A DISCUSSION ON THE ALIGNMENT PROBLEM

Riley Harper
riley.harper@unc.edu
GitHub

Abstract: This paper explores current Artificial Intelligence (AI) research through the lens of ethics. In particular, the ethics of utilizing AI, particularly through the examination of real-world examples posited by supplementary research papers. Most notably, the paper bases its methods of exploration and discussion on the seminal paper "Sparks of Artificial General Intelligence: Early experiments with GPT-4" with additional data provided by the authors of "BBQ: A Hand-Built Bias Benchmark for Question Answering."

Keywords: Machine Learning, Artificial Intelligence, Deep Learning, Natural Language Processing, Convolutional Neural Networks.

Introduction

Artificial Intelligence (AI) is a field that many nowadays would consider to be novel, yet the underpinnings of its effects in many fields such as psychology, law, and medicine can have deep ethical concerns. In fact, although many would consider this field to be a new one, it has been studied back to the '60s when the man many would consider to be the father of Computer Science—English scientist Alan Turing—published a seminal paper titled "Computing Machinery and Intelligence." In this paper, he posits an activity titled the "Imitation Game." The game involves three players—A, B, and C—with player C being blinded from the other two. By communicating back and forth via written notes with the players they cannot see, can player C correctly determine which of players A and B is human and which is machine? If not, then the computer wins the game. The cleverness of this game is that it sidesteps the need to provide a formal definition for what it means to "think" or to be intelligent by instead asking the machine to act as similar as a human would to a level at which it is indistinguishable. Since humans are considered to be sentient and capable of thinking, does this mean that the machine is too since the two are, absent of sight, indistinguishable from one another?

A highly discussed concern with the rapid advancements made in the research of AI is that of the alignment problem. The alignment problem can be defined as the challenge of ensuring that the objectives and actions of artificial intelligence systems are aligned with human values and ethical principles. This problem becomes increasingly complex as AI systems become more autonomous and capable of making decisions without direct human oversight. The concern is that if an AI's goals are not properly aligned with human values, it could lead to unintended consequences, potentially causing harm. Brian Christian, in "The Alignment Problem: Machine Learning and Human Values," delves into the nuances of this issue, exploring the intersection of machine learning (ML) technologies and the ethical considerations they necessitate. Christian highlights that as ML models become more integrated into various aspects of daily life, from healthcare decision-making to criminal justice and beyond, the stakes of misalignment grow exponentially. One of the key points Christian makes is the necessity of embedding ethical considerations into the very fabric of AI development processes. This involves not only technical adjustments, such as the design of algorithms that can understand and prioritize human values, but also broader societal engagement to determine which values should guide AI development. The book argues for a participatory approach to AI ethics that includes diverse voices in the conversation about what it means for AI to be aligned with human values, and emphasizes the role of transparency and accountability in AI systems. By making AI systems more interpretable and their decision-making processes more transparent, stakeholders can better assess whether these systems are aligned with ethical principles and human values. This transparency is crucial for building trust between AI systems and the people they impact.

Analysis of Methods

AI Development Approaches

The research analysis commenced by reviewing the primary methodologies used in AI development, including machine learning (ML), deep learning (DL), and natural language processing (NLP). This involved a comprehensive literature review to understand the technical foundations of each approach and their specific applications in various sectors. The focus was on identifying how each method could potentially contribute to or exacerbate the alignment problem. Special attention was given to the ways in which ML, DL, and NLP technologies make decisions, learn from data, and interpret human language, which are all critical factors in understanding their alignment with human ethical standards.

Ethical Considerations in AI Design

The investigation then extended to the ethical frameworks currently integrated into the AI development process. This involved analyzing existing literature, policy documents, and AI ethics guidelines proposed by both academic and industrial entities. The aim was to ascertain the extent to which ethical considerations—such as fairness, accountability, transparency, and avoidance of bias—are accounted for in the design and deployment of AI systems. We also explored the mechanisms by which AI developers and companies seek to embed these ethical principles into their AI models and the tools and techniques employed to audit AI systems for ethical compliance. Of note, it was found that while many instances of bias and overfitting are present with many types of artificial intellegence these problems are not unique to AI and can be seen at similar and sometimes elevated rates to humans. It will be interesting to weigh

the benefits and drawbacks between the two in various settings such as those outlined in the introduction.

