Discussion on Artificial Intelligence Safety and Ethical Issues

Xinyu Chen[1], *Tianfang* Hui[2*],*Yanlin* Li[3,] and *Haoyuan* Yang[4]

[1]Shanghai Shangde Experimental School, 201100 Shanghai, China
[2]Guangzhou No.7 Middle School, 510000 Guangzhou, China
[3]Qunxing Foreign Language School, 322000 Yiwu, China
[4]Luoyang No.1 Senior High School, 47100 Luoyang, China

Abstract. As artificial intelligence (AI) is increasingly integrated into society, people are relying on it more and more, and higher requirements are put forward for the safety and ethical standards of AI. This article explores the development of artificial intelligence technology and its potential safety and ethical challenges in various fields. In terms of security, the risk of adversarial attacks is analyzed in depth, and the robustness of the model is enhanced through adversarial training and data enhancement techniques. In addition, it is recommended to adopt measures such as data encryption and differential privacy to address data privacy and security issues. Regarding ethical considerations, this paper identifies the origins of algorithmic bias and argues for mitigating it through rigorous testing, validation, and regulatory frameworks. It also highlights the importance of increasing the transparency and explainability of AI to enhance public trust. Finally, the paper emphasizes the importance of defining accountability for AI behavior and suggests establishing laws and regulations that effectively govern AI applications. In conclusion, the study argues that the development of AI should emphasize safety and ethical considerations. Through the combination of technical intervention, legal supervision and social responsibility, the sustainable development of artificial intelligence is effectively promoted.

1 Introduction

In recent years, with the development of technology, artificial intelligence (AI) has been gradually improving and has been involved in various fields, such as autonomous driving, the financial industry, and healthcare. AI technology not only improves production efficiency but also drives society towards digital transformation. For example, in the development of autonomous driving vehicles, AI technology is developing new traffic patterns to make transportation more convenient and efficient. In the financial industry, AI technology is used for risk prediction and market analysis, while in the medical field, AI technology uses image recognition and data analysis to help doctors make more precise diagnoses and treatments. However, as AI technology is widely applied, safety and ethical issues have become increasingly prominent.

[*] Corresponding author: bennet99@nwfsc.edu

As a strategic technology leading the future, AI is the core driving force for a new round of scientific and technological revolution and industrial change and an important engine for the development of new quality productivity. For example, in the medical field, AI can propose high-quality solutions in drug development and reduce the time to drug production. AI can make corresponding suggestions for doctors to improve the quality and efficiency of medical services. In the financial sector, AI can provide services 24 hours a day, provide customers with more personalized services by calculating large amounts of data, greatly improve the efficiency of banks and companies, and help banks and companies automate processes such as processing customer needs and loans. In the future, AI will be more widely used, the application of large models will be more in-depth in the industry, promote industrial upgrading and new technological revolution, more industries will be completely changed by AI, and profoundly change people's social environment and living habits. The rapid development also brings some risks. However, with the proper application of people and reasonable policies and supervision, the risk will be reduced.

Therefore, this article will explore the challenges and solutions to the safety and ethical issues of AI technology in today's society.

2 AI safety

Adversarial attacks interfere with the judgment of AI models through special data, seriously affecting their security and reliability. The robustness of a model refers to its ability to maintain performance in the face of attacks and noise, and improvements include adversarial training and data enhancement. At the same time, data privacy and security go hand in hand, with the former focusing on the protection of personal information and the latter ensuring that data is protected from unauthorized access and tampering. Together, these three aspects constitute the core elements of AI security, ensuring the security and reliability of AI systems in technology, decision-making, and data management.

2.1 Adversarial attack

An adversarial attack is a type of attack that causes an AI to make bad judgments and behaviors by feeding it special data. This attack will seriously affect the model to produce the wrong output, affecting the security and reliability of the system. For example, adversarial attacks can be used for signal jamming and spoofing. The anti-drone device interferes with the drone's communications by transmitting signals on the same frequency as the controller, forcing loss of connection or an emergency landing. This adversarial attack will cause the AI to make wrong judgments, affecting its normal operation.

