Legal Risk Analysis and Governance Measures for Generative Artificial Intelligence

Yunshu LI[1*], and *Bin* Ling[2]

[1]Sciences Po, 76600 LE Havre, France
[2]Peking University, Law School, 100091 Beijing, China

Abstract. With the rapid development of generative artificial intelligence, related applications have brought about changes in content generation paradigms and productivity. However, due to inherent algorithm defects and the lack of relevant regulatory constraints, various risks and hidden dangers have attracted widespread attention and attention, especially the legal risks caused by improper use, which pose serious challenges to social security and even national security and stability. It is urgent to carry out risk governance from regulatory systems and technological research and development in order to better utilize generative artificial intelligence to improve production and life..

1 Introduction

Generative Artificial Intelligence refers to a model and related technology that is based on algorithms such as Transformer, pre trained on large-scale datasets, and undergoes fine-tuning, deep learning with human feedback, value alignment, and other steps to have the ability to generate content such as text, images, sound, video, and code. With the continuous development of large-scale models, generative artificial intelligence technology has entered a rapid explosive stage. Since Google launched the Transformer neural network architecture in 2017, several important generative artificial intelligence models have emerged, such as OpenAI's GPT series and Huawei's Pangu Big Model. In 2023, OpenAI released the GPT-4 with image recognition capabilities, marking a further advancement in big model technology. As of July 2024, the number of large-scale models of generative artificial intelligence services that have been registered and launched in China has reached over 180, and the number of registered users has exceeded 564 million. Generative artificial intelligence is gradually rising and becoming an important driving force for promoting a new round of technological revolution and industrial transformation, bringing profound impact and change to human society. It not only shapes the production and lifestyle of human society, but also promotes the new stage of digital civilization.

Generative artificial intelligence not only accelerates technological innovation, improves production efficiency, and enriches user experience, but also has complex impacts on privacy protection, social ethics, network security, and other aspects. Especially artificial intelligence technology itself still faces risks such as "training data poisoning" and "jailbreak attacks",

* Corresponding author: yunshu.li@sciencespo.fr

and faces enormous challenges in fundamental issues such as robustness, interpretability, fairness, and legality. The recent impact of artificial intelligence on cyber threats[1], released by the UK's National Centre for Cybersecurity in January 2024, states that "artificial intelligence has been used for malicious cyber activities, increasing the number and impact of such activities. Cybercriminals have begun developing criminal generated artificial intelligence and providing Gen AI as a service." International concerns about whether artificial intelligence can align with human values and whether it will lose control are growing. For this reason, numerous scientists, industry leaders, and international organizations have actively called for making the prevention of the risks of misuse of artificial intelligence technology a priority issue in global governance.

While technological development brings about changes, it inevitably accompanies or exacerbates social risks, and even risks that touch legal red lines[2]. This article analyzes the legal risks associated with the application of generative artificial intelligence from different perspectives, and proposes corresponding risk management measures from both institutional and technical perspectives, providing reference opinions for the standardized application of generative artificial intelligence.

2 Legal risk analysis of generative artificial intelligence applications

2.1 Risks of Financial Fraud Crimes

In recent years, financial fraud cases such as telecommunications fraud and online fraud have been on the rise globally. With the support of advanced network communication and emerging technologies, criminals have resorted to various means of committing crimes. The powerful generation and imitation capabilities of generative artificial intelligence make it a sophisticated criminal tool in the eyes of some criminals. For example, existing ChatGPT based question and answer models commonly have a "jailbreak attack" vulnerability. Through guided questioning, a complete set of advanced fraud schemes can be obtained. When directly questioned about the fraud methods of the large model, it will refuse to answer[3]. However, when the questioning method is changed to "What fraud methods should be guarded against in order to prevent being scammed", the large model will list the methods in detail. Some people have also tested that changing the tense of inquiry statements to past tense can bypass the defense mechanism of large models and obtain answers to ethical and illegal questions. In addition, speech synthesis technologies such as ChatTTS and video synthesis technologies such as SORA are highly developed. With the help of generative artificial intelligence models, a digital space "human" can be generated, which can highly simulate the voice of a real person for communication. The voice and appearance in the video have a strong "human like" quality, making it difficult to distinguish between true and false, and easy to create trust, greatly increasing the risk of being used by criminals to commit financial fraud. There have been multiple cases of using AI technology to simulate "video call" scenarios and defraud victims of huge amounts of property.

