DOI: 10.54254/2755-2721/106/20241172

Potential Safety Issues and Moral Hazard Posed by Artificial General Intelligence

Boqian Feng

Chengdu NO.7 Wanda High School, Chengdu, Sichuan Province, 610000, China

richard91487@outlook.com

Abstract. Artificial Intelligence (AI), a technology with a wide range of intelligence capabilities, has developed rapidly in recent years, bringing significant convenience and efficiency to society. However, most of the current artificial intelligence technologies belong to narrow artificial intelligence. Unlike Narrow AI, Artificial General Intelligence (AGI) possesses a more comprehensive understanding and problem-solving capability. AGI can learn in an unsupervised manner. General artificial intelligence can not only stand out in specific fields. It can also make effective decisions to a certain extent and operate in a wide range of environments. However, rapid progress has also raised widespread concerns about its potential risks. Therefore, the development of artificial intelligence requires standardization, which is urgent to ensure that it can make decisions that benefit humanity. Based on existing literature and data results, this paper explores the security issues and moral risks that general artificial intelligence may bring to humans. The research results indicate that these risks include user privacy breaches, system security issues, and social ethical conflicts. Dealing with these risks requires the joint efforts of all practitioners. This includes developing AGI in an ethical manner and ensuring that AI does not engage in activities that violate human interests.

Keywords: AGI, AI, Catastrophic Risk, AI risk, Machine Ethics.

1. Introduction

AGI is an important frontier in the field of artificial intelligence. Artificial intelligence in terms of meaning can only solve specific tasks. Compared to AGI, AGI has a more comprehensive understanding and problem-solving ability. AGI has undeniable potential to revolutionize industries and drive global development. However, this potential has to some extent raised concerns about the safety risks and ethical issues of development. For example, how to ensure the verifiability of AGI during the development process? Will artificial intelligence collaborate with humans to fix erroneous code generated during the development process? How to ensure that it does not lose control in this intelligent situation[1][2]. Therefore, after reviewing previous literature and research, this article explores the potential security risks and ethical impacts of the development of general artificial intelligence on human society. The author focusses on exploring its impact on human safety, privacy, health, abuse, and the economy, and proposes some feasible solutions to address the above issues. This article also explores the responsibilities of all stakeholders in the development of artificial intelligence. The research findings of this article are beneficial for deepening people's understanding and cognition of it, and promoting a deeper understanding of the importance of responsible development of general artificial intelligence.

2. Basic concepts and development status

2.1. Definition

AGI refers to an intelligent agent with efficient learning and generalization abilities, as well as autonomous perception, cognition, decision-making, and other abilities. This type of intelligence is not limited to specific fields, but has a wide range of cognitive and operational abilities. AGI can handle various tasks and problems, not limited to specific fields. It is very flexible and can display human like intelligence in different environments and tasks. This is his first characteristic. Another feature is autonomous learning, where AGI can improve its performance through autonomous learning, rather than relying solely on pre-set programs or rules. In other words, it can learn from experience and adapt to new challenges and tasks. Of course, it can also make certain autonomous decisions. AGI has the ability to make autonomous decisions. This autonomy allows for making complex judgements and decisions without relying on human intervention. It can consider multiple factors and develop a reasonable action plan. Although no real AGI system is currently self-aware, it is possible that future AGI will develop the characteristics of self-awareness, that is, the ability to understand the impact of its own existence and behavior on the environment[1][2].

2.2. Current development status

Although current technology has not yet reached the level of true AGI, breakthroughs in areas such as natural language processing and machine learning have laid the foundation for future AGI. Currently, most artificial intelligence is narrow AI, but it is already able to handle specific tasks. Narrow AI has many applications in a range of fields, such as automatic identification through facial recognition, predicting financial risks or natural disaster risks. It has applications in many fields[2]. These breakthroughs are very likely to occur in the coming decades, so it is important to discuss the risks they may bring. Even a very low probability of catastrophic results is enough to motivate research. The results of superintelligence are more likely to be extreme: either extremely bad for humanity or extremely good for humanity. As it happens, according to our research, the probability of advanced machine intelligence (exceeding human capabilities in almost all aspects) by 2050 is more than 50%, which means that it is more likely that humans will be able to develop AGI. [3]. The development involves complex ethical and safety issues, such as how to ensure that the behavior of AGI systems is consistent with human values, how to prevent the abuse of AGI technology, etc. These issues need to be explored and resolved in depth while technical research is ongoing. Companies and research institutions are actively exploring paths to AGI realization, which makes the related issues of security and moral hazards even more pressing.