2.2 Model robustness

In this section, the definition of model robustness and several methods to improve robustness are briefly explained. Robustness refers to the ability of the model to maintain its performance and predictive power in the face of counterattack and noise interference. Model robustness can make the right choice when encountering noise interference or other disturbances, and can still operate normally without being disturbed by the outside world in a complex environment. On this premise is an excellent model robustness. The methods to improve the robustness of the model include adversarial training, data enhancement, and input preprocessing. These methods work together to significantly improve the ability of the model to deal with counterdisturbance. Next, teammates will go into details of these methods.

Adversarial training is a common way to improve robustness. When adversarial samples are introduced into the training process, the model is trained with the original sample, making the right choice between the two, and performing well in the face of both "normal" and "abnormal". Enhancing Data is another way to solve this problem. Different transformations on the original sample, for example, can change the color of the sample, shape, data changes, etc. In this way, the model can adapt to various input changes when facing different samples, and improve the resistance to data disturbance. Before the model is processed on the data, the noise or adversarial attack that may exist in the processing process is removed. Common ways include noise reduction, edge detection, etc. This enables the model to show good coping measures in the face of noise disturbance or other adversarial attacks.

2.3 Data privacy and security

Data privacy is the idea that information directly or indirectly contained in the data, involving individuals or organizations, is not suitable for public disclosure, and needs to be protected in the process of data collection, data storage, data investigation and analysis, and data release. The ability to protect data privacy, usually using data anonymization, data perturbation, data encryption, differential privacy, and other technologies.

This means users have management rights and control over their personal data, and users can accept or reject the collection of data. Under the premise, personal data can only be used in the correct and legal way, data privacy is protected by law, is also restricted by law, and may not be used arbitrarily. Data security refers to a set of measures and techniques that protect data from unauthorized access, use, destruction, disclosure, or tampering. The purpose of data security is to protect the confidentiality, integrity, and availability of data. Data security not only needs to ensure the security of the whole life cycle of data from the perspective of science and technology, combined with a multidisciplinary approach, but also needs to start from the perspective of government governance, cooperate with multiple departments, and provide legal basis and policy guarantee for data security by improving laws and regulations and establishing a regulatory system.

While data privacy focuses on how data is used, ensuring that personal data is protected within a legal, ethical framework, and emphasizing users' right to information and consent, data security focuses on how to protect data from unauthorized access and prevent data leakage, tampering or loss, emphasizing the defense of technical means. The two complement each other, data privacy can only be achieved by protecting data security, and ensuring data privacy compliance can also help reduce data security risks.

3 AI Ethics issues

3.1 Algorithmic bias and discrimination

Algorithmic bias refers to the unjust or discriminatory behavior of AI systems when making decisions. Bias and discrimination may cause distress and harm to minorities. The formation of bias and discrimination has multiple factors, including data, models, development processes, and the social environment. The sources of algorithmic bias are mainly influenced by several factors such as data bias, human intervention and developer bias, and social structural bias. Human intervention and developer bias happens because developers may unconsciously or consciously inject their biases into the development process of an AI system, leading to AI's thinking and decision-making being biased during operation. For example: iTutor Group Inc., an English tutoring company, faced legal consequences for using an AI application software to automatically reject older job applicants. The system was programmed to exclude female job applicants over 55 years old and male job applicants over 60 years old, regardless of their self-sufficiency or experience.

