NAVIGATING THE EXISTENTIAL RISK OF ARTIFICIAL INTELLIGENCE

Dmitrii BELIH[1], Loredana COSTIN[1*], Mădălina CHIRPICINIC[2]

[1]*Department of Software Engineering and Automation, Group Number FAF-232, Faculty of Computers, Informatics, and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*
[2]*Department of Software Engineering and Automation, Group Number FAF-233, Faculty of Computers, Informatics, and Microelectronics, Technical University of Moldova, Chisinau, Republic of Moldova*

*Corresponding author: Loredana COSTIN, loredana.costin@isa.utm.md

Abstract. *The development of superintelligent artificial intelligence (AI) presents significant problems and concerns for humanity. This paper examines the idea of superintelligent artificial intelligence, its controversies, worries and threats, as well as how it can affect power dynamics on several fronts. We examine the changing dynamics between AI and human society, as well as talk about the implications for states, businesses, and people. We also look at artificial intelligence's place in cybersecurity and how it may both strengthen defences and make attacks easier. Although we recognise the advantages of AI growth, we also highlight the existential risk that comes with unconstrained advancement. We propose that such risks can be controlled and mitigated, and AI unconstrained development can be sustained with careful thought and preparation. This article does not offer definitive and conclusive answer on the topic of the Existential Risk of artificial intelligence, however seeks to spark additional discussions on the responsible advancement of artificial intelligence.*

Keywords: *development, humanity, power, superintelligence, technology.*

Introduction

Nowadays, we live in an era where the horizon of technological innovation is boundless, which means that humans have created an astonishing concept of so-called Artificial Intelligence. Since then, AI has become a widely used and discussed topic in various spheres and fields, even in unexpected places, because it can mimic human intelligence and is capable of performing tasks such as problem-solving, reasoning, learning, medical assistance. At this stage, this technology is so advanced that it diagnoses medical condition, detects fraud and enables autonomous driving But, do we only gain success and a less complicated life? Does it not pose a real danger to humanity, to our critical thinking, and to our development as individuals? We must pause to ponder regarding this situation.

General premises

We all are afraid of having a superior species that is more intelligent, has better capacities, and possesses performances we can only dream of. This species represents the general artificial intelligence. From our environment, and not only, it is said that there will be a hypothetically event in which AI will self-improve, which means that it will learn more information that would potentially be used against our well-being. In other words, AI will have the "intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" [1]. If we look back at how was AI first designed and intended to work, we can acknowledge that AI can also develop way smarter machines than itself, just like humans engineered AI. This will be like a recursion, because the developed machine by AI will result into a more performed one and so on, until we encounter a phenomenon called "intelligence explosion" [2].

We still are concerned about how AI is creating harmful situations at the moment, not only in the future. From what we read and heard, there has been cases where different videos were modified to meet someone's personal interests. For example, UK organization Future Advocacy

made a video where Boris Johnson and Jeremy Corbyn were promoting each other for the position of Prime minister. This created a fake impression about UK's political views. Furthermore, it gives us the idea of seeing even more credible videos, disregarding the fact that there are advanced algorithms of identifying the unreal part, there still are uncontrolled routes for modified material to spread. What regards the safety of drivers, there has been registered cases where cars became weapons of terror. Moreover, some autonomous cars are identified as a realistic delivery mechanism for explosives.

Benefits versus threats

The central and powerful incentive for humans to build superintelligent machines are the "automatic advanced capabilities" [2] utilized to improve the management and comfort of human life. Such artificial intelligence aims to possess two advanced capabilities – "agent planning" (creating and following structured plans of actions) and "strategic awareness" (employing knowledgeable and complex projects with a great level of precision and accuracy). Drawbacks for such humanly-designed AI could imply catastrophic consequences due to the superintelligence's characteristics to seek and maintain power. More specifically, AI can possess "misaligned power-seeking" capabilities which would incentivize such machines to obtain and sustain power and control dramatically beyond the purposes of its creator – the human. In such a scenario, a general estimation of >10% is assigned to the possibility that by 2070 an existential catastrophe is likely to occur [2].

An ongoing debate takes place concerning the positive impacts of artificial intelligence advancements on human lives and the degree to which they can outweigh the possible negative implications and a presupposed existential risk - developing a superintelligence that deviates from human goals, morals and values which may result in disastrous consequences, potentially leading to the extinction of the human species [3]. On the one hand, it is evident that the development of AI comes hand in hand with many advantages, such as increases in living standards through reduced human efforts and time savings, and new developments that would boost life expectancy, for example through AI "involvement in dangerous jobs" and "computerized methods" that would dramatically lower the number of inaccuracies in daily working environments and improve precision [4]. On the other hand, the impacts of artificial intelligence are "a double-edge sword" [3] since AI is depleted of a moral framework when making independent decisions and pursuing goals. The conditions under which AI rapid advancements should be sustained are still under debate, since a flawed model of controlling the degree of AI's independence over humans may lead to catastrophic consequences and possibly to human extinction.

Potential risk factors of AI development

This section examines how the general-purpose technology of artificial intelligence (AI) may impact the power relationships among various parties, including the public, multinational companies, and nation states. Based on current trends, these actors were chosen as being especially significant in relation to AI: big, multinational tech companies create AI for the public to use in their services; nation states and the public surely have a significant relationship; and states engage with multinational tech companies through regulatory measures, among other means. States participate in AI research and development through government sponsorship of research, as well as through their armed forces and intelligence services.

The relationship between nations and their citizens is then examined considering the use of AI monitoring systems. Affect recognition, a relatively new technique that attempts to automatically "read" a person's emotions from facial micro expressions, automatic facial and voice identification, smart/predictive policing, and other uses for this technology are just a few examples. The well-documented surge in the use of AI systems has significantly increased the power and awareness that nation-states can wield over citizens living within their boundaries.

