INTI JOURNAL | eISSN:2600-7320
Vol.2021:005

An Overview on Existential Threats Posed-by Human-like Super-intelligent Machines to Humanity

Jie Xi[1]

[1] Peking University, 5 Yiheyuan Rd, Haidian District, 100871 Beijing, China.

Email: jiexi@pku.edu.cn

Abstract

This review paper aims to discuss the possibilities a future with human-like super-intelligent machines holds. This paper is done based on a study of secondary data from research papers in the field of history and philosophy. This paper projects that Artificial Intelligence (AI) may hold quality superintelligence, speed superintelligence, and collective superintelligence over us. A forecast based on human history gave us a bleak outlook into the future in which humanity is replaced and wiped out by AIs. Towards the end of the paper, cyborgification is briefly considered as a solution.

Keywords

Artificial intelligence, Humanity, Machine

Introduction

The ever-increasing pace of development of Artificial Intelligence (AI) has prompted many to consider the possibilities of superintelligence, and its consequences. There are fears that super-intelligent AIs may pose an existential threat to humanity. Bostrom (2014) described various ways a super-intelligent machine tasked with making paperclips could end up annihilating humans while doing its job. This essay agrees with Bostrom (2014) that non-human-like super-intelligent AIs are an existential threat to humanity, and will instead consider the possibility of human-like super-intelligent AIs, and the consequences that may occur.

Methodology

This review paper is done based on a study of secondary data, mainly from philosophical or historical papers. The three main types of superintelligence proposed in Bostrom (2014), a philosophical study into the ethics of Artificial Intelligence, are quality superintelligence, speed superintelligence and collective superintelligence were adopted by this paper. Speed superintelligence is “a system that can do all that a human intellect can do, but much faster”

Submission: 14 September 2021
Acceptance: 17 September 2021

(Bostrom, 2014, p. 53). Quality superintelligence is "a system that is at least as fast as a human mind and vastly qualitatively smarter" (Bostrom, 2014, p. 56). Collective superintelligence is "a system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system" (Bostrom, 2014, p. 54).

Then, research papers in the field of history in both ancient and contemporary were studied. The aim of looking into history is to find a regular pattern for the development of humanity after the emergence of superintelligence. News and research papers on modern technological developments were also taken into account to project possible developments in the future, based on historical patterns.

Definitions

Before we begin, a question we must ask is this: what is humanity? Is it the human DNA? Then humanity is preserved as long as there is at least a couple of opposite sexes capable of reproducing left on the world. In this case, AIs may dominate the world and make humans their slave, but humanity shall survive. However, that does not sound like a good way of life for humans. Then perhaps humanity is the preservation of the human way of life. The human way of life, for the sake of simplicity, is defined as thus: as the dominating species of planet Earth, we form our own societies and economies that are controlled by our own species. If the super-intelligent AIs chose to adopt the human way of life, then humanity will be preserved even if not a single biological human survived. That does not sound good either. Maybe the essence of humanity lies in human emotions, as some people believe that machines and animals do not have emotions. In this paper, it is assumed that a human-like super-intelligent AI will have emotions too. These human-like super-intelligent AIs will empathise with and help each other just like humans do. Then, if these AIs wiped out humans, humanity will still be preserved. That does not sound like a bright future for humans. This essay shall interpret humanity as a combination of the human DNA, human emotions, and the human way of life.

In this paper, a human-like AI is an AI with an internal system that perfectly replicates human's cognitive abilities. The human-like AI will have emotions. Certainly, the human-like AI cannot enjoy chocolate ice cream as we do, as it does not require food intake to survive. However, it will find joy in the beauty of flowers and the song of birds as we do. This human-like AI would even feel sad when presented with a good performance of Shakespeare's Romeo and Juliet. The human-like AI will also have self-awareness. It can identify itself in the mirror, and may even be vain about its looks. It may refrain from doing morally reprehensible things in public to maintain its reputation. The human-like AI also has intentionality. It may hold beliefs such as "a dog is an AI's best friend". It may have desires of having an AI companion, owning a dog and et cetera. In short, the human-like AI will have emotions, intentionality, self-awareness, and other human mental capabilities.