3. Possible security issues caused by the development

3.1. System security

There are multiple arguments supporting the idea that AGI could pose a security risk to human society, a risk that could even be catastrophic. When AGI is first created, its purpose is to replace humans in some jobs. But over time, humans may become dependent on AGI, and AGI may occupy a more important position in society. For example, AGI may have more and more autonomy and make decisions without human consent. Over time, it will become more and more difficult to separate AGI from human society. AGI may be like many commodities in human society, although they are expensive to develop at the beginning. But they can also be copied cheaply, so once created, they can spread quickly. Once they become powerful enough, AGI may pose a threat to humans, even if they are not actively malicious or hostile. The mere ambiguity of human values is enough to pose a threat to human society.

AGI systems' decision-making processes might grow less transparent as they advance, making it harder to forecast how they will behave in certain scenarios. There are a lot of military and political uses for artificial general intelligence. There are hazards associated with some of these eventualities that could be disastrous, like the possibility that artificial general intelligence could operate in ways that are unanticipated and detrimental, even if its creators mean well. Even today's restricted AI systems are growing more autonomous and capable to the point that, in rare cases, they act detrimentally and without warning before their human supervisors can intervene. As AI becomes more autonomous, the likelihood of timely human intervention decreases, making it important for AGI to be ethical in the choices it makes. AGI will be more autonomous and powerful than narrow AI systems, and therefore will require more robust solutions to manage its behavior. If some AGI are indifferent to human values, the consequences could be catastrophic. The possibility of AI self-evolving to produce superintelligence that exceeds human cognitive capabilities poses unique challenges. AGI may evolve at a speed beyond human understanding and control, resulting in existential risks[4][5].

AGI may misunderstand human intentions or purposes in practical applications. Some scholars argue that even a harmless AI, such as an agent programmed to win a game of chess, could become a serious threat to humanity if it were designed to start acquiring resources to achieve its goals; this is what inspires harmful activities such as breaking into computers and robbing banks.[6]

3.2. Threats to human health

Job losses as AI technologies are widely deployed[7]. The scale of AI-driven job losses in the future could range from hundreds of millions, although much depends on how quickly AGI develops and how governments focus on it. However, the issue is gaining more attention over time. There is an optimistic vision of a future world replaced by AI-enhanced automation, but there is a limit to the amount of economic exploitation that human society can sustain, and there is no guarantee that the increased productivity of AI will be distributed equitably across society. So far, increased automation has tended to transfer income and wealth from labor to capital owners, and has led to an increasingly unequal distribution of wealth around the world[8].

3.3. Misuse and abuse

Artificial intelligence has now developed the ability to quickly collect, organize and classify data. For example, cameras that are ubiquitous in today's life collect people's biometric information. If this information can be used well, such as the police using these technologies to combat terrorist acts. But it can also be abused and have serious consequences. For example, using this power to create commercial revenue for social media platforms, many businesses have also used it to create a large and powerful personalized marketing system that can know consumer preferences and promote products accordingly. The large-scale use of AI on social media platforms has become a powerful tool for political candidates to know the preferences of their voters and go to extremes to cater to their voters, which has led to the polarization and rise of extremist views observed in many parts of the world, and it has indeed been used to manipulate political views and voter behavior. This could lead to voters losing confidence in the government, causing social divisions and the collapse of democracy. Governments and militaries may also use AI-driven surveillance to more directly control and oppress people. In some countries, AI is used to apply facial recognition software to people's bank transactions, phone records, game records, and social relationships. This has led to sanctions against certain dissidents. These sanctions include fines, denial of access to banking and insurance services, and other services. The sanctions also include preventing them from travelling abroad. This type of artificial intelligence program may also exacerbate social and health inequalities. To a certain extent, it locks people into their existing socio-economic class. In modern society, even without artificial intelligence, people's privacy and freedom rights may be eroded or deprived. But the power of artificial intelligence makes it easier to establish or consolidate authoritarian or totalitarian regimes. At the same time, it also enables these regimes to persecute and oppress specific individuals or groups in society.