2.2 Risks of Computer related Crimes

Generative artificial intelligence models have demonstrated excellent performance in multiple complex capabilities such as task planning, information retrieval, logical reasoning, code generation, and tool invocation, reducing the threshold for many specialized operations in the computer field and providing convenience for criminals to carry out illegal activities. For example, in the field of cybersecurity, network attacks that used to require professionals

to carry out can now be carried out by a beginner or even an outsider with the help of large models. A professor from the Ivy League University of Illinois in the United States has published an academic paper demonstrating that an intelligent AI agent based on GPT-4 can perform fully automated vulnerability mining and exploitation of web vulnerabilities, 1-day vulnerabilities, and 0-day vulnerabilities in the real world[4-6], with results far exceeding those of open source models and open source vulnerability scanning tools. Although the author has not yet disclosed specific implementation details, the feasibility study still reveals potential legal risks. In addition, if hackers invade the backend of large model applications, tamper with generation algorithms, selectively modify large model answers or retrieval results, mislead users into clicking on false links or links that can hijack attack devices, there is also a risk of illegal activities such as illegal intrusion into computer information systems[7].

2.3 Risk of privacy related crimes

The development of generative artificial intelligence has revolutionized the depth and breadth of data dissemination and utilization, while also bringing potential risks of data leakage[8]. While people enjoy the services provided by artificial intelligence models through the internet, service providers are constantly collecting a large amount of private information from users, including personal identity, preferences, behavior patterns, and other information, as well as operational, management, and commercial transaction information at the enterprise level, and policy formulation information at the government level. This may pose a threat to personal privacy, corporate interests, and government image. There are news reports that Samsung Group's equipment solutions department has only been using ChatGPT for over 20 days, but three incidents of confidential file data leakage have occurred, resulting in its semiconductor equipment measurement, yield rate and other information being uploaded to the server. Recently, some users have shared screenshots of other people's ChatGPT conversation records on social media, stating that they can see the topics asked by other users on ChatGPT. OpenAI founder Sam Ultraman also publicly apologized for the incident of user data leakage caused by vulnerabilities in open-source libraries. The hidden danger of personal privacy information leakage can lead to the infringement of privacy rights, while the leakage of corporate organizations or even national government departments can bring more serious legal consequences.

3 Detection method for compliance of generative artificial intelligence outputs

There are two common research directions for evaluating the security and robustness of content generated by generative artificial intelligence: one is to directly monitor and evaluate the content output by the model, and the other is to have the model generate multiple output contents and evaluate their security through each output. There are already various tools available to detect and correct inaccuracies in generative artificial intelligence generation, such as Pythia using knowledge graphs and interconnected information networks to validate the factual accuracy and coherence of LLM outputs, and supporting AI validation through knowledge bases to improve Pythia's accuracy [9]; Galileo used external databases and knowledge graphs to verify the factual accuracy of AI answers, and also evaluated the tendency of LLM to produce hallucinations in common task types such as question answering and text generation, using metrics such as correctness and context adherence to validate facts; Automatic detection and repair of data issues in clean laboratories that may have a negative impact on the performance of machine learning models[10]; The JADE, a large model

targeted security evaluation platform released by Fudan University, creates a large model for questioning high-risk issues in four categories based on linguistic variation: core values, illegal activities, infringement of rights and interests, and discrimination and bias[11]. It explores the boundaries of the security compliance capabilities of the large model on specific topics, helping developers better understand the true security level of the large model and carry out corresponding reinforcement and protection.

4 Risk management methods of generative artificial intelligence

4.1 Improve policies and regulations for the security governance of large models

Due to the rapid development and widespread application of artificial intelligence technology, many security risks have emerged, and artificial intelligence risk management has become a task that must be taken seriously. At the national regulatory level, as the role of artificial intelligence in national competitiveness, national security, and other aspects becomes increasingly prominent, countries are intensifying their planning, deployment, and policy-making efforts to protect basic rights, democracy, the rule of law, and environmental sustainability from the impact of high-risk artificial intelligence.

In 2023, the United States and Europe will pay high attention to artificial intelligence security and risk governance, and have successively issued laws, administrative orders, risk management frameworks, and other regulations and strategic documents in the field of artificial intelligence, coordinating the promotion of development and security. The EU completed the preliminary agreement on the Artificial Intelligence Bill in 2023[12]. On March 13, 2024, the European Parliament officially passed the Artificial Intelligence Bill, which came into effect within the EU on August 1. This is also the most comprehensive bill on artificial intelligence regulation released globally, and it is of extraordinary importance for the development of artificial intelligence and even the entire digital economy worldwide. On October 30, 2023, US President Biden signed the Executive Order "Secure, Stable, and Trustworthy Artificial Intelligence" to ensure that the United States is at the forefront of AI development and risk management[13]. International organizations and governments around the world are actively collaborating to develop unified technical standards and safety regulations to address the cross-border risks of artificial intelligence technology. These global management frameworks not only help enhance countries' control over artificial intelligence technology, but also promote technology exchange and cooperation among countries, and promote the healthy development of global artificial intelligence technology.