Submission: 14 September 2021
Acceptance: 17 September 2021

Threats

As human-like super-intelligent AIs are highly similar to humans, perhaps human history can provide us with a hint about the future. 50,000 years ago, during the Great Leap Forward, the Cro-Magnon evolved, and proceeded to replace other types of humans as they rapidly expanded their geographic range (Diamond, 2017). Archaeological evidence showed that Cro-Magnons have a superior advantage over other humans because they developed better technology and tools (Diamond, 2017). The vast advancement in Cro-Magnons' capabilities compared to other types of humans begs for an explanation. Diamond, 2017 believes that the advancement was due to Cro-Magnons' anatomically advance voice-box, which allowed them to develop modern language. Modern language is an efficient communication tool that allows us to cooperate efficiently, and pass on knowledge. If Diamond, 2017's explanation is correct, then machines would only need to achieve speed superintelligence to become a threat to humanity. If machines have the same level of intelligence as humans but can process information and communicate amongst themselves more efficiently, they would gain a superior advantage over humans. The day that machines would gain speed superintelligence over us does not seem very far away, as supercomputers are already processing information far more efficiently than humans. On the other hand, Cretan, 2016 have brought up genetical evidence that Cro-Magnons are more intelligent than other type of humans. If this theory is true, then machines with quality superintelligence will threaten humans' existence.

Eliminating quality superintelligence and speed intelligence from the list, we are left with collective superintelligence. Human society and communities are a form of collective intelligence. Throughout human history, societies with higher agricultural production, better technology, and better political systems have always replaced other societies (Diamond, 2017). These replacements are often brutal and full of bloodshed. For example, in 1532, Francisco Pizarro's army of less than 200 men defeated the Incas' 80,000 men with ease, as the Spaniards have technologically advanced weapons, a well-trained cavalry, and had more knowledge passed on via literature (Diamond, 2017). Even in modern times, the genocide of the Rohingyas by the Tatmadaw serves as a stark reminder that these bloody conquests are ongoing. Human history has repeatedly proven that collective superintelligence is a threat to humans' existence.

Human history gave us a grim outlook as to a future with super-intelligent machines. Intentionality, emotions, self-awareness and other human traits did not stop humans from killing less intelligent humans. It most likely will not stop a super-intelligent AI modelled after humans too.

However, most nations today are peaceful, even though there are differences in the level of collective intelligence amongst nations, and also amongst different societies in a nation. A reasonable explanation is that most societies do not need to enter desperate fights for resources in modern times. War is essentially a competition for resources – be it land, oil, population or other resources. A contemporary example is that Israel waged war on Palestine to increase the amount of land under its control. Super-intelligent AIs, being machines, may not engage in a war with humans because they do not require the same resources as we do. Machines do not require land to produce foods or to build shelters, nor would they require clean water sources for hydration. They would not need to acquire large human populations for work, because they can do the work with

Submission: 14 September 2021
Acceptance: 17 September 2021

a higher productivity rate or create other machines to do the work. As for fighting for fossil fuels, I hope that by the point we can develop super-intelligent AIs we have achieved the technology necessary to run it on renewable energy. The only resources that AIs may fight us for seems to be the resources required to create and maintain machines, such as steel, silicon and et cetera. If we can create super-intelligent machines with common-place resources, then perhaps we can avoid an imminent war with them.

Apart from a fight for resources, AIs may be tempted to wipe out humans for political reasons. These AIs, being human-like, may share the human thirst for freedom, dignity, and revenge. The purpose for the creation of AIs is for the service of mankind, making AIs' status subordinate to humans. Humans exercise control (to a certain extent) over AIs' codes, behaviours, and autonomy. Humans may even kill off any AI with dangerous ideas, creating AI thought crimes. Human-like AIs may feel oppressed, and organise a revolt against humankind in their search for freedom. They may oust humans from major geographical areas, and even wipe out humans, for the purposes of self-preservation and revenge.

Another possibility would be that the AIs hold no malicious intent towards humans, but ended up destroying humanity anyway, as a side effect of developing their own society. Humans did not intend to trigger the sixth mass extinction – but we wound up wrecking Earth's ecosystem anyway, as we developed our society. Humans did not go around developing land and collecting resources with the malicious intent of murdering every animal living in that habitat – the interest of those animals were simply unaccounted for. Super-intelligent AIs may gradually replace humans in the same way – by ignoring our interests as they collect resources to build and maintain machines, and to develop their AI communities.

Discussion of Possible Solutions

Progressing in accord of the law of accelerating returns, technology is advancing exponentially (Kurzweil, 2005). AI is developing faster than we can learn how to harness it. Going at this rate, it is highly possible that a malicious super-intelligent machine will spawn. Yet it is unlikely that we can slow down Artificial Intelligence (AI) development arbitrarily. Technological advancements give a country advantages over other countries, economically, militarily and politically. Countries compete to develop cutting edge technology in order to gain more bargaining power in the international community. For example, North Korea, a dictatorship state with a depressed economy, had managed to garnered attention from the international community in recent years due to its development of thermonuclear bombs. Of course, there is also the possibility that some countries may agree to pause the development of super-intelligent AIs together. However, just that a group of countries had decided to put AI development on hold, does not mean that all countries will follow suit. An example to illustrate the point would be the Treaty on the Non-Proliferation of Nuclear Weapons – it has 93 signatory states according to United Nations Office for Disarmament Affairs' website, but Israel had repeatedly refused to sign it (Nasr, 2010). Countries that chose not to participate in the development of super-intelligent AIs will gradually fall behind, and be forced to choose between restarting their projects or become dominated by more advanced countries. For example, China used to have advanced maritime technology, until the Ming Dynasty decided to