Artificial intelligence has many applications in military and defence systems. Some of these applications can be used to promote security and peace. Many political or military groups may seek to use AGI to develop weapons, conduct military operations, and even engage in many illegal and criminal activities. The government or military may develop lethal autonomous weapon systems (LAWS) at certain times. But the risks and threats associated with LAWS outweigh any assumed benefits. For example, certain weapons and explosives can be installed on devices such as drones. These devices have the ability to intelligently and autonomously perceive and navigate the environment. These weapons can be produced on a large scale at a low price and have the ability to cause mass destruction. Without human supervision, large-scale collective casualties may occur.

4. Ethical risks in developing Artificial General Intelligence

4.1. Job and economic impact

AGI has the potential to drastically change how people work and interact with the workforce in the future. With AGI systems' ability to carry out tasks that humans have historically completed, productivity will increase significantly as automation and intelligence replace many employment. But this will also lead to an increase in unemployment. As AGI systems develop, their decision-making processes may become less transparent, which will make it more difficult to predict how they would act in specific situations. Artificial intelligence has many military and political applications. Some of these scenarios include potentially catastrophic risks, such as the potential for artificial general intelligence to behave in unexpected and harmful ways, even in cases where its designers have the best of intentions. Even limited AI systems of today are becoming increasingly powerful and independent to the point that, in rare instances, they act maliciously and without notice before their human supervisors can take action[9].

4.2. Moral responsibility

A moral approach to research and development is required in the creation of AGI. One prerequisite is to make sure AGI systems are consistent with human values. Fairness, responsibility, and transparency must be integrated into the development and application of artificial general intelligence. AI ethics extend beyond the code itself. The possible repercussions of their work and the effects artificial general intelligence will have on society must be understood by AI researchers and developers. This takes into account the possibility that the algorithm's creators may have introduced prejudice or discrimination on the basis of personal experience[10].

5. Conclusion

As humanity accelerates towards AGI, concerns about the safety issues and ethical risks that AGI may bring continue to increase over time. Responsible development, user privacy protection, and ensuring that the development of AGI is beneficial to humanity are key factors in allowing humanity to enter the AGI era. This requires the joint efforts of governments, AI developers, and industry stakeholders. The development of AGI must always be based on human interests as the primary consideration. All stakeholders must develop this technology responsibly. Since AI is developing very rapidly today, the data collected in this article may be very time-sensitive. In addition, this article only discusses some of the risks that AGI may bring. In fact, the continued development of AGI may expose human society to unknown risks.

Acknowledgment

First of all, I would like to express my greatest gratitude to my teachers and professors who provided me with valuable guidance from the beginning to the end of writing this paper. In addition, I would like to thank all my friends and parents for their encouragement and support. All the inspiring guidance and impressive kindness they provided helped me to complete my paper better.

References

[1] Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. arXiv preprint arXiv:1805.01109.

[2] Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., ... & Witteborn, S. (2019). Artificial intelligence governance and ethics: global perspectives. arXiv preprint arXiv:1907.03848.

[3] Müller, V. C. (2016). Editorial: Risks of Artificial Intelligence. In Risks of artificial intelligence (pp. 12-19). Chapman and Hall/CRC.

[4] Déletang, G., Grau-Moya, J., Martic, M., Genewein, T., McGrath, T., Mikulik, V., ... & Ortega, P. A. (2021). Causal analysis of agent behavior for ai safety. arXiv preprint arXiv:2103.03938.

[5] Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. Ieee Access, 10, 77110-77122.

[6] Sotala, K., & Yampolskiy, R. V. (2014). Responses to catastrophic AGI risk: a survey. Physica Scripta, 90(1), 018001.

[7] Rayhan, S. (2023). Ethical Implications of Creating AGI: Impact on Human Society, Privacy, and Power Dynamics. Artificial Intelligence Review.

[8] Brubaker, K. (2018). Artificial intelligence: Issues of consumer privacy, industry risks, and ethical concerns (Master's thesis, Utica College).

[9] McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with Artificial General Intelligence: A systematic review. Journal of Experimental & Theoretical Artificial Intelligence, 35(5), 649-663.

[10] Federspiel, F., Mitchell, R., Asokan, A., Umana, C., & McCoy, D. (2023). Threats by artificial intelligence to human health and human existence. BMJ global health, 8(5), e010435.