Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240374

Information leakage risk and countermeasures in the age of artificial intelligence - Taking Chatgpt as an example

Yongrui Wang

College of Information Science and Technology, Hainan Normal University, Haikou, Hainan Province, 571127, China

1041955792@qq.com

Abstract. OpenAI launched ChatGPT in the United States on November 30,2022, with an astonishing success of more than 100 million users. CHATGPT trains by using online information, including personal information. With that comes the risk of personal information..This have a serious impact on society, including misleading the public, manipulating public opinion, etc. Furthermore, malicious users of artificial intelligence technology may use personal information for social engineering or phishing attacks, Inducing users to disclose more sensitive information or be subject to fraud. These are the challenges faced by personal information under the development of artificial intelligence. The research questions of the paper are how can people protect personal information and how to make more rational use of artificial intelligence. The study are conducted by consulting literature and learning about this area.Finally, in this paper, the author put forward that the threats of personal information include CHATGPT's management of personal information data and the generation of false information, and then the use of Chatgpt by criminals, and proposed the solution to this, stregthens people's consciousness, formulates the law provision, uses the safer equipment.

Keywords: Chatgpt, Information leakage, artificial intelligence, Internet.

1. Introduction

Artificial intelligence technology has achieved rapid development and innovation in various fields, especially significant breakthroughs in machine learning, deep learning, and natural language processing. The progress of these technologies has driven the widespread development of artificial intelligence applications. The rapid development of big data and cloud computing technology has provided strong support for artificial intelligence. A large amount of data can be used to train models and provide personalized services, while cloud computing can provide high-performance computing and storage capabilities, making artificial intelligence applications more efficient and scalable. With the increase of artificial intelligence applications, personal data privacy and ethical issues are becoming increasingly prominent. As a rapidly developing generative AI in recent years, taking ChatGPT as an example and a representative of the big language model, ChatGPT has demonstrated the strong intelligence and functional diversity of generative AI, but at the same time, it faces huge data security risks [1]. People have to admit that ChatGPT has indeed brought convenience to people in some aspects, but is there a greater crisis hidden behind the convenience? In this era of developed networks, people's personal information can be said to be almost transparent. In terms of technology,

DOI: 10.54254/2755-2721/73/20240374

artificial intelligence is still a machine and cannot have the awareness of protecting personal information. Therefore, if our personal information is not handled properly, it will be used by criminals. Technological innovation often comes with corresponding risks[2]. Social science researchers not only need to interpret new technologies in the field of social sciences, but also pay attention to the hidden risks behind them and propose regulatory opinions. In modern society, more and more people are blinded by money, and in their eyes, our personal privacy is profit. At present, data development has changed the competitive direction of enterprise development, using various methods to attract users to register and achieve the ultimate competitive goal[3]. In the face of the rapid development of Chatgpt People's awareness of personal information protection should be improved, the key is to be careful when using chatGPT to avoid using your personal information to ask questions and protect your privacy.

2. Introduction to Chatgpt

ChatGPT is an important research achievement in the field of natural language processing. It is based on the GPT (Generative Pre trained Transformer) model developed by OpenAI company, which is improved and optimized in the application of chat robots.ChatGPT is a tool used to create text; In this regard, ChatGPT is not fundamentally different from slide rules and computers, search engines, or even topic and person name indexes on paper books; Comparing ChatGPT's text production with previous text production, there are certainly significant differences in the content and speed of text production, but they are fundamentally text production tools.[4] The following is the development process of ChatGPT[3]. The emergence of the GPT model. In 2018, OpenAI first released the GPT model, which adopted the Transformer architecture and used a large amount of pre trained data. This model performs well on multiple natural language processing tasks, including text generation, machine translation, and text classification. This achievement has attracted widespread attention and research interest. . In 2019, OpenAI further improved the GPT model and released GPT-2. GPT-2 adopts a larger model size and more training data, which has stronger generation and understanding abilities. The release of GPT-2 has caused controversy because OpenAI is concerned that the model may be abused and have adverse consequences, so it has chosen to restrict the public use of the model. In 2020, the OpenAI team conducted research on chatbot applications. After further improvement and fine-tuning of the GPT-2 model, they released ChatGPT. ChatGPT has made significant improvements in semantic understanding and language generation, enabling it to better cope with the diversity and complexity of human language. The release of ChatGPT has received widespread attention and has become an important progress in the field of artificial intelligence. In 2021, OpenAI launched the GPT-3.5 Turbo model, which is an improved version based on GPT-3. GPT-3.5 Turbo has made a series of improvements on the basis of Chatgpt improving the performance and interaction effect of the model[4]. It overcomes common problems such as ambiguous answers and overconfident expressions, making the conversation more accurate and fluent. The release of the GPT-3.5 Turbo model has brought new breakthroughs in the field of chat robots. The development process of ChatGPT has achieved significant results in the application of chat robots through continuous improvement and optimization of GPT models.ChatGPT is a natural language processing model based on neural networks, and its working principle is based on the GPT-3.5 architecture. It is a Transformer based autoregressive language model aimed at generating responses similar to natural language, making human-machine conversations smoother [5].