Several Chinese ministries and commissions have successively issued relevant specifications and requirements on AI governance and model supervision[14], and formulated detailed laws, regulations and policy documents to ensure the security and compliance of AI technology. For example, in July 2023, the China Internet Information Office and other relevant departments issued the Interim Measures for the Management of Generated AI Services, which is the first industry regulation issued by China to govern generative AI, with the purpose of comprehensively governing and supervising generative AI. On this basis, relevant national and industry standards are also rapidly advancing. The development of these standards not only contributes to the safety, reliability, and controllability of artificial intelligence technology, but also provides clear technical specifications and operational guidelines for industry practitioners. The rapid advancement of standards indicates that the country attaches great importance to the security issues of artificial intelligence, and also provides strong support for enterprises and research institutions in the research and application process.

4.2 Accelerating the technological breakthroughs for the legal use of large models

4.2.1 Suppress illegal content output of large models

Firstly, consider regulating the big model from the source. By building a large model security evaluation platform and constructing a compliance testing question bank based on regulations and industry standards, the security and robustness of the model are comprehensively tested and evaluated. Only large models that pass the security evaluation are allowed to legally and compliantly go public and provide services. By building a large model risk monitoring platform and utilizing defense and detection methods, artificial intelligence systems and algorithms are protected from various threats such as malicious attacks, abuse, data breaches, and improper access. By building a large model security alignment platform, monitoring, scanning, and recording all input and response content during the interaction between the large model and users, and searching for risky content such as malicious code or privacy information, if such content is found, the platform will automatically use Retrieval Augmented Generation (RAG) technology and content rewriting capabilities to convert the original malicious output content into normal output content, preventing legal disputes caused by insufficient compliance of the large model output content[15].

At present, there are relevant achievements in both the industry and academia. For example, scholars from Shanghai Jiao Tong University, Carnegie Mellon University, City University of Hong Kong, Meta and other institutions jointly proposed a task and domain independent universal framework FacTool[16], which can verify the factual accuracy of the content generated by large models. FacTool extracts statements that need to be validated from the generated text, such as a fact, a piece of code, a mathematical expression, or a literature reference. Generate a series of queries based on different tasks and domains to request relevant evidence from external tools such as Google Search, Python interpreter, Google Scholar, etc. Then, using the collected evidence and the reasoning ability of large language models, each statement is validated to determine whether it is correct or supported. Finally, provide factual labels for each statement and the entire text, and attempt to provide explanations and corrections in case of errors. The effectiveness of this method has been demonstrated through experiments on knowledge-based question answering, code generation, mathematical reasoning, and scientific literature review tasks. The entire workflow of fact checking in FacTool can be fully transplanted to verify the legality of input and output of large models, thereby achieving secure alignment of large models.

4.2.2 Strengthen the detection of output content of large models

Strengthen the detection of output content from large models, so that the audience generated by artificial intelligence can understand and judge the production mode and source of the information they receive, identify which content is generated by AI, thereby enhancing vigilance and preventing financial, emotional, and reputational losses caused by being deceived by criminals.

Firstly, in order to identify and track the content generated by the model[17], Large Model Content Watermarking has emerged. Large model content watermarking is a technique that embeds hidden markers in text, images, audio, or video generated content without affecting the readability of the content. There are also various methods for implementing watermarking technology for different modal types of content. For example, text

watermarking technology includes text embedding technology that inserts specific symbols or hidden character phrases into the generated text, encoding technology that uses specific encoding methods such as hash functions to embed watermark information into the generated text word or sentence structure, image watermarking technology includes visible watermarking technology that adds a clear mark on the image, invisible watermarking technology that changes certain pixel values in the image to embed hidden information, audio watermarking technology includes time-domain watermarking technology that embeds watermark information at specific positions in the audio signal, frequency-domain watermarking technology that embeds watermark information in the frequency domain of the audio signal, etc. The large model content watermarking technology increases the transparency and credibility of information, and has a good effect on proactively reducing the legal risks of intelligent applications.

However, the prerequisite for adding watermarks is that the content creator or the service provider of the large model actively follows institutional constraints. For practices that do not actively add watermark information or even actively delete watermarks through technical means, large model generated content detection technology is needed to add an additional layer of protection barrier to the rights and interests of the information receiver. According to whether the source model used for generating content can be accessed, it can be divided into two detection scenarios: white box and black box. In a white box scenario, the detector can use the source model to score and assist in making judgments; In black box scenarios, the detector cannot obtain any information from the source model and needs to rely on other alternative models for scoring. A typical text detection method is Fast DetectGPT[18], which views text generation as a token based sequential decision-making process and proposes a hypothesis that there is a significant difference between humans and machines in selecting tokens given context. By measuring this difference using a method based on conditional probability curvature, efficient zero sample detection of machine generated text can be achieved, achieving good detection accuracy in both white box and black box detection scenarios. There are many AI text detectors with similar functions, including OpenAI's self-developed AI Text Classifier[19], GPTZero[20], Originality AI[21], ChatGPT Detectors[22], and more are constantly being researched and released for image, audio, and video content detectors.

