Department of Electrical & Computer Engineering

ENCS4210 - Computer Engineering Ethics

Ethics of AGI Systems Development

Prepared By : Amro Abu Hashish – 1180620

Instructor : Abdel Salam Sayyad

Date : 18-6-2023

Introduction

Artificial intelligence (AI) has rapidly developed over the past few years, and experts predict that it will continue to evolve at a staggering pace. As AI technologies advance, there is increasing concern about the ethical implications of their development, particularly within the scope of generalization and reliability, which are the hallmarks of "Artificial General Intelligence".

Let's take a step back and define Artificial General Intelligence (AGI), AGI refers to intelligent machines that possess human-like reasoning, perception, and problem-solving abilities with the potential to perform a wide range of complex tasks, including those that require creativity, deep abstract thinking, and adaptation to new situations that have not been explicitly programmed or encountered before, differentiating it from narrow AI, which is usually designed for a specific task.

The achievement of AGI is becoming the topic of debate within the scientific community, raising many questions about its potential benefits and dangers and how to address them, questions such as: Will AGI be really achieved? What could be the dangers of AGI? and who should be in control? This paper focuses on exploring the ethics of AGI systems development and attempts to answer some of these critical questions.

Achievement of Artificial General Intelligence

One of the most fundamental questions surrounding the development of AGI is whether or not it will be achieved, and if so when, and how can we measure it?

The rapid technological advances in recent years, paired with the emergence of large language models such GPT, led some experts to believe that AGI is just around the corner, given we have endless sources of data, and increasingly greater computational power according to Moore's Law, enabling more complex and sophisticated algorithms development. However, others argue that it may never be

possible to achieve a genuine AGI, given the complexity and the current unpredictability of human cognition due to our lack of understanding of human consciousness, resulting in the difficulty of representing and replicating it in machines, furthermore, an AGI must be ethical and incorruptible, ensuring value alignment, and avoiding biased or harmful behavior, which requires a deeper understanding of human values and ethics that we humans have yet to fully grasp.

The way deep neural networks are designed and function makes them extremely unpredictable and uninterpretable, like a black box with gibberish mathematical equations in it that take an input and gives an output based on said equations, but there is no way for us to really grasp the actual flow in which the machine made the decision, making it extremely challenging to debug models, to gauge and measure how close we are to achieving AGI, as well as what potential dangers we may face once it is completed.

For an algorithm to be considered an AGI, it must exhibit various characteristics, including linguistic fluency, intent, empathy, and an understanding of context and complex ideas far superior to the narrow AI systems that we have today, it must also be able to solve mathematical problems, compare and analyze data, quickly learn from personal experience.

GPT-4, a language model developed by OpenAI, has recently astounded the world with its ability to generate coherent and seemingly human-like content, leading some to speculate that it could be the closest humans have got to achieving AGI [1], with its ability to understand context, express language fluently, learn quickly from personal experiences, and reason across domains and make decisions, but GPT falls short in many other areas necessary to be classified as a true AGI, such as the ability to exhibit empathy, or the capacity to understand complex ideas and situations that require human-like understanding and dissecting.

GPT-4 was tested through various benchmark tests, and its performance varied widely. Starting with critical thinking and problem-solving, various GPT versions were tested on Leet-Code problems categorized into three levels of difficulty. Although GPT-4 had a much better performance than GPT-3 in solving them, GPT-4 was only able to solve 6% of hard problems, 26% of medium difficulty problems, and 75% of easy problems, suggesting that something is going wrong either in the way GPT comprehends problems, or problem-solves technical problems, which was a surprise as algorithm coding is well supported with resources online, and a major concern when it comes to developing AGI systems.

At the same time, GPT-4 performed impressively answering Multiple-choice questions in 57 subjects (professional & academic) with an evaluation score of 86.4% compared to 70% for GPT-3, and in Commonsense reasoning around everyday events with an accuracy rate of 95% compared to 85.2% for GPT-3.

With the quick improvement of GPT-4 in comparison to GPT-3, it seems like we're taking big steps towards possibly achieving AGI, but it's important to keep in mind that we still have a long way to go in developing a true AGI that can match or exceed the abilities of human intelligence in various domains and situations.

Exploring the Potential Dangers of AGI Realization

As AGI development progresses, it is crucial to examine the potential dangers that could arise from its realization. One of the primary dangers is that AGI may surpass human intelligence and pose a threat to humanity. For example, an AGI system that has access to vast amounts of information and data could potentially make decisions that have catastrophic consequences for humanity. The creation of an AGI system that is not governed by ethical principles and values could lead to severe implications for society. Moreover, the development of AGI raises concerns about job displacement and its effects on society. In addition, the implementation of AGI systems may also present security risks, information gathering, and privacy

concerns that could lead to a violation of individual rights. in many recent cases, customers were surprised by the data leakage, and privacy concerns that could lead to a violation of individual rights. For example, Open Ai developers had to make their model dumber for safety reasons that highlight the importance of considering and mitigating potential risks and ethical concerns associated with AGI development.