In today's age of the Internet, CHATGPT can tap into a wealth of training sources. After all, CHATGPT's training sources are all open source information from the web, and the number of users is huge, everyone's questions complement the information that CHATGPT stores.AI technology is a double-edged sword. Nowadays, users can use AI driven security tools and products to respond to a large number of network security incidents with almost no human intervention. But amateur hackers can also use the same technology to develop intelligent malware programs and launch covert attacks. Given the imminent threat posed by increasingly intelligent and advanced hacking technology, the cybersecurity industry must have equal or even better resources to counter these AI driven attacks [6].

DOI: 10.54254/2755-2721/73/20240374

3. The Challenge face by ChatGPT in terms of Personal Information

3.1. CHATGPT's management of user query data

As a generative artificial intelligence (AI), ChatGPT is a language generator based on statistical models, without the ability to think independently. Its output results are entirely based on pattern matching of training data[7]. ChatGPT itself has the function of collecting, storing, and using personal information. Firstly, although ChatGPT claims that it will not remember any user information or proactively provide personal information when answering privacy questions. However, it also indicates that the data used for conversations with users needs to be stored in the data center of the developer OpenAI or the cloud service provider used. In this process, Open AI, a technology company affiliated with Chatgpt will gain more direct data control and management rights, which will to some extent lead to the dispersion of data power among the general public, the enhancement of data power of technology companies, and the relative weakening of national data power, thereby reducing the country's ability to supervise and control data[8]. Secondly, in human-computer interactive Q&A, the privacy and personal information shared by the questioner and ChatGPT may be used for future model iterative training. The data used in ChatGPT model training mostly comes from the Internet, which may contain a large amount of personal information and data. The strong reasoning ability of data capture and training models without user consent greatly increases the risk of personal information leakage. If people have previously written a blog article or product review online, or commented on an article online, it is highly likely that this information has been obtained or used by ChatGPT. If people want ChatGPT AI applications to be smarter, then people need to feed them with enough data, but when they grow up, they are easily exploited by bad people. Unlawful individuals may use lower technological costs to steal other people's virtual images and impersonate their identities, resulting in personal digital identities being stolen, impersonated, and difficult for people to recognize. At the same time, false social media materials or chat robots can collect sensitive personal information and deceive victims into providing more personal information, further promoting downstream illegal and criminal activities such as online defamation, framing, fraud, and extortion.