This case of age discrimination driven by AI resulted in the Equal Employment Opportunity Commission paying $356,000 in settlement. The example demonstrates how developers amplified age discrimination during the coding process of AI. This disadvantaged older job applicants in the competition and subjected them to unfair treatment. Human intervention has a significant impact on algorithmic discrimination and can cause harm to users. This case shows that during the design and development of AI systems, developers' conscious or unconscious biases may be amplified and affect AI's judgment results, especially in applications involving credit scoring and healthcare, where algorithmic bias can have significant negative impacts on specific groups of people. Therefore, the an urgent need for AI's fairness and transparency [1]. Data bias: when developers train AI, there is bias in the uploaded data, and AI will inherit the bias in the data during the training process. For Instance, a Brookings Institution study highlights how AI-based financial services perpetuate socioeconomic inequalities in credit scores. The study found that existing credit scores, such as FICO scores, were strongly correlated with race, with white home buyers having an average credit score 57 points higher than black applicants and 33 points higher than Hispanic applicants. These differences lead to significant differences in loan approvals and interest rates. More than 1 in 5 blacks had FICO scores below 620, compared to 1 in 19 whites. Even if race is not explicitly included as a factor, AI systems tend to look for alternative factors due to existing income and wealth gaps between racial groups. While new AI models that use alternative data such as cash flow analysis may reduce some of the bias, they are still somewhat correlated with income and wealth. Addressing socioeconomic biases in AI credit scoring systems is a complex challenge, and efforts to improve accuracy sometimes inadvertently amplify existing gaps.

This suggests that even when AI systems do not directly include race as a decision-making variable, they may still inadvertently widen racial inequality due to implicit biases in the data. The root of the problem is that the data itself is unequal. When data is used directly to train AI, these biases can be inherited by the AI, thereby affecting specific racial or social groups [2].

Society and structure: AI is run within the social structure, he will to a certain extent. Reflecting the social structure may be affected by socio-political factors, which will also be reflected in the AI algorithm and the answer to the question.

Example: Most of the AI models in China cannot publish sensitive topics related to politics, and the AI system does not support or answer sensitive topics related to Chinese politics, such as national division and socialist policies.

As a product of human society, AI systems inevitably inherit social values and ideologies. AI's "silence" on certain sensitive topics is actually a response to the norms of social behavior, thereby reinforcing existing constraints and limitations at the technical level. Therefore, the design and training of AI may also be influenced by cultural context, resulting in AI showing a bias or avoidance when answering questions.

3.2 Transparency and explainability

Transparency and explainability are important in the operation process of AI systems. Transparency means that the operation process, decision-making mechanism, data flow, and model behavior can be understood and traced by people. People can see clearly and transparently how AI makes decisions; Interpretability means that people can logically figure out how AI solves this problem and answers this question through AI's operational principles, and how to produce explanations for AI's own answers. Organizations are increasingly using complex black-box machine learning models for high-stakes decisions. A popular approach to solving the opacity problem of black-box machine learning models is to use a post-interpretability approach. These methods approximate the logic of the underlying machine-learning models and aim to explain their inner workings so that human examiners can understand them [3].

3.3 Accountability and responsibility

AI's machine behavior is different from the traditional sense of machine behavior,
traditional Machines do not have the ability to think and make decisions, just different parts assembled by the motor to run

Unlike machines that have no autonomous consciousness, AI can think and make decisions, which is the ability of its machine to act autonomously without human interference.

Therefore, there is also a moral responsibility for the behavior of AI machines. Therefore, in order to avoid the ethical risks of educational AI, the United States has formulated relevant standards for educational AI, strengthened the supervision of educational AI algorithms, introduced data protection guidelines and bills, and standardized the application of intelligent technology in education. China should learn from the governance strategies of the ethical issues of educational AI in the United States, formulate and improve the corresponding policies of educational AI, establish an all-staff linkage supervision mechanism throughout the whole life cycle, develop an educational AI ethical framework, and cultivate and improve the AI literacy of teachers and students [4].

4 Current solutions and progress

In terms of Algorithmic bias and discrimination, human intervention discrimination, the use of open source toolkits in the development and testing process to test the unit and reduce AI bias, as well as the release of relevant laws, Microsoft issued a dialogue when AI robots treat humans fairly guidelines, Relevant developers should develop AI according to guidelines and laws [5]. Data bias can be fixed by detecting whether there is bias in the data and checking the fairness of the data, before uploading data to train AI.