"50% of AI researchers believe there's a 10% or greater chance that humans go extinct from our inability to control AI", according to a survey conducted by the University of Oxford. Everything around us can be translated into data, and if an AGI system is not developed with the proper safeguards in place, it could have disastrous consequences. Given the potential risks, it is essential for companies and individuals involved in developing AGI systems to accept responsibility for ensuring these systems are developed ethically and with safety as a top priority.

Responsibilities of Companies and Individuals in AI Development

As AI systems continue to gain prominence in various domains, companies and individuals involved in their development must be conscious of their ethical responsibilities. They have a moral obligation to use technology responsibly and ethically, with the ultimate aim of benefiting society as a whole. This involves ensuring that the AI systems they develop adhere to ethical principles and values, such as fairness, transparency, and accountability. Additionally, companies and individuals must prioritize the safety and well-being of humans over profits or personal gain when developing AI systems. The ethical principles outlined by the IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems, such as transparency, incorruptibility, responsibility, audibility, and usability, must be integrated into the AI development process. Furthermore, it is crucial for individuals and companies involved in

AI development to engage with the wider community and stakeholders to understand their concerns about the technology.

Analyzing the Intelligence of GPT-4: Is it AGI?

While GPT-4 may exhibit impressive capabilities in language processing and other tasks, it falls short in areas necessary to be classified as a true AGI system. It is important to note that having advanced language processing abilities does not necessarily equate to possessing human-level intelligence in various domains and situations. Therefore, while GPT-4 may be a substantial development in AI language models, it is not yet considered AGI and still has a long way to go in bridging the gap between AI and human intelligence. GPT 4 was only able to solve 4 hard-leet code questions out of thousands, GP3 was able to solve none, and this is just one example demonstrating the limitations of language models and their ability to perform tasks beyond language processing. Depending on how intelligence is measured, For example, GPT 4 has no real-time memory and also lacks the ability to solve most of the problems that require thinking several steps in advance. But on the other hand, how it was trained for example to classify images or perform certain specific tasks can be considered a form of specialized intelligence. and how the working methodology it uses simulates the human brain, it is clear that GPT-4 may exhibit certain forms of intelligence, but it lacks the general problem-solving abilities of a true AGI system. One more crazy piece of information to be considered is that the experts predict the ability of these models to translate the activity of human brains, by this granting the ability to understand what a person has dreamed about without even the person saying anything about it.

Open-source vs. Closed-source Debate for Advanced AI Systems and LLMs

There is much debate surrounding whether advanced AI systems, such as GPT-4 and ChatGPT, should be open-source or closed-source. Open-source systems allow for greater transparency and

collaboration, which can lead to faster innovation, whereas closed-source systems offer more control over the technology. However, the potential dangers associated with advanced AI systems make it crucial to prioritize transparency and collaboration. Therefore, it is recommended that advanced AI systems and LLMs be open-source to ensure greater transparency in their development and to allow for collaboration among various stakeholders.

Transparency and Accountability are essential components in the development of AI systems, as they ensure the responsible use and deployment of these technologies.

Collaboration and Innovation are also important factors, as they allow for a wide range of perspectives and ideas to be considered in the development process.

Security and Intellectual Property concerns need to be addressed, but should not be prioritized over transparency and collaboration when it comes to the development of advanced AI systems.

Control of Advanced AI Systems: Private Companies vs. Non-Profit Organizations vs. Government

Given the potential dangers associated with advanced AI systems, it is important to consider who should be in control of their development and deployment. Private companies may prioritize profits over ethical concerns, and non-profit organizations may lack the necessary resources and expertise. Therefore, the government should play a role in regulating and overseeing the development of advanced AI systems to ensure that ethical principles are upheld and societal benefits are prioritized over individual gain. The role of the government in regulating AI systems is crucial for ensuring that these technologies are developed and implemented responsibly. The government can establish guidelines and regulations that hold developers accountable for the ethical use and deployment of AI systems. Additionally, the government can also establish ethical standards for AI development that prioritize human values and

societal benefits over individual or commercial interests. The government can also invest in the research and development of advanced AI systems, ensuring that they are developed in a responsible manner. However, it is important to ensure that the government does not stifle innovation and progress in AI development by implementing overly restrictive regulations. It is important to strike a balance between regulation and innovation so that the potential benefits of advanced AI systems can be realized without compromising ethical principles or jeopardizing societal welfare. In conclusion, the development of advanced AI systems requires a responsible and collaborative approach that prioritizes ethical considerations over individual gain. Governments should play a central role in regulating and overseeing the development of these systems to ensure that they align with human values and promote social good. Sources suggest that the governance of AI systems requires a collaborative effort across different stakeholders, including non-governmental organizations, industry, and governments. Effective governance controls and audits should be established to enforce ethical principles, and inclusive and transparent governance processes should be fostered to shape the development of AI technologies in a responsible manner. Furthermore, ethical standards and regular auditing are key to maintaining compliance with regulatory frameworks governing the use and adoption of AI. Overall, it is essential to approach the development of AI systems with a sense of responsibility and accountability to ensure that these technologies are used for societal benefit. "This is an exciting time for AI governance. The AI community is moving beyond high-level principles and starting to actually implement new governance measures. I believe that structured access could be an important part of this broader effort to shift AI development onto a safer path. We are still in the early stages, and there is plenty of work to be done to work out exactly how structured access should be implemented" *Toby Shevlane says.* In line with Toby Shevlane's statement, there is a need for continued efforts to develop new governance measures that promote

responsible AI development. While high-level ethical principles are important, they alone cannot ensure that AI systems operate in an ethical and trustworthy manner. Strong governance controls and auditing mechanisms are required to enforce ethical principles and reduce the impact of any potential ethical exposure.