Case Studies

COMPAS Algorithm in Criminal Justice
The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm is a risk assessment tool used within the criminal justice system to predict the likelihood of a defendant reoffending. It has been widely used in various jurisdictions across the United States to inform bail, sentencing, and parole decisions. However, the COMPAS algorithm has been at the center of a significant controversy regarding its ethical alignment, particularly concerning bias and transparency.

Alignment Issues: The primary ethical alignment issue with the COMPAS algorithm involves accusations of racial bias. An investigation by ProPublica in 2016 found that the algorithm was more likely to falsely label black defendants as likely to reoffend than white defendants, raising concerns about fairness and equality in the criminal justice process. Additionally, the proprietary nature of the COMPAS algorithm means its decision-making process is not transparent, making it difficult for defendants and the public to understand or challenge its assessments.

Mitigation Methods: In response to these concerns, some jurisdictions have called for increased transparency and accountability in the use of such algorithms. Proposals include independent audits of risk assessment tools for bias, the development of more transparent AI systems in criminal justice, and the inclusion of fairness constraints in algorithm design to reduce bias.

Outcomes: The debate over the COMPAS algorithm has sparked broader discussions about the use of AI in criminal justice. While some reforms and audits have been initiated, the issue of balancing predictive accuracy with fairness and transparency remains unresolved. The case highlights the complex challenges in aligning AI technologies with ethical principles and human values in sensitive applications.

Tesla's Self-Driving Cars and Regulations
Tesla's self-driving car technology represents a pioneering effort in the field of autonomous vehicles. Tesla's Autopilot and Full Self-Driving (FSD) systems aim to enable vehicles to navigate and operate with minimal human intervention. However, regulatory challenges and safety concerns have led to scrutiny and restrictions in some jurisdictions, such as Texas.

Alignment Issues: The ethical alignment concerns with Tesla's self-driving cars primarily revolve around safety and accountability. Incidents involving Tesla vehicles operating on Autopilot or FSD leading to accidents have raised questions about the technology's current readiness and reliability. Additionally, there are concerns about the clarity of communication to consumers regarding the capabilities and limitations of the technology, potentially leading to overreliance and misuse.

Mitigation Methods: Tesla has continuously worked on improving the safety and reliability of its self-driving technology through software updates and enhanced sensor capabilities. The company also emphasizes driver education and the need for driver vigilance even when using these systems. Regulatory bodies and Tesla have engaged in dialogues to address safety standards and testing protocols for autonomous vehicles.

Outcomes: The situation with Tesla in Texas and other jurisdictions underscores the ongoing challenges in the regulation and public acceptance of self-driving car technology. It highlights the need for clear safety standards, transparent communication of the technology's capabilities, and robust testing and oversight mechanisms. As regulatory frameworks evolve to accommodate advancements in autonomous vehicle technology, ensuring the alignment of these systems with public safety and ethical standards remains a critical focus.

Evaluation of Effectiveness

Finally, the effectiveness of the various methods used to ensure AI alignment with human values were compared to one another. This evaluation was based on criteria such as the ability to prevent or mitigate ethical issues, the adaptability of methods to different AI systems and contexts, and the sustainability of alignment efforts over time. We also considered the limitations and challenges encountered in implementing these methods, drawing on evidence from the literature review and case study analyses. This evaluation will provide insights into the current state of AI alignment, highlighting areas of progress and identifying gaps where further research and development are needed.

Analysis of Normative Consideration

Ethical Frameworks

The ethical implications of AI technologies like GPT-4 are ubiquitous, touching on various ethical frameworks including utilitarianism, deontological ethics, and virtue ethics. Utilitarianism evaluates AI outcomes based on the greatest good for the greatest number, considering the widespread benefits of AI in enhancing productivity, creativity, and solving complex problems. Deontological ethics focus on the duties and rights associated with AI, emphasizing the importance of developing and using AI in ways that respect human dignity and rights, such as privacy and autonomy, regardless of whether it results in the most optimal outcome in terms of productivity. Virtue ethics highlight the character and virtues both of AI systems and their creators, promoting the development of AI that embodies virtues like honesty, justice, and empathy. The development of GPT-4, as reported in [Bubeck et al., 2023], will need to involve considerations of minimizing harm and bias, promoting fairness, and ensuring that the AI acts in ways that are beneficial to humanity since there have been cases of adversarial users utilizing the AI in a malicious manner.

