Abstract

This paper discusses the ethical concerns surrounding the increasing autonomy of AI systems and their potential to make decisions without human intervention, which has sparked a heated debate about the need for human oversight and control in AI development and deployment. The paper explores various strategies to ensure human oversight and control in AI, including limiting the AI system's learning scope, avoiding centralized AI, establishing a central committee for AI model approval, and developing a new field of study for enhanced AI model explainability and control. Through a review of existing literature and original research, the paper evaluates the effectiveness of current strategies and proposes potential new approaches to address this critical issue. The paper seeks to contribute to the ongoing discourse on the ethical and societal implications of AI and to inform policy and practice to ensure the safe and responsible development and deployment of AI. Index Terms—artificial intelligence, AI development, internet access, centralized AI, AI model approval, AI explainability, AI control, ethical implications, unintended harm, responsible AI development, human control, well-being, society

Keywords: artificial intelligence, AI development, internet access, centralized AI, AI model approval, AI explainability, ethical implications, responsible AI, human control AI, society and AI

1. Introduction

The rapid development of artificial intelligence (AI) has brought about numerous benefits and opportunities for society, from improved healthcare and transportation to more efficient manufacturing and financial services. However, the increasing autonomy of AI systems and their potential for making decisions without human intervention has raised ethical concerns about the implications of AI for human autonomy, control, and decision-making(University, 2021).

The risk of AI systems taking over human decision-making processes has sparked a heated debate about the need for human oversight and control in AI development and deployment(Statista, 2023a). While some argue that AI should be allowed to operate autonomously to achieve its full potential, others caution that the risk of autonomous decision-making and supremacy could threaten human autonomy and dignity.

As artificial intelligence (AI) technology advances, it has become increasingly important to ensure that AI systems remain under human control and are used for ethical purposes. This requires addressing various concerns, such as the potential misuse of AI models, the development of AI weapons, and the risk of AI surpassing human intelligence. In this context, several strategies have been proposed, including limiting an AI system's learning scope by restricting internet access, avoiding centralized AI, establishing a central committee for AI model approval, and developing a new field of study for ethical AI(Naughton, 2023)(Peacock, 2021), enhanced AI model explainability, and control. Each of these strategies aims to safeguard against the autonomous decision-making and supremacy of AI and ensure that it remains a tool for human benefit.

2. Motivation

The rapid advancement of Artificial Intelligence (AI) has the potential to revolutionize our world in unprecedented ways. However, this progress has also raised serious concerns about the possibility of AI surpassing human intelligence and becoming a threat to our existence. As noted by renowned AI researcher Geoffrey Hinton, if AI becomes much smarter than humans, it could be very adept at manipulation and may be difficult to control(Naughton, 2023). Therefore, there is a pressing need to address the challenge of preventing AI from surpassing human intelligence and becoming a potential danger(Naughton, 2023). To achieve this, various approaches must be considered. One approach is to limit the scope of AI systems by restricting their internet access. This would ensure that AI models are not exposed to destructive information that could lead to harmful decision-making. Additionally, avoiding centralized AI models and assigning specific tasks to AI would increase accuracy and prevent the potential for one AI model to become dominant and destructive. Furthermore, establishing a central committee for AI model approval could enable better control over the use of AI models, preventing their misuse for potentially harmful applications. Additionally, a new field of study should be developed to enhance AI model explainability and control, enabling us to identify better ways of using AI while minimizing the risk of it surpassing human intelligence. Lastly, ethical concerns related to the development and deployment of AI weapons must be addressed. The potential loss of human control over these systems could have catastrophic consequences, and thus, there is a need to carefully consider the implications of developing AI weapons. Overall, these various approaches are essential to ensure that AI remains a safe and beneficial technology for humanity. This paper aims to contribute to the ongoing debate on

this important topic, by exploring various strategies for preventing AI from surpassing human intelligence and ensuring that it remains under our control.

3. Discussion

AI is advancing at a fast pace, which has raised concerns about the need to limit and prevent its potential negative consequences. The aim is to prevent AI from becoming like nuclear weapons due to the potential risks it poses. The Internet is a full of false information according to a recent study by the Central Statistics Office (CSO), around 62% of all information on the internet is unreliable(BusinessDIT, 2023). AI can mislead a generation based on unreliable information. Even people may overly depend on AI which will make them hinder their research potential and critical thinking skills.

To illustrate, consider the following examples: With around 2.96 billion monthly active users and 1.96 billion daily active users, Facebook is undeniably one of the most popular social media platforms(Statista, 2023b). However, the content shown on Facebook is personalized based on the user's interests(Statista, 2023b)(Lee et al., 2022). This means that if a user reads an article or story that spreads hate about a particular religion or country, they are likely to receive more similar content regularly. Consequently, this may lead to the user developing a negative bias towards that religion or country. The same things apply to other social media as well.

