HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 2 (2022)

What are the Ethical Considerations Involved in the Creation of a Superintelligent AI?

Justin Bonneau-Diesce[1] and Alex Chan[#]

[!]Eastbourne College, United Kingdom
[#]Advisor

ABSTRACT

IBM Cloud Education (2021) has defined artificial intelligence (AI) as a system that "leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind". Therefore, we could assume that similarly to humans, AI comes with advantages as well as disadvantages. However, if we consider that one of the most common daily uses of AI is in translating a text from one language to another (European Parliament, 2021), it may seem that there is little risk involved with AI. In reality, though, according to Harris (2016), humanity is standing before two doors. To go through one door would mean that we would stop the production of intelligent machines and our technology would cease to progress – marking the end of the technological era and potentially seeing one of the greatest failures in human history. Yet, from the evolution of humanity, we might conclude that it is not in our nature to cease developing meaning that the only remaining option is behind the second door. Following the second door, we continue to improve our technology until eventually, it starts improving itself, leading to an 'intelligence explosion', according to Harris[1] (2016), a scholar famous for his controversial views regarding the future of AI. Consequently, it is only a matter of time before we reach a new era – one where humanity could lose its position as the apex predator to a machine. In this paper, I will be assessing some of the potential risks and ethical considerations involved in the development of artificial intelligence and proposing ways in which we can avoid them.

Introduction

In terms of methodology, this paper is of the review type on a topic of interest. The data was collected through reading over tens of research and debate papers on the relationship between the fields of artificial intelligence and ethics. Published in 2016 in the journal Nature, the paper called 'Program good ethics into artificial intelligence' by Jim Davies is what inspired the creation of this review due to the novel reflections that it discussed. Furthermore, his position as a professor of Software Engineering at Carleton University's Department of Cognitive Science makes him a knowledgeable source on the subject. Following that article, I filtered my research to keep only relevant papers from reliable sources or authors associated with prestigious research institutions. For example, Nick Bostrom, who is a Professor at the University of Oxford and Director of the Future of Humanity Institute, is also renowned for his numerous award-winning papers and books on existential risk and superintelligence risks – asserting his opinion as very informed, making him a valuable source. To categorise my research, I structured this review into an abstract, an introduction, a refute, detailed paragraphs on potentially dangerous scenarios, paragraphs on the ways to mitigate the risks of those scenarios from happening, and a conclusion, which enabled me to assign each article to one or more parts of this essay before writing. Considering this is a review paper, no software programs or statistical formulas were

[1] Despite being famous for his controversial view concerning the future of artificial intelligence, Sam Harris' arguments were used in this paper to demonstrate extreme examples and prove points. Furthermore, he is a renown TED-Summit speaker and philosopher with a B.A. from Stanford University and a PhD in Neuroscience from the University of California.

used to help in the research. Instead, the varied opinions of many, diverse researchers have allowed me to situate myself in the debate and make my own conclusions. The rationale behind selecting this topic of research is due to its current and relatable issues to many people as we live in a very connected world in which autonomous machines are more present than we think. Following the divided public opinion on the matter of artificial intelligence, some governments and media have tried to appease fears by telling the people of the greater good that AI could bring – but they have left out half of the story. This paper has for purpose to inform of the worst-case potential risks that we may encounter if some conditions are not met. The sources used in this essay include authors such as Paul Lukowicz, a Professor and a Scientific Director of the Embedded Intelligence Research Group, German Research Center for Artificial Intelligence (DFKI); Karim Jebari, a researcher at the Institute for Futures Studies who presented and argued his doctoral thesis on applied ethics and how it relates to the risks and opportunities of technological innovation at the Royal Institute of Technology; or even Professor Csaba Szepesvári who is affiliated with the University of Alberta, and mainly focuses his research on the fields of Learning Theory, and Reinforcement Learning. However, this also includes philosophers like Sam Harris, a renowned TEDSummit speaker and philosopher with a B.A. from Stanford University and a PhD in Neuroscience. Despite being famous for his controversial view concerning the future of artificial intelligence, his arguments were used in this paper to demonstrate extreme examples and prove points. Independently of their opinions, each author selected in this article was done depending on their knowledge, fields of study, and affiliations with institutions.

Since the Industrial Revolution, humanity has constantly strived for technological advancement, leading to "the most important dimension of humanity's progress to date" (Nowak et al., 2018)[2]: artificial intelligence. However, despite the benefits that AI can bring, it is important to be vigilant of its potential dangers if precautions are not followed.

Over recent years – with the help of the media – a growing fear that artificial intelligence might develop consciousness has emerged and raised public attention. However, this fear is misplaced, according to Professor Jim Davies from the Institute of Cognitive Science at Carleton University. He argues that consciousness is not the characteristic in AI that is likely to lead us to an existential threat, continuing to say that "in humans, one process that puts the brakes on immoral behaviour is 'affective empathy': the emotional contagion that makes a person feel what they perceive another to be feeling. Maybe conscious AIs would care about us more than unconscious ones would." (Davies, 2016) Furthermore, if a superintelligent AI is made and given a predetermined goal, as long as it considers humans to be detrimental to its task, it could still initiate major problems for us – conscious or not.