Transparency and interpretability: explainable reinforcement learning (XRL), or explainable reinforcement learning, is an AI. XAI subproblem is used to enhance human understanding of the model and optimize the performance of the model, so as to solve the above four types of problems caused by the lack of solvability. There are commonalities between XRL and XAI, while XRL has its own uniqueness [6,7].

Third, the attribution of responsibility for AI: The law of the state or related organizations issuing the attribution of ethical responsibility for AI. Clear attribution of responsibility for AI by relevant groups can avoid unnecessary disputes in times of conflict [8,9]. On March 3, 2021, the European Commission adopted the AI Act, which aims to introduce a common regulatory and legal framework for AI and clarify the attribution of responsibility for AI [10].

5 Conclusion

This paper aims to discuss potential safety issues and ethical challenges to promote the sustainable and responsible development of AI. In the foreseeable future, society's acceptance of AI will further increase, and new problems will arise. However, the balance between technological progress and social responsibility should always be the basic principle for measuring issues.

At present, adversarial attacks are considered to be the main threat, and the robustness of models has become an important criterion for evaluating AI security performance. Data privacy has become a consensus of the present, people also use differential privacy and other means to prevent leaks.

In terms of ethical issues, AI's inheritance of human prejudice and discrimination has been a concerned for all sectors of society. This bias, whether intentional or not, often brings great losses to society. The lack of transparency and explainability of AI will also affect people's attitudes towards AI. Through manual intervention, targeted training, and other means, such problems can usually be solved. The improvement of relevant laws and regulations is also the obligation of each country.

AI's outstanding performance in various fields shows its important impact on the development of the world but also shows that in-depth research on AI safety and ethics is urgent. The responses also need to evolve to adapt to more complex and volatile situations.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

Reference

1. G. Denison, 8 shocking AI bias. Prolific. August 24 (2023) https://www.prolific.com/resources/shocking-ai-bias
2. D. Vale, A. El Sharif, M. Ali, Explanatory Artificial Intelligence (XAI) Post hoc Interpretability Methods: Risks and Limitations in Non Discrimination Law (2024)
3. T. Xiaoling, Z. Chuyue, Examining the Ethical Issues of Artificial Intelligence in American Education from the Perspective of Technological Ethics: Realistic Representation and Avoidance Strategies. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition) (2024)
4. J. Silberg, J. Manyika, Notes on the Frontiers of Artificial Intelligence: Addressing Bias in Artificial Intelligence (and humans). McKinsey Global Institute. **1**(6), 1-31 (2019)
5. X. Liu, S. Liu, Y. Zhuang, Y. Gao, Explorations on the explainability basis of reinforcement learning and a review of methods. Journal of Software. **34**(5), 2300-2316 (2019)
6. K.-C. Chen, Artificial Intelligence in Wireless Robotics. Routledge, 9788770221184 (2023)
7. M. Shah, A. Doshi, Transforming Agricultural Technology by Artificial Intelligence. Routledge, 9781032072425 (2023)
8. T.H. Davenport, N. Mittal, All-In On AI. Harvard Business Review Press. 9781647824693 (2024)
9. M. Coeckelbergh, AI Ethics. MIT Press. 9780262538190 (2020)
10. K.C. Santosh, L. Gaur (Eds.), Artificial Intelligence and Machine Learning in Healthcare. Springer. 9789811667671 (2022)
11. G. Viggiano, D. Puthal (Eds.), Convergence: Artificial Intelligence and Quantum Computing. Social, Economic, and Policy Impacts. 1394174101 (2023)
12. C. Guger, B.Z. Allison, G. Edlinger (Eds.), Clinical Neurotechnology Meets Artificial Intelligence: Modern Advances and Future Challenges. Springer. 3030645894 (2021)

13. P. Raj, J.M. Chatterjee (Eds.), Artificial Intelligence in Cyber Security: Impact and Implications. Springer. 9783030880392 (2022)
14. M. Shah, A. Doshi (Eds.), Transforming Agricultural Technology by Artificial Intelligence. Routledge. 9781032072425 (2023)