Google is a highly utilized software, with over 5.16 billion internet users worldwide as of January 2023, representing 64.4% of the global population(Statista, 2023a)(Walter, 2019). However, a significant issue with Google is that search results may differ based on the location and personalization of the user(Statista, 2023a)(Walter, 2019)(Lee et al., 2022). For instance, searching for the same topic from different countries may yield varying results, which may not always be objective.

ChatGPT is even more worst its providing direct answers to students can create dependency and hinder their research potential and critical thinking skills(Abdullah et al., 2022). This could have negative consequences on an entire generation. It's important to encourage students to seek out answers and develop their own problem-solving abilities.

Artificial intelligence (AI) can be a valuable tool in advancing society, but it's crucial to have a governing body that evaluates its impact and determines which AI technologies will have a positive impact on society. Additionally, we need to develop a new field of study focused on improving AI models through enhanced model explainers, which can analyze and interpret the decision-making processes of AI models. This will help us to find better solutions and ensure that AI is used in a responsible and ethical manner.

4. Survey Data on Human Oversight and Control over AI

The data from the surveys conducted by the Pew Research Center and Monmouth University reflect the diverse perspectives and opinions regarding the impact of artificial intelligence

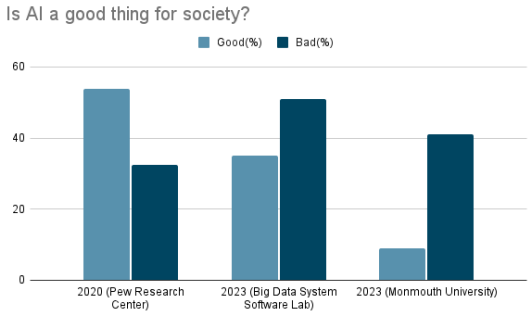

Figure 1: Public Perception of AI's Impact on Society: A Shift in Views from 2020 to 2023

(AI) on society. These findings highlight the ongoing debate surrounding the potential benefits and drawbacks of AI technology.

The bar chart displays varying opinions on the impact of AI on society. In 2020, the majority (53.75%) viewed it positively, but in 2023, opinions shifted with a decrease in positive views (35%) and an increase in negative views (51%).

In the 2020 Pew Research Center survey with 150,000 participants, around 32.45% of respondents expressed concerns about the negative impact of AI, while 53.75% believed it would have a positive influence(Center, 2020). This indicates that a significant portion of the population recognizes the potential benefits AI can bring, such as advancements in various industries, increased efficiency, and improvements in quality of life. However, there are also concerns about uncontrolled AI development and its potential negative consequences, such as loss of control, unintended harm, and ethical dilemmas.

In 2023 Both the Big Data System Lab and Anthony J. James independently conducted surveys on LinkedIn, collecting data by asking the same question, which included 10,038 participants, revealed that 53% of respondents viewed AI as harmful, while 35% considered it to be beneficial. This indicates a shift in public opinion, with a larger proportion of people expressing concerns about the potential negative impact of AI compared to the previous survey. This shift may be influenced by increased awareness of the risks associated with unregulated AI development or specific incidents related to AI technologies.

Additionally, the Monmouth University survey included 805 participants and showed that 41% believed AI would cause more harm than good, while 9% believed it would have more positive than negative effects. Interestingly, 46% of respondents perceived an equal balance between the positive and negative aspects of AI(Institute, 2023). This suggests a more nuanced perspective, acknowledging both the potential benefits and risks associated with AI technology.

In light of these findings, the paper emphasizes the importance of responsible and ethical AI development. It suggests several strategies to mitigate the risks associated with AI, including limiting the learning scope of AI systems, avoiding centralized control, establishing oversight committees for AI model approval, and enhancing AI model explainability and control. The paper also touches upon the ethical implications

of AI weapons, emphasizing the need to maintain human control and minimize unintended harm.

5. Ensuring Safe and Ethical Implementation of Artificial Intelligence

As artificial intelligence (AI) continues to evolve, the question of whether it could one day surpass human intelligence has become a topic of concern for many researchers and experts in the field. The potential risks associated with creating an AI that is much smarter than humans are numerous, including the possibility of it becoming uncontrollable and making decisions that could be harmful to society. Therefore, it is essential to explore ways to prevent AI from surpassing human intelligence while still maximizing its potential for good. In this paper, we propose several strategies for preventing AI from becoming too intelligent, drawing on insights from leading AI experts and existing literature. These strategies include limiting AI's learning scope by restricting internet access, avoiding centralized AI, establishing a central committee for AI model approval, creating a new field of study for enhanced AI model explainability and control, and considering the ethical implications of developing AI weapons. We will explore each of these strategies in more detail, highlighting their potential benefits and limitations, as well as offering recommendations for how they can be implemented effectively.