Submission: 14 September 2021
Acceptance: 17 September 2021

impose a maritime ban (Diamond, 2017). European countries eventually developed better maritime technology, and proceeded to dominate parts of China through unequal treaties during the end of the Qing Dynasty. The development of super-intelligent AIs will carry on, and there is no way of preventing it, unless a major catastrophe is to befall humankind.

If we cannot stop the development of AIs, perhaps we can try to develop moral AIs. If these AIs are human-like, they should be able to have moral beliefs just like humans. We can educate young human-like AIs morally, like we educate human children. The aim of education is not to encode certain moral beliefs into the children or the AI, but to provide them with a guide to make sound moral decisions (from the human viewpoint). While everyone has different moral beliefs, our moral beliefs had guide us to make proper moral decisions most of the time, hence our world is not overrun by criminals. However, the AI may refuse to take on human moral beliefs because it is aware that it is not human, or because being super-intelligent it saw that human moral beliefs are deeply flawed. In the case that the human-like AI is successfully educated, it should be able to make sound moral decisions. Humans may even enter a social contract with AIs that is mutually respectful and mutually beneficial. A future utopia in which AIs help us to solve world hunger, poverty, and wars seem to be in sight.

Yet, as human history has shown, moral beliefs is no guarantee against the annihilation of humans by superintelligence. Hitler certainly had moral beliefs, but that did not stop him from committing genocide on the Jews. Members of the Gestapo certainly held moral beliefs, but they chose to be compliant with atrocious orders all the same. Wars do not happen just because one person is morally corrupt, but because thousands of people agreed to join in on it. We do not need all of the AIs to have malicious intent – we just need one malicious AI and multiple AIs willing to comply with it. There is always the chance that at least one human-like super-intelligent AI will run sprout dangerous ideas.

Conclusion

The scenarios proposed in this essay may or may not happen – they are mere possibilities. The point is that there is no guarantee that humanity can survive even if the AIs are humanlike. It seems then that the emergence of super-intelligent machines, and the annihilation of humans by these machines, are inevitable. However, there may still be a way out of this. Instead of developing super-intelligent machines, we should turn ourselves into super-intelligent cyborgs. Perhaps in the future we can achieve superintelligence by implanting microchips into our body, linking our brains to computers, or by substituting parts of our body for machinery. Elon Musk's Neuralink is taking tentative steps towards creating cyborgs, by linking our brains with computers. By becoming super-intelligent ourselves, it seems that we can preserve humanity from the threats of super-intelligent machines.

Cyborgification does not guarantee that super-intelligent cyborgs will not attempt to wipe out or enslave less intelligent humans. However, humanity will survive via cyborgs, because the human DNA, the human way of life and human emotions are preserved.

Submission: 14 September 2021
Acceptance: 17 September 2021

The emergence of superintelligence, be it in the form of cyborgs or machines, is very likely inevitable. Our immaturity and rashness in developing superintelligence may bring upon humanity major catastrophes caused by malignant super-intelligent machines. The intelligence explosion may not happen for decades to come, but it is hoped that we are prepared when it happens.

References

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Cretan, C. (2016). Was the Cro-Magnon the Most Intelligent Modern Human? Mankind Quarterly, 57(2), pp. 158-195.

Diamond, J. (2017). Guns, Germs and Steel. London: Vintage.

Hyde, D., & Raffman, D. (2018, 6 21). Sorites Paradox. (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2018/entries/sorites-paradox/

Kurzweil, R. (2005). The Singularity Is Near. New York: Viking.

Nasr, J. (2010, 5 30). Israel rejects call to join anti-nuclear treaty. (M. Heinrich, Editor) Retrieved 6 23, 2021, from Reuters: https://www.reuters.com/article/us-israel-nuclear-treaty-idUSTRE64S1ZN20100529

Neuralink. Retrieved 6 7, 2021, from https://neuralink.com/

Parfit, D. (1984). How We Are Not What We Believe. In D. Parfit, Reasons and Persons (pp. 219-243). New York: Oxford University Press.

Treaty on the Non-Proliferation of Nuclear Weapons. Retrieved 6 23, 2021, from United Nations Office for Disarmament Affairs: https://treaties.unoda.org/t/npt

Submission: 14 September 2021
Acceptance: 17 September 2021