Stakeholder Perspectives

The perspectives of various stakeholders including developers, users, ethicists, and policymakers highlight diverse views on what it means for AI to be aligned with human values. Developers focus on technical challenges and the potential of AI to solve complex problems. Users are concerned with the practical benefits of AI, its usability, and ethical implications such as privacy and job displacement. Ethicists raise concerns about long-term impacts, bias, and the moral responsibilities of AI developers. Policymakers grapple with regulating AI to protect public interests while fostering innovation. There's a consensus on the need for AI like GPT-4 to be developed and used responsibly, with ongoing dialogue among stakeholders to navigate ethical dilemmas and societal impacts.

Policy and Regulation

Existing regulations, such as the GDPR in Europe, provide frameworks for data protection and privacy, while proposals for new guidelines aim to address emerging issues related to autonomy, accountability, and transparency of AI systems. The development of GPT-4 considers these regulatory landscapes, with efforts to embed ethical considerations into AI design and deployment. Future regulations may need to address more complex issues such as AI personhood, intellectual property rights generated by AI, and the use of AI in sensitive decision-making processes.

Future Implications

The long-term implications of normative considerations on AI development are significant. Successful alignment of AI with human values could lead to AI systems that enhance human well-being, foster economic growth, and solve intractable problems. However, failure to address ethical and societal concerns may result in mistrust, societal harm, and exacerbation of inequalities through a self fulfilling algorithm. The exploration of GPT-4's capabilities and its impact on society underscores the importance of ethical AI development that accounts for a wide range of future scenarios, both positive and negative.

Conclusion

The exploration of GPT-4 has unveiled a monumental leap in the journey towards Artificial General Intelligence (AGI), offering a glimpse into the potential that lies within large language models (LLMs) when they are endowed with extensive training across a vast spectrum of data and tasks. Our analysis has demonstrated that GPT-4, even in its early stages, represents a significant advancement in AI's capability to perform a wide array of tasks that were previously thought to be exclusive to human intelligence. This includes, but is not limited to, complex problem-solving across various domains such as mathematics, coding, natural language understanding, and even creative tasks such as music composition and graphic design.

One of the most striking revelations from our experiments is GPT-4's ability to understand and generate content that resonates with human cognitive functions, suggesting that it is inching closer to achieving a form of general intelligence. However, it is imperative to note that GPT-4, while showcasing traits akin to AGI, is not without its limitations. The model still faces challenges in areas such as planning, common sense reasoning, and maintaining coherence over extended interactions. These limitations underscore the necessity for ongoing research and development to

overcome these hurdles and further refine the model's capabilities.

The societal implications of GPT-4 and subsequent models are profound, encompassing both positive and negative dimensions. On one hand, GPT-4 has the potential to revolutionize various sectors by enhancing efficiency, fostering creativity, and aiding in complex decision-making processes. On the other hand, the advent of such powerful technology raises ethical concerns, particularly related to privacy, security, misinformation, and the displacement of jobs. It is crucial for the research community, policymakers, and society at large to engage in a collaborative dialogue without siloed stakeholders to navigate these challenges and ensure that the development and deployment of AI technologies like GPT-4 are aligned with ethical standards and human values.

Appendix

References

[Bubeck et al., 2023] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.

[Christian, 2020] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, New York.

[Fedus et al., 2022] Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. _eprint: 2101.03961.

[Parrish et al., 2022] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. (2022). Bbq: A hand-built bias benchmark for question answering.

[Srivastava et al., 2022] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A., Abid, A., Fisch, A., Brown, A., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A., Safaya, A., Tazarv, A., and Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

[Turing, 1950] Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.