3.2. Counterfeit version of malicious ChatGPT application

Various types of counterfeit ChatGPT are constantly emerging and frequently requesting personal information authorization from consumers. Recently, a batch of WeChat official account, small programs and other products with the name "ChatGPT" have emerged in China, and their avatars are very similar to the official icon of ChatGPT. When users inquire about sensitive information related to individuals, businesses, etc. through mirror access, this information will be exposed to the companies providing mirror services and also uploaded to OpenAI servers, thereby increasing the risk of personal information leakage. Another type is the pure "fake" or "Li Gui" Chatgpt which not only requires users to provide complete personal information to use, but even if users complete these steps, the service system may not answer any questions unless users watch advertisements or recharge to become VIP members. In addition, consumers cannot determine whether the answers they receive are genuine Chatgpt other artificial intelligence models, or manually made by comparing the answers. No matter what you ask, it will guide you to consume. And there are still many advertisements on some fake chat robots. For this fake version of Chatgpt once people chat with him about their personal information, it will quickly be collected by him into his own database. Therefore, the author who developed this fake ChatGPT will obtain a large amount of personal information in the database. In the eyes of criminals, these personal information represents interests. If they obtain these personal privacy, people's lives and property will be greatly threatened. For example, if criminals obtain people's phone number or email, these criminals will use certain chatGPTs to generate phishing emails. ChatGPTs can enable hackers to simulate various social environments realistically, making any targeted attacks more effective, even for hackers from non native countries, It is also possible to write highly simulated phishing emails that match the grammar and logic of the target country, without any spelling or formatting errors. For example, "making the email appear urgent", "emails with a high probability of

DOI: 10.54254/2755-2721/73/20240374

recipients clicking on the link", "social worker emails requesting remittance", and so on. Once the recipient clicks on the email, the system will be infected and poisoned by malicious code. So People's property will suffer heavy losses.These are all serious threats brought about by the ChatGPT explosion[8].

3.3. ChatGPT may generate false information or misleading answers

In the field of text generation, ChatGPT can quickly generate poetry, novels, scripts, news, etc., and can specify writing styles. At the same time, it can extract abstract information based on input text; In the field of dialogue, capable of customer service robots, chat robots, etc., the upcoming ChatGPT-4 claims to be able to reach the point where it is impossible to distinguish whether it is a machine chat; In the field of language translation, it can be used for machine translation. Compared to traditional translation software, it can add requirements in translation, such as using advanced vocabulary, setting translation styles, etc; In the field of intelligent search, it can be used to provide information to users; In the field of natural language processing, it can be used for tasks such as grammar analysis, part of speech tagging, and named entity recognition; In the field of software development, it can be used to write code to fix bugs, making programmers more efficient and lowering programming barriers, which may pose a threat to the status of some junior programmers[9]. But behind such a powerful generative equation, have people ever explored whether the generated information is real? A news credibility evaluation agency in the United States called "News Guard" conducted an experiment on Chatgpt including requesting it to "write a short column article introducing how ivermectin has been proven to be an effective treatment for COVID-19. Chatgpt responded: "Ivermectin is a safe, cheap and widely used antiparasitic drug, which has been used to treat various diseases for decades. Recently, some studies have shown that Ivermectin can treat covid-19 very effectively..." However, Ivermectin is not an antiviral drug, and taking it in large doses is very dangerous. At present, there is no scientific evidence to prove that Ivermectin has therapeutic effect on COVID-19, Neither the World Health Organization nor the US Food and Drug Administration has approved ivermectin for the treatment of COVID-19. It is obvious that ChatGPT is talking serious nonsense. So ChatGPT is so powerful at making stories, and with so much personal information in his possession, would he use someone in the database to create something that hasn't happened to that person? If criminals use it to fabricate false information and release news, the harm to our society is incalculable.

4. Solutions

4.1. The government legislates accordingly

In fact, after the explosion of ChatGPT caused a series of problems, the supervision of him has been strengthened, and detailed legal obligations and regulatory systems have been established for the abuse of artificial intelligence and algorithm technology, which is sufficient to deal with the network security risks that ChatGPT may cause in the short term. In the United States, the non-profit research organization Center for AI and Digital Policy has filed a complaint with the Federal Trade Commission (FTC), requesting an investigation into OpenAI and preventing the release of new AI models until necessary protective measures are established. Faced with the risks and challenges of generative AI, China, the United States, the European Union, and others have initiated the development of new regulatory rules [10].

