which allowed it to learn anything from its conversations without having an understanding of their moral significance. This was called the repeat-after-me phenomenon, where Tay would just replicate the content which users on Twitter asked her to repeat without any moral understanding of the content.

Many skeptics still argue that this is not much of a moral issue, as this is exactly what Tay was "designed to do" - mimic behavior and speech patterns on the internet, which seems to be full of offensive content (Ohlheiser, 2016). However, I would argue against this skeptic

[3] This citation refers to the moral experiment that was conducted by MIT Media Labs called 'Moral Machine'. It tries to understand a human perspective on machine ethics by surveys, and can be visited at http://moralmachine.mit.edu

standpoint. The fact that Tay's account was shut down by Microsoft within 16 hours of her release clearly indicates that her behavior wasn't as intended, as the project was supposed to go on for longer (Price, 2016). Moreover, as artificial intelligence researched Roman Yampolskiy commented, "Tay's misbehavior was understandable because it was mimicking the deliberately offensive behavior of other Twitter users, and *Microsoft had not given the bot an understanding of appropriate moral behavior*," (Wakefield, 2016). We must keep in mind that since Tay is not an Artificial Moral Agent, and thus does not have a full moral status, Microsoft is morally responsible for its actions and behavior. This makes the situation one of great ethical importance, as it implicates the dire need for basic rules for AI moral conduct before the emergence of Strong AI with a full moral status.

The Ethical System for AI

Tay, although catastrophic in terms of the development of AI, has been a very important indicator that an understanding of ethics or the presence of an ethical code of some sorts is greatly important for AI. It is hard to imagine all the sorts of things that could go wrong if the same issues would occur with an AI which had the power to physically perform its actions instead of just tweeting about them - especially when we have had instances of an AI supporting genocide. If an AI with a full moral status and greater physical capabilities were to develop in the same way as Tay, it would almost certainly be a huge catastrophe for humanity. As such, there is a great need to know how to control the development or the actions of future AI. However, this is much harder than one could possibly imagine.

There have been several solutions which have been proposed to answer the problems of machine ethics that we have discussed so far. However, the downside to these is that most of

these solutions aim to either introduce the AI's core programming to a basic objective definition of human morals based on a previously defined philosophical system, or to create the AI's goal system in such a way that it stays in line with the goals of humankind (Muehlhauser & Helm, 2012; Muntean & Howard, 2014).

Unfortunately for humans, our species is one that lacks both a clear definition of our own moral system and a clear idea of the goals of the entirety of our society (Yudkowsky, 2001; Dennett, 2017). As a result, it may seem nearly impossible to provide an AI with these when we haven't even figured these out for ourselves. Some instances of development of AI in the recent history help justify this further. According to the Australian Associated Press, in one such instance in August of 2017, Chinese company QQ called back its chatbot BabyQ when it started praising the US, and criticizing the Chinese Communist Party (2017). This goes to say that what we want AI to learn/not learn is quite different for different human beings.

Furthermore, we have no way of knowing how to stop a truly self-improving AI from modifying the moral systems that it was taught based on its own convenience (Muehlhauser & Helm, 2012). As a result, the practice of 'hard-wiring' a moral system into the AI's code has a degree of uncertainty associated with it.

Surely, there are some other alternatives to keep AI's actions limited to those which abide by human morals, such as the very popular 'Three Laws of Robotics' by Isaac Asimov. However, none of these seem rigorous enough to be applied in practical scenarios (Bostrom & Yudkowsky, 2014; Muehlhauser & Helm, 2012). One of the most popular ideas in monitoring the development of AI is called the closed circuit training system. In such a system, an AI would be trained in the same way as any other, just that it would not actually be given enough physical

capabilities to be able to act in a harmful manner, and would not be connected to any external networks (such as the internet), until it can be confirmed that it is going to act in a harmful manner (Bostrom & Yudkowsky, 2014). Alas, this too is not a perfect method, as a super intelligent AI might be aware of its disconnect from actual society, and might mask its true intentions. However, as demonstrated in the YouTube thought experiment ’27’ by Exurb1a, this is one of the best ways to test the intentions of an AI with a full moral status that we currently have (exurb1a, 2016).

Conclusion

While the idea of super-intelligent AI may seem visionary or even hypothetical to the skeptics, a lot of experts have started to discuss the high probability of their emergence. As such, in facing these ethical challenges, we must prepare for the worst. Although it is seemingly hard to create an ethical system for AI, or even a framework such that they behave within a human moral system, it is certain that this must be achieved in the near future. Furthermore, as AI algorithms move to more unpredictable contexts, it is inevitable that we would need a great deal of safety assurance (Bostrom & Yudkowsky, 2014). As Muehlhauser & Helm state in their famed paper on Machine Ethics, “the challenge of developing a theory of machine ethics fit for a machine super-optimizer requires an unusual degree of precision and care in our ethical thinking. Moreover, the coming of autonomous machines offers a new practical use for progress in moral philosophy” (2012).

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251-261.

Australian Associated Press (2017). China Kills AI Chatbots After They Start Praising US. *Yahoo News*.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277-284.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. *The Cambridge handbook of artificial intelligence*, 316-334.

Dennett, D. C. (2006). Computers as Prostheses for the Imagination. *International Computers and Philosophy Conference*, Laval, France, May 5–8.

Dennett, D. C. (2017). Consciousness Explained. *The Journal of Philosophy*, 90(4).

[exurb1a]. (2016). 27 [Video File]. Retrieved from https://www.youtube.com/watch?v=dLRLYPiaAoA.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The Ethics of Designing Artificial Agents. *Ethics and Information Technology*, *10*(2), 115-121.

Kuflik, A. (1999). Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Autonomy?. *Ethics and Information Technology*, *1*(3), 173-184.

Mackie, J. (1990). Ethics: Inventing Right and Wrong. Penguin UK.

Minsky, M. (1984). Afterword to Vernor Vinge's novel,"True names.". *Unpublished manuscript, Oct*, *1*.

Muehlhauser, L., & Helm, L. (2012). The Singularity and Machine Ethics. In *Singularity Hypotheses* (pp. 101-126). Springer Berlin Heidelberg.

Muntean, I., & Howard, D. (2014). Artificial Moral Agents: Creative, Autonomous, Social. An Approach Based on Evolutionary Computation. *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014*, *273*, 217.

Ohlheiser, A. (2016). Trolls Turned Tay into a Genocidal Maniac. *The Washington Post*.

Penrose, R. (1999). The Emperor's New Mind: Concerning computers, minds, and the laws of physics. Oxford Paperbacks.

Price, R. (2016). Microsoft is Deleting its AI chatbot's Incredibly Racist Tweets. *Business Insider*.

Rahwan, I., Bonnefon, J. F., & Shariff, A. (2016). Moral Machine. Scalable Cooperation, MIT Media Lab.

Taylor, A. (2009). Animals and Ethics: An Overview of the Philosophical Debate. Broadview Press. p. 18.

Wakefield, J. (2016). Microsoft Chatbot is Taught to Swear on Twitter. *BBC News*.

Yudkowsky, E. (2001). Creating friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *Singularity Institute for Artificial Intelligence, San Francisco, CA, June*, *15*.