The ethics and governance of Artificial Intelligence large language models

Zhihong Liu
School of Artificial Intelligence, Hebei Oriental College, Langfang 065900, China;
Publishing House of Electronics Industry, Beijing 100036, China;
Bress Business School, France Brest, 29200

Abstract: After the construction of large language artificial intelligence (AI) models and their provision as a social service, potential risks quickly become real risks that affect our society. In the past, we have always responded to ethical and safety issues with AI in a reactive manner, rather than proactively. As a result, we encounter some of the social problems brought about by large language AI models. This paper aims to focus on the existing problems caused by large language AI models and promote the implementation of ethical and safety governance.

Keywords: artificial intelligence, large language models, ethical security, governance.

I. Introduction

Artificial intelligence (AI) large language models, such as ChatGPT and Wenxin Yiyan, have had a profound impact on the world. These models can understand and generate human language, perform complex reasoning and problem-solving, and even create works of art. However, with their enhanced capabilities, they have also raised a series of ethical and governance issues. These issues involve data privacy, algorithmic bias, decision transparency, technical regulation, and more.

II. Ethical Issues of AI large language models

According to Darkreading's report on April 26, 2023, a security researcher lured ChatGPT to build a sophisticated data theft malware tool that bypasses the anti-malware protection of chatbots using signature-based and behavior-based detection tools. The researcher admitted that he had no experience in developing malware code and had only guided ChatGPT through multiple simple prompts to ultimately produce a malware tool capable of silently searching for specific documents in a system, decomposing them and inserting them into image files, and sending them to Google Drive.

Forcepoint's solution architect and malware author Aaron Mulgrew said that it only took about four hours to compile the malware from the initial prompt into ChatGPT, with zero detection on Virus Total.

In addition, Samsung has reportedly told employees in internal memos to stop using ChatGPT and Bard, among other generative AI tools, due to concerns about security risks. According to reports earlier, Samsung had experienced three incidents involving ChatGPT within less than 20 days of introducing the chatbot, two of which were related to semiconductor equipment and one was related to meeting content. The internal memos stated that Samsung feared that data stored on external servers transmitted to artificial intelligence platforms such as Google Bard and Microsoft Bing would be difficult to retrieve and delete, potentially leaking to other users. The internal memos also showed that Samsung conducted a survey on the use of AI tools internally last month, with 65% of respondents believing that such services posed security risks.

Samsung is not the first or the last company to take measures such as disabling or suspending ChatGPT due to security concerns. The finance industry was among the first to reject ChatGPT. Shortly after its popularity surged, in February 2023, JP Morgan Chase, Bank of America, Citigroup, and Wells Fargo announced that they would ban or restrict the use of ChatGPT. This was not only because ChatGPT is a third-party application software and not an internal system for the company, but also because it poses potential regulatory risks related to sharing sensitive financial data with AI. On April 30th, 2023, the G7 Digital and Technology Ministers agreed on "risk-based" regulation of AI. In a joint statement, the ministers stated that such regulation should also "maintain an open and favorable environment" for the development of AI technology. Recently, the European Union proposed new legislation that requires developers of AI tools such as ChatGPT to disclose copyright materials used in building AI systems. This rule will empower publishers and content creators to seek profit sharing.

Through these security issues caused by large language models, we understand that like any powerful tool, AI large language models can be used for good or bad purposes and may be accidentally used for the latter. For example, privacy data may be accidentally leaked through machine learning (ML) algorithms or data engineering pipelines. Therefore, to adapt to the times and become a comprehensive AI engineer, it is necessary to understand the ethical requirements of working in the field of AI-related fields. This usually includes privacy and bias.

The use of machine learning algorithms is becoming increasingly prevalent in people's daily lives, and an important issue to consider is the ethical implications of using these powerful tools.

Data Privacy: AI large language models typically require a large amount of data for training. This data may contain private information about users, and how to use this data while protecting user privacy is a significant ethical concern.

Algorithmic Bias: The training data for AI large language models may contain human biases, which can lead to biases in the model's decision-making. For example, if an AI recruitment system's training data mainly comes from male candidates, the system may produce unfair bias against female candidates.

Transparency of decision-making: The decision-making process of AI large language models is usually a black box operation. It is

difficult for users to understand the basis of their decision-making. This may lead to a lack of trust in the decision-making of AI systems, and it may also make it difficult to hold AI system errors accountable.

III. Governance Issues of AI large language models

In view of the risks of large language artificial intelligence models mentioned above, we should focus on real-time process deep learning, leverage the computing power of clouds to complete analysis more quickly, and establish a global backbone network with ultra-low latency to build a more professional AI/ML implementation for real-time threat defense. On the basis of continuous, complete, and accurate security data, we should use large language elastic scalable computing power to build a more accurate, user-friendly, and inclusive next-generation intelligent security. In this regard, we should start from the following three aspects.

Good data can create great artificial intelligence. By analyzing security events and network effects, we can build a huge high-quality security data lake.

AI+ML achieves precise detection of zero-day attacks. The constantly evolving security threat situation fundamentally requires AI engineers to continuously explore and discover new solutions.

Online real-time prevention of attacks. We can explore online real-time security detection methods, and even more precise identification and protection against network attacks.

In the era of AI large language models, security integration management is the only way out. We also need to focus on the following aspects.