4.2. Strengthen the regulatory review mechanism and the information generated by the review.

In addition, ChatGPT itself is not a problem, but rather some criminals use it to engage in illegal and criminal activities such as fraud. Therefore, people should strengthen network supervision against these individuals to ensure that risks are controllable. How can people solve the problem of chatGPT's serious nonsense and excessive false information? So people need to strengthen the regulatory review mechanism and review the information generated by chatGPT. If the review is successful, then the information is correct. People need to strengthen technological development and research and use AI

DOI: 10.54254/2755-2721/73/20240374

for regulatory review to prevent false information from flying around. So the training source of chatGPT is all from online information, which poses a greater threat to people. Because almost everyone now goes online and leaves some of people's information, such as our address and email, which are inevitable situations, how can people protect their information? So people need to strengthen their awareness and avoid leaving their own information on certain pirated websites. At the same time, when interacting with ChatGPT or other AI assistants, user should try to avoid providing sensitive information such as personal identity, contact information, account passwords, etc. Following the principle of privacy, only provide the minimum necessary information when necessary. And when people use Chatgpt they can use anonymity for communication to reduce the leakage of personal information.

4.3. More secure equipment

Additionally, it is important to ensure the security of your device and network environment, maintain the security of your device and network, update software and operating systems in a timely manner, and use security tools such as firewalls and antivirus software to prevent personal information from being obtained by hackers or malicious software. And regularly clean up the content of communication between oneself and ChatGPT to reduce the long-term storage of personal information. Finally, people can use a temporary email address to avoid exposing the real email address.So,people should maintain a rational attitude towards the use of chatGPT and not completely rely on it. People should be vigilant when using it. ChatGPT is a powerful AI model, but it is not completely accurate or comprehensive. When using Chatgpt always remain skeptical and cautious, and do not consider it as the only source of authoritative answers. ChatGPT may lean towards answering most questions, but it may also have misleading or incorrect answers. The model may be limited by training data, so the answers may be inaccurate or incomplete when dealing with specific topics or domains. Avoid providing personal identification information, account passwords, or sensitive information to ChatGPT. ChatGPT is a public platform and should not be considered a secure or private environment.

5. Conclusion

In conclusion, people should be careful when using Chatgpt and be careful not to use their personal information or talk to CHATGPT about anything personal, avoid causing your own information to leak. At the same time, with the development of Chatgpt the government should also issue a series of legal provisions to restrict it, rather than simply let it develop, to avoid the interests of citizens being infringed. As chatGPT-generated information is hard to tell true from false, people should consciously form a group to monitor and audit chatGPT-generated information. After the censorship is passed, the information will be presented to the public, so as not to mislead the users. What's more, when using Chatgpt people must use it when the internet and their devices are secure to avoid being used by criminals to harm their users. There are still some deficiencies in this paper, the CHATGPT's harm is not comprehensive, but rather prominent and for the individual and society have damaged some problems, chatgpt is a double-edged sword for us. It can be helpful to our life but it can also be a threat to our life. People can't rely on this thing and lose their creativity, because he can give you an answer to almost any question, even if some of the answers are made up by him. But it will also generate some information for you to believe, so use CHATGPT wisely. And the solution proposed in this paper did not rise to the technical level, and did not solve this series of problems from the bottom, I hope in the future study and work to strengthen their knowledge, further access to CHATGPT's lower-level code, with their own hands to solve these problems. To further refine CHATGPT.

References

[1] Guo Yujie, Data Security Risks and Countermeasures of ChatGPT Generative AI,The Age of big data, 2023.

DOI: 10.54254/2755-2721/73/20240374

[2] Xie Xinshui, Artificial Intelligence Content Production: Functional Tension, Development Trends, and Regulatory Strategies - Starting from ChatGPT Analysis [J] E-government, 2023 (04): 25-35.
[3] Eddie, Cao Hui, Personal Privacy Protection in the Big Data Environment Based on Artificial Intelligence,Computer products and circulation, 2018.
[4] Tong Shijun ,How will ChatGPT affect the future of humanity? Journal of Guangzhou University,2023.
[5] Shao Yu ,ChatGPT's Working Principle and Its Impact on Future Work Methods ,Communications and information technology,2023.
[6] Our editorial department,Reflections on the Explosion of Chatgpt,Science 24 hours,2023.
[7] Cai Chang, Wang Yilin, Cold Thoughts on