Role of Governments in AI System Regulation

Governments have a crucial role to play in regulating AI systems. As AI increasingly influences many aspects of society, it is important to promote the development of trustworthy and ethical AI systems. Governments can establish regulations and standards for AI development, testing, and deployment to ensure that they are safe, reliable, and transparent. In addition to regulation, governments can also invest in research to better understand the potential risks and benefits of AI systems, as well as encourage collaboration between various stakeholders to ensure that AI systems and LLMs are developed in a way that promotes ethical and societal values. However, government regulation must also strike a balance between promoting innovation and ensuring ethical considerations are upheld. Moreover, governance mechanisms that bring together various stakeholders must complement ethical frameworks to ensure responsible AI development. Furthermore, effective governance must include audits that enforce ethical principles and ensure compliance with regulatory frameworks. To sum up, the role of governments in regulating AI systems is vital for ensuring that these technologies are developed and deployed in an ethical and responsible manner. One of the critical aspects that governments must consider in regulating AI systems is to ensure that these technologies remain human-centered. To achieve this, governments must promote transparency and inclusive by encouraging public participation in policy discussions related to AI, developing ethical guidelines for AI development, and establishing mechanisms for holding organizations accountable for the potential negative impacts of their AI systems. In conclusion, the ethics of AI system development is a critical issue that requires

sustained attention and action from various stakeholders, including governments, private organizations, researchers, and the general public. Developing robust regulatory and legal frameworks for machine learning is essential to address the ethical and societal implications of AI technologies. These frameworks should be accompanied by investment in research, public engagement, and collaboration between various stakeholders to ensure responsible AI development. Only through a collective effort can we ensure that AI systems and LLMs are developed and deployed in a way that promotes human values, societal well-being, and sustainable development. Moreover, the regulation of AI systems is not a responsibility that governments can delegate to private organizations. Effective governance of AI systems requires a combination of ethical, legal, regulatory, and technical approaches that promote transparency and inclusive while balancing innovation with ethical considerations.

Limitations and Exclusions of the Current Study on AI Ethics

As with any research, there are limitations and exclusions to the current study on AI ethics. One limitation is that the study focuses on a few specific questions related to AI ethics, whereas there are many other important ethical and societal considerations surrounding AI development and deployment that should be explored. Another limitation is that the study may not fully account for the diverse perspectives and experiences of various stakeholders, including individuals from marginalized communities who may be disproportionately impacted by Ai systems. Furthermore, the study may also exclude ethical considerations that are specific to certain cultures or regions. Therefore, it is important for researchers and policymakers to continue exploring the ethical implications of AI systems from diverse perspectives and to ensure that all voices are heard in the conversation. Additionally, education and training are crucial components in regulating AI systems. As AI technology

continues to evolve and become more complex, it is important that individuals, organizations, and governments are equipped with the knowledge and skills necessary to understand and regulate such systems. Lastly, it is important to note that the current study may become outdated as technology continues to rapidly advance and new ethical considerations emerge. Thus, ongoing research and dialogue surrounding AI ethics are imperative to stay up-to-date with the latest developments in this field. In conclusion, developing AI systems and LLMs presents both opportunities and challenges for society. As AI technology becomes more advanced, it is imperative that ethical considerations are taken into account to ensure that these systems are developed and used responsibly. This includes the need for human oversight, guidance, and responsible design and operation for generative AI applications. Furthermore, it is important for companies and individuals involved in AI development to take full responsibility for the impact of their systems on society.

References

[1] R. An, J. Shen and Y. Xiao. "Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies". Dec. 2022.

[2] Sebastien Bubeck. “Sparks of AGI: early experiments with GPT-4” [YouTube

Video]. Published Apr 6, 2023. URL:

“https://www.youtube.com/watch?v=qbIk7-JPB2c”

[3] Tristan Harris and Aza Raskin. “The A.I. Dilemma” [YouTube Video]. Published

Apr 5, 2023. URL: “https://www.youtube.com/watch?v=xoVJKj8lcNQ”

[4] Many signatories. “Pause Giant AI Experiments: An Open Letter” [Online

Petition]. Published March 22, 2023. URL: “https://futureoflife.org/open-

letter/pause-giant-ai-experiments/”

[5] Toby Shevlane. “Sharing Powerful AI Models” [Research Brief]. Published

January 20, 2022. URL: “https://www.governance.ai/post/sharing-powerful-ai-

models”

[6] Open-Ai Gpt3