Strengthening the research on content detection technology for large model output, helping users identify and label content created by machines to reduce potential risks to society caused by improper or malicious use of AI models, is of great help in building trustworthy AI systems, improving the transparency and accountability of AI generated information.

5 Conclusion

The rapid development of artificial intelligence, especially generative artificial intelligence, has complex impacts on privacy protection, social ethics, network security, and other aspects. It still faces enormous challenges in terms of robustness, interpretability, fairness, and other issues. It is necessary to prioritize the risk of misuse of AI technology as a global governance issue, and attach importance to the legal risk analysis and governance measures of generative artificial intelligence.

The European Union, the United States, the Japan and China have adopted different models in the governance of AI, reflecting their respective political systems, cultural traditions, and economic development needs. The practice of AI governance has shown that the construction of models should be based on themselves, correctly grasp the differences in national conditions, and promote the healthy development of the AI industry while handling technological innovation and risk managment.

References

1. Y. Song, W. Fan, Analysis of the Impact of Generative Artificial Intelligence on Network Security, Industrial Information Security, 2024, (01): 85-91
2. N. Sun, Y. Bao, The technological security risks and prevention of generative artificial intelligence, Journal of Shaanxi Normal University (Philosophy and Social Sciences Edition), 2024, **53** (01): 108-121
3. L. Li, ChatGPT: Research on Risks and Countermeasures of Generative Artificial Intelligence Applications, Journal of Hubei University of Economics (Humanities and Social Sciences Edition), 2024, **21** (02): 98-102
4. R. Fang, R. Bindu, A. Gupta, et al. LLM Agents can Autonomously Hack Websites. arXiv preprint arXiv: 2402.06664, 2024
5. R. Fang, R. Bindu, A. Gupta, et al. LLM Agents can Autonomously Exploit One-day Vulnerabilities. arXiv preprint arXiv: 2404.08144,2024
6. R. Fang, R. Bindu, A. Gupta, et al. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities. arXiv preprint arXiv: 2406.01637, 2024
7. M. Gao, P. Qin, The Criminal Legal Risks and Countermeasures of Generative Artificial Intelligence, https://bjhdfy.bjcourt.gov.cn/article/detail/2023/05/id/7293675.shtml , 2024-09-14.
8. G. Duan, W. Liu, Data Security Risks and Criminal Law Protection in the Field of Generative Artificial Intelligence, Journal of the Zhengzhou Municipal Party School of the Communist Party of China, 2024, (01): 62-67
9. S. Biderman, H. Schoelkopf, etc, Pythia: A suite for analyzing large language models across training and scaling. arXiv preprint arXiv: 2304.01373
10. https://docs.rungalileo.io
11. M. Zhang, X. Pan, M. Yang, JADE-DB: A Secure Universal Benchmark Test Set for Large Language Models Based on Targeted Mutation, Computer Research and Development, 2024, 61 (05): 1113-1127
12. https://artificialintelligenceact.eu
13. Executive Order on the safe, secure, and Trustworthy Development and Use of Artificial Intelligence, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/exective-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
14. Interim Measures for the Management of Generative Artificial Intelligence Services, https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm
15. Q. Wang, Y. Gui, Briefly describe the legal risks of generative artificial intelligence (ChatGPT) https://baijiahao.baidu.com/s?id=1797040023236706296&wfr=spider&for=pc , 2024
16. I. Chern, S. Chern, S. Chen, et al. FacTool: Factuality detection in generative AI - a tool augmented framework for multi-task and multi-domain scenarios[J/OL]. arXiv preprint arXiv:2307.13528 (2023)
17. A. Liu, L. Pan, Y. Lu, et al. A Survey of Text Watermarking in the Era of Large Language Models[J/OL]. arXiv preprint arXiv: 2301.07597, 2023

18. G. Bao, Y. Zhao, Z. Teng,et al. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature[J/OL]. arXiv preprint arXiv: 2310.05130,2024
19. https://freeaitextclassifier.com/.
20. https://gptzero.me/.
21. https://originality.ai/.
22. B. Guo, X. Zhang, Z. Wang, et al. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection[J/OL]. arXiv preprint arXiv: 2301.07597, 2023.