5.1. Limiting AI System's Learning Scope by Restricting Internet Access

The internet is an endless resource of information, and numerous authors have written about how machine learning can breach limitations and acquire knowledge independently. Similarly, there are several movies like I, Robot, Terminator, and The Matrix that explore this concept. However, if an AI system were to learn everything from the internet and these movies, it may begin making destructive decisions autonomously. Therefore, it is crucial to ensure that we monitor what our AI systems are learning and where they are learning from.

5.2. Avoiding Centralized AI

It is recommended to avoid having a centralized AI, as it has the potential to gain control over all other AI models and amalgamate their collective learning, which may result in it becoming both dominant and destructive. AI should be assigned a specific task to focus on, as this would result in more accurate outcomes and prevent the AI from making decisions outside its designated area of learning.

5.3. Establishing a Central Committee for AI Model Approval

One solution to prevent the potential misuse of AI models is to establish a central committee responsible for evaluating and approving any AI models before publication. The committee would review the models and hold a vote to determine their potential for both good and bad applications, similar to an election. Based on the outcome of the vote, the committee would decide whether or not to allow the model to be published. This would ensure that the control of AI remains in the hands of the committee, enabling them to halt any potentially harmful applications without requiring permission from the developer company.

5.4. A New Field of Study for Enhanced AI Model Explainability and Control

A novel field of study should be developed to explore ways of preventing AI from surpassing human intelligence. This field would focus on creating Ethical AI which will understand social norms and will be more social and also enhance AI model explainers that could analyze and interpret the decision-making process of AI models. These studies would aim to discover new methods of controlling AI and maximizing its potential while minimizing the risk of it becoming more intelligent than humans. By understanding the underlying mechanisms of AI decision-making, we can identify better ways to use AI and ensure that it remains under our control.

5.5. The Ethical Implications of Developing AI Weapons

The development and deployment of AI weapons raise significant ethical concerns. One major issue is the potential loss of human control over these systems, as AI weapons could make decisions and take actions without human intervention. Additionally, the use of AI weapons could lead to unintended consequences, such as targeting the wrong individuals or causing harm to civilians.

6. Summary and conclusions

The advancement of artificial intelligence (AI) technology presents a myriad of opportunities and challenges for society. On one hand, AI has the potential to revolutionize industries, improve efficiency, and enhance quality of life. On the other hand, the uncontrolled development of AI may lead to negative consequences, such as loss of control, the potential for unintended harm, and ethical concerns.

This paper has explored several strategies to mitigate the risks associated with AI development, including limiting an AI system's learning scope by restricting internet access, avoiding centralized AI, establishing a central committee for AI model approval, and developing a new field of study for enhanced AI model explainability and control. The ethical implications of developing AI weapons have also been discussed, including the loss of human control and the potential for unintended harm.

Overall, this paper highlights the importance of responsible and ethical AI development. By implementing these strategies, we can ensure that AI development remains beneficial for society while minimizing the risks associated with uncontrolled development. It is crucial that we work together to ensure that AI is developed in a way that benefits humanity and does not pose a threat to our well-being.

References

Abdullah, M., Madain, A., Jararweh, Y., 2022. Chatgpt: Fundamentals, applications and social impacts, in: Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), Milan, Italy. pp. 1–8. doi:10.1109/SNAMS58071.2022.10062688.

BusinessDIT, 2023. 62 percent of unreliable information on the internet in 2022. BusinessDIT URL: https://www.businessdit.com/fake-news-statistics/ .

Center, P.R., 2020. Publics express a mix of views on ai, childhood vaccines, food and space issues. Pew Research Center URL: https://www.pewresearch.org/science/2020/09/29/publics-express-a-mix-of-views-on-ai-childhood-vaccines-food-and-space-issues/ps2020

Institute, M.U.P., 2023. Monmouth university poll: Few americans believe 2020 election outcome was legitimate. Monmouth University Polling Institute URL: https://www.monmouth.edu/polling-institute/reports/monmouthpollus021523/

Lee, J., Kim, C., Lee, K.C., 2022. Exploring the personalization-intrusiveness-intention framework to evaluate the effects of personalization in social media. International Journal of Information Management 66, 102532. doi:10.1016/j.ijinfomgt.2022.102532.

Naughton, J., 2023. Geoffrey hinton, 'godfather of ai', quits google and warns of dangers of machine learning. The Guardian URL: https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning .

Peacock, A., 2021. Is ai good or bad, and who decides? Engineering and Technology Magazine URL: https://eandt.theiet.org/content/articles/2021/08/is-ai-good-or-bad-and-who-decides/ .

Statista, 2023a. Google - statistics and facts. Statista URL: https://www.statista.com/topics/1001/google .

Statista, 2023b. Number of monthly active facebook users worldwide as of march 2023. Statista URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide .

University, B., 2021. Artificial intelligence is changing how we interact with the world. Brown University News and Events URL: https://www.brown.edu/news/2021-09-16/ai100

Walter, R., 2019. Google engineer reveals search engine bias. Mind Matters URL: https://mindmatters.ai/2019/07/google-engineer-reveals-search-engine-bias .