Thus, could AI ever become a threat to humans? The answer is simple: it depends. In a scenario where an artificial intelligence had the goal of decreasing suffering, it could try to eliminate humanity for the good of the rest of life on the planet. (Bostrom, 2016) In this scenario, the problem is not consciousness – or perhaps its lack of it – it is rather a cruel absence of good ethic within the machine. (Davies, 2016) In Professor Davies' views, the most crucial point that artificial intelligence would need to acknowledge would be that we, humans, exist and that it was created to benefit us, and not the reverse. In the future, he argues, when an AI reaches a point where it has surpassed us intellectually, only good ethics could stop a post-apocalyptic Hollywood-esque scenario. Additionally, we can safely assume that if such a machine does come to exist, it will be used in a broad range of circumstances – making it a "general agent (i.e., an agent capable of action in many different contexts)." (Jebari & Lundborg, 2020) Though, by doing that, the risks only become greater since the chances for the machine to pose a serious threat grow exponentially.

Alternatively, many other scenarios could lead to the creation of a hazardous AI. For example, if researchers attempted to engineer a new type of intelligent machine and resulted in designing an artificial intelligence independent of human cognition, this could instead act as a method of replacement for it. The machine would start by developing its own highly efficient systems of knowledge and reasoning which would be incompatible with those of humans. As it would be strictly impossible for the AI to support and cooperate with us, we would gradually be replaced by the

[2] Paul Lukowicz, one of the three main authors, focuses his research on cyber-physical systems which further helps to assert the reliability of their paper.

machine in more and more tasks – initiating an increased loss of control from humans. This is where implementing a considerable amount of morals within the AI becomes essential to make sure that it knows its role and continues to work in our interests despite humans not fully comprehending what the machine is trying to achieve or having taken monopoly of some areas. If not, eventually, the AI could start seeing us as competition in the fields that it does not control yet and upgrade itself accordingly; "the principle of self-preservation would lead to the development of defensive strategies on the part of the AI". Those strategies, such as "hiding itself, replication, and resource maximization" (Nowak et al., 2018), would become too advanced for us to anticipate and follow, causing our overall loss of control of the world, of our freedom and, in an extreme case, lead us to an existential threat. Some people may argue that the first stage has already been reached as research from MIT Lincoln Laboratory (2021)[3] demonstrated that when a variety of people played two games with an advanced AI model, one with a general AI and the other with a manually programmed one specifically for this game, all reported that both machines were "unpredictable, unreliable, and untrustworthy, and felt negative even when the team scored well", making cooperation near impossible.

However, some factors could mitigate the likeliness of those risks from occurring. Firstly, nearly all AIs are designed with the objective of, within given parameters, yielding the highest scores, and are benchmarked by their objective performances. This is an enormous problem because it makes the machines unable to operate with humans on a task since they were not intended to meet human preferences but instead output high results, alone. Associated with reinforcement learning, which is the "learning paradigm concerned with learning to control a system so as to maximize a numerical performance measure that expresses a long-term objective" (Szepesvári, 2010), it gives the impression to humans, that tend to think about the short-term consequences, that the AI's decisions are not only unpredictable but also random. The main problem to address, however, is the current lack of morals within those machines. When a superintelligent AI finally appears, if not already programmed with our own codes and laws but provided with sufficient abilities, it will redefine its environment, in some cases that will be the world, with its own ones to set up the optimum conditions for the success of its task regardless of the safety or wellbeing of humanity. One safety precaution would be to fund a project to make sure that the first superintelligent AI is a friendly one with which we can cooperate, and this will only be possible with a "well-funded team of ethics-minded" (Davies, 2016) programmers and researchers that can prioritise the AI's integrity and morals over the profits that it will produce.

Alternatively, there are also more radical ways to regulate AI agents. On one end of the spectrum, we could completely prohibit autonomous machines from operating and being developed, due to the risks and uncertainties that they pose. Nevertheless, if this method proves to be successful, it would impede the development of many beneficial AIs that could drastically impact our lives in positive ways. On the contrary, we could permit the development and deployment of autonomous machines and accept the risks and costs at a social level, without constructing a more effective framework for the regulation of these agents. This is the permissive approach. (Chopra & White, 2011)[4] It would allow the operation of many beneficial AIs but also many dangerous ones. Ipso facto, this would eventually lead to a backlash and a pessimistic public opinion concerning the fate of artificial intelligence, and technologies in general, because of the harmful nature of some of the machines that were permitted to exist. Each of those two opposing methods is different but deals with the same problem: how to contain an artificial intelligence that has been coded with a lack of morals and, therefore, has become a nuance and a danger to society.

Furthermore, a government also has the necessary jurisdiction to devise and implant measures to prevent the development of a potentially dangerous artificial intelligence in its country. For example, a stricter liability policy for companies or even governmental organisations that work to develop a superintelligent AI. There are "two liability

[3] This study was performed by one of the leading institutions in the world on the field of Technology: MIT – making this source unquestionably reliable.