Technological supervision: As the capabilities of AI large language models increase, so do their potential risks. How to effectively regulate these models to prevent abuse or negative consequences is a major challenge.

Policy making: The development and application of AI large language models require corresponding policy support. How to formulate reasonable policies that can protect users' interests while promoting the development of AI large language models is a problem that needs to be solved.

Community participation: The development and application of AI large language models involve multiple stakeholders, including developers, users, governments, etc. How to involve these stakeholders in the governance of AI large language models is a question that needs to be explored.

IV. Solution

1.Establish a data privacy protection mechanism: You can protect user data and use it for training through technical means, such as differential privacy.

Data privacy is an important part of the ethics of data science. When working in the field of artificial intelligence, although what we see are just some numbers, it's important to remember that these numbers represent people. A strong example is the Titanic dataset, which is often used as an ML learning material. The dataset contains data on passengers on the Titanic, and is usually used for classification exercises, where the goal is to classify whether passengers survived or not on the Titanic. When processing data, AI engineers can easily get lost in the numbers and details of implementing ML algorithms, but they should remember that each data point is a person just like you and me.

Data privacy can be compromised in various ways, such as data breaches (e.g., hacked or stolen data), anonymizing people by combining data from multiple sources, extracting information from ML algorithms, data mining, and publishing results of small datasets (even aggregates).

Privacy breaches and data leaks can expose a person's sensitive data, such as their medical condition they may have, their web search history, etc. Of course, data breaches caused by hacker attacks are a clear way to compromise data privacy. Therefore, people should properly protect their data, for instance, by securing it on a secure system with appropriate authorization (e.g., multi-factor authentication with strong passwords). Similarly, data can leak out from within the organization. Proper data governance and employee management will help prevent internal data theft.

A more subtle way to invade privacy is by combining public data with so-called de-identified data. For example, identifying someone by combining data sets that combine zip codes, age, gender, and other demographic statistics. In this case, we can use k-anonymity to protect the privacy of a data subject with at least k records that cannot be distinguished from each other. A more advanced approach to this is l-diversity, which aims to ensure that sensitive classes are diverse by at least l factors. L can simply be the count of unique class labels. The distribution of sensitive classes can be inferred from the data, and the probability that someone may have a sensitive class label can be inferred.

An improved version of l-diversity is t-closeness. This method ensures that each equivalence class of sensitive classes is distributed within t factors of the overall table distribution. There are several ways to measure the distance between distributions, but the simplest is variational distance. In this case, the difference in unique values between two distributions is taken. An advanced method is Kullback-Liebler (KL) distance, which uses entropy and cross-entropy computations.

2.Establish a fairness check mechanism: Anti-bias elements can be introduced into training data or the fairness of models can be reviewed periodically.

There are several methods to eliminate these biases in machine learning (ML) algorithms:

- Collect more data, especially for balanced datasets

● Create synthetic data using GAN (Generative Adversarial Networks), SMOTE, or ADASYN

● Use oversampling or undersampling techniques such as SMOTE and ADASYN

Data collection processes are often expensive, so collecting more data may be difficult to achieve. However, one can balance classification problems during data collection or increase the amount of data for underrepresented groups. A simpler method to balance data groups is to use SMOTE or sampling techniques for simple classification problems. Data can also be undersampled, but generally, it is best to use as much data as possible. The method to create more data by synthesizing is using GAN. These neural networks generate fake data that follows the distribution of the data they were trained on. If there is some data on underrepresented groups, synthetically generating more data with GAN can be done. However, it will follow the underlying distribution of the existing data, so it should ensure diversity within the original dataset.

3.Improve decision transparency: users can understand the basis of AI system decisions through interpretable AI technology.

Establish technical supervision mechanisms: government and industry organizations can develop relevant technical standards and norms to regulate the development and application of AI large language models. Establish a diversified stakeholder participation mechanism: multiple stakeholders can be involved in the governance of AI large language models through open forums and public consultations. The evaluation of multiple indicators of machine learning algorithms should be carefully considered, and which indicator is the best to use. In addition, the communication of the final results with end-users should be clear and accurate. For example, facial recognition systems can output matching confidence from classification algorithms, and perhaps some indicators such as accuracy and recall (or false positive rate) can also be good references. The interpretable results of machine learning can also be added to dashboards, which can show which parts of the input data are used to derive ML results.

V. Conclusion

The ethical and governance issues of AI large language models are a complex and important topic. Only by deeply understanding these problems can we find effective solutions, and I hope this article can provide some useful insights for research and discussion on this topic.

References

[1] Zero-experience - Four hours! Researchers Use ChatGPT to Build a VirusTotal Zero-Detection Steganography Malware□EB/OL□.http://www.hackdig.com/04/hack-958424.htm, 2023-04-06

[2] https://www.darkreading.com/attacks-breaches/researcher-tricks-chatgpt-undetectable-steganography-malware

[3] https://www.forcepoint.com/blog/x-labs/zero-day-exfiltration-using-chatgpt-prompts

[4] Nathan George. Practical Data Science with Python: Learn tools and techniques from hands-on examples to extract insights from data[M]. Packt Publishing Private Limited,2021

[5] Emergency! Trillion giant moves: Forbidden! [EB/OL]. https://new.qq.com/rain/a/20230502A05MGI00, 2023-05-02

[6] https://www.cs.purdue.edu/homes/ninghui/papers/ t_closeness_icde07.pdf