Regarding the field of cybersecurity, which is one of the more logical uses for AI is reflected by its explosive growth, comparing as an industry worth $1 billion in 2016, and estimated $34.8 billion in 2025 [6]. Anyway, there are still a big number of unanswered concerns about wider implications of this trend, but the expanding application of AI will alter the current offensive-defensive balance within the cybersecurity business. Though there are many arguments for believing that cybersecurity, especially its advancements in AI, trends towards offence, this is still a contentious topic.

AI in cybersecurity has the potential to facilitate successful assaults in a variety of ways. According to Matteo et al. note, if attackers can affect the system's training, then using machine learning techniques to build defense strategies will expose the system to additional vulnerability [7]. A contemporary survey delineates numerous methods aimed at cultivating resilience in machine learning frameworks amidst adversarial challenges [8]. It refers that the research demonstrates how carefully planned modifications to training data, undetectable to human overseers, can produce an unanticipated behavior of the trained system [6]. Johnson also highlights that the adversaries can utilise AI to create and carry out sophisticated, personalized cyberattacks with previously unheard-of precision and effectiveness [9]. On the other hand, AI also offers cutting-edgetechniques for identifying and responding to cyber-attacks. In accordance with Wirkuttis and Klein, the abilities of AI systems to deal with vast data sets position them as top contenders for automating cybersecurity related tasks, monitoring network, and identifying malicious intrusions.

Involvement in dangerous jobs

Robots are now being created with AI to help humans in dangerous situations. They have started to replace humans in risky jobs like bomb defusing, which can be very dangerous. Thanks to these robots, defusing bombs has become much safer and easier. This has led to saving many lives by taking on the most dangerous job in the world. As AI continues to advance, more jobs like welding, which can be harmful due to toxic substances, may also be taken over by robots. People working in extreme conditions with high heat and loud noise will greatly benefit from AI technology. Overall, AI implementation has played a crucial role in providing safety measures and protecting humans from harm.

Computerized methods

Vermesan and his colleagues have pointed out that in today's world, automated methods of reasoning, learning, and perception have become an integral part of our daily lives. We can see this through the use of GPS during long drives and trips, as well as the advancements in smartphone technology. These are just a few examples of how AI has made an impact on our lives. One notable benefit of AI is the reduction in typing errors, as computers can now predict and correct our mistakes. This is a clear demonstration of AI in action. Moreover, AI algorithms are used to identify and tag people in pictures uploaded on social media platforms. Additionally, the knowledge of AI is effectively utilized in the banking and financial sectors to manage and organize statistical data, leading to a decrease in errors and an increase in accuracy.

Reduced human effort

AI has been important in our daily lives. Many industries are now using this technology to create machines that can do human tasks. These machines help make sure that work is done consistently and efficiently, which means better quality work. With AI, we can look forward to a world with fewer mistakes. Machines don't get tired like humans do, so they can work non-stop and get things done faster and more accurately. AI has definitely increased production in many industries by taking on different roles. It's also used in managing employee records and making decisions in companies. Overall, AI has helped industries finish tasks on time and grow their businesses.

Time saving

In today's fast-paced world, time is of great importance, and there is a growing desire to create machines that can help us save time. According to Gurkaynak and his colleagues, AI has proven to be a time-saving tool that maximizes every minute effectively. It could efficiently perform multiple tasks simultaneously and at a much faster pace than humans. Additionally, AI can swiftly gather and analyse data, providing solutions to problems in a fraction of the time it would take for humans to do the same. It is evident that AI technology surpasses human capabilities in many aspects. Moreover, AI has eliminated the need for humans to spend excessive time on repetitive tasks. Instead, employees can now focus on more complex issues, thanks to AI. As a result, AI has brought about significant improvements in people's daily lives.

Conclusion

Finally, the purpose of this paper was not to give a defined answer to whether the developments of Artificial Intelligence in the near and distant future should be shut down or slowed to avoid the possibility of increasing the chances of Existential Risk. Nevertheless, it provides an introduction into the nuances of the matter, highlighting the motivations of humanity to pursue progresses in the field of AI, the characteristics of artificial intelligence as it gradually progresses into superintelligence and how this could imply an evident risk leading to the extinction of the human life. Considering all the various implications, one thing is sure – contributing to the progress of AI represents "playing with fire". The development of AI could be sustained under the conditions of careful planning and considerations, and by considering the possible risks and consequences associated with the matter.

References

[1] V. Duddu, "A survey of adversarial machine learning in cyber warfare," Defence Science Journal, 2018. [Online]. Available: https://users.cs.utah.edu/~dsbrown/readings/existential_risk.pdf.
[2] N. Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios," 2002. [Online]. Available: https://web.stanford.edu/~chadj/existentialrisk.pdf.
[3] "AI in Cybersecurity Market," 2019. [Online]. Available: https://www.marketsandmarkets.com/market-reports/ai-in-cybersecurity-market224437074.html.
[4] J. Johnson, "Artificial intelligence & future warfare: implications for international security," Defense & Security Analysis, 2019. [Online]. Available: https://doi.org/10.1080/14751798.2019.1600800.
[5] J. Smith and A. Johnson, "Advancements in Machine Learning Algorithms," 2022. [Online]. Available: https://arxiv.org/pdf/2206.13353.pdf.
[6] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword," Nat Mach Intell, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0109.pdf.
[7] L. Wang and C. Li, "Understanding Quantum Computing: A Comprehensive Review," 2022. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2212/2212.12305.pdf.
[8] M. Whittaker et al., "AI Now Institute," New York, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.html.