[4] Samir Chopra is a Professor of Philosophy at Brooklyn College, CUNY, with an expertise in Foundations of Artificial Intelligence. Laurence F. White is a lawyer and policymaker specialised in law, technology, and financial markets regulation. Together, they have an in depth understanding of what these different approaches would lead to from a social and economical point of view, as well as from a Computer Scientist point of view.

frameworks in the law, civil and criminal" (Asaro, 2016)[5]; however, artificial intelligence is hard to legally hold liable for its actions considering that it is not a real person. Currently, even the developers of the machine can only be held liable if "provided with evidence of a foreseeable risk of harm rising to the level of criminal negligence." (Asaro, 2016) In the context of an autonomous AI that can learn and adapt, it is also very difficult to prove that someone knew the intent or foresight of the action that the AI would take. The addition of regulations is good, to a certain extent, as it would discourage large capital companies from entering the market – motivated by the high profits – and accepting the risks. As a result of less competition, companies can concentrate fully on the careful development of AI and the implementation of ethics and moral values, rather than on the intelligence race.

Finally, even with those safety measures followed, many scientists agree on the idea that we are nowhere near the summit of possible artificial intelligence. According to Sam Harris (2016), a neuroscientist and philosopher, "worrying about AI safety is like worrying about overpopulation on Mars". Factually, we are aware that our current technology is nowhere near advanced, and powerful enough to develop a superintelligent AI and that it will take decades at least for it to happen. Nevertheless, this is exactly what makes the problem so precarious, it is due to all the uncertainty and the unknowns that we will eventually have to face, ready or not.

Conclusion

In conclusion, there are several ethical aspects to consider to make autonomous systems safer. Artificial intelligence will always pose a threat to humanity if not built carefully with "subtle and complex ethics." Therefore, we must put substantial efforts into "programming goals, values and ethical codes" (Davies, 2016) in AI which will benefit us in the long term and avoid creating threatening machines that could harm us. Furthermore, creating advanced AI systems should only be attempted by accomplished programmers who understand that they "are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications" (Future of Life Institute, 2017)[6], to maintain it as impartial as possible. I believe that with a supervised legal framework and proper ethics, artificial intelligence could become a very powerful ally to us and diminish the number of hypothetical scenarios that we fear. The systems should also "be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity." (Future of Life Institute, 2017) Lastly, we need to realise that we are currently in the process of developing some sort of god, so "now would be a good time to make sure it's a god we can live with." (Harris, 2016)

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic

References

Asaro, P. (2016, March). *The Liability Problem for Autonomous Artificial Agents*. AAAI Spring Symposia.

[5] Professor Peter Asaro, the author, has held research positions at the Center for Information Technology Policy at Princeton University, Center for Cultural Analysis at Rutgers University, the HUMlab of Umeå University in Sweden, and the Austrian Academy of Sciences in Vienna. I have chosen this paper partly due to his prestigious associations which highlights the extent of his knowledge in this field of study.

[6] This report is from the Future of Life Institute which focuses on reducing near and long-term threats from artificial intelligence. It was initially found on a report by the European Parliamentary Research Service for a study on the ethics of artificial intelligence. Furthermore, the ASILOMAR AI PRINCIPLES has been signed by 1797 AI/Robotics researchers and 3923 others, confirming its credibility.

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies* (Reprint ed.). Oxford University Press.

Chopra, S., & White, L. F. (2011). *A Legal Theory for Autonomous Artificial Agents* [E-book]. University of Michigan Press. Retrieved 21 May 2022, from https://doi.org/10.3998/mpub.356801

Davies, J. (2016). Program good ethics into artificial intelligence. *Nature*. https://doi.org/10.1038/538291a

European Parliament. (2021, March 29). *What is artificial intelligence and how is it used?* [Press release]. https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

Future of Life Institute. (2017). *ASILOMAR AI PRINCIPLES*. https://futureoflife.org/2017/08/11/ai-principles/

Harris, S. (2016, September 29). *Can we build AI without losing control over it?* [Video]. TED Talks. https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it?referrer=playlist-talks_on_artificial_intelligen

IBM Cloud Education. (2021, June 30). *Artificial Intelligence (AI)*. IBM. Retrieved 12 December 2021, from https://www.ibm.com/uk-en/cloud/learn/what-is-artificial-intelligence

Jebari, K., & Lundborg, J. (2020). Artificial superintelligence and its limits: why AlphaZero cannot become a general agent. *AI & SOCIETY*, *36*(3), 807–815. https://doi.org/10.1007/s00146-020-01070-3

MIT Lincoln Laboratory. (2021, October 4). *Artificial intelligence is smart, but does it play well with others?* MIT News | Massachusetts Institute of Technology. Retrieved 11 December 2021, from https://news.mit.edu/2021/does-artificial-intelligence-play-well-others-1004

Nowak, A., Lukowicz, P., & Horodecki, P. (2018). Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance ? or the Biggest Threat [Opinion]. *IEEE Technology and Society Magazine*, *37*(4), 26–34. https://doi.org/10.1109/mts.2018.2876105

Szepesvári, C. (2010). Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *4*(1), 1–103. https://doi.org/10.2200/s00268ed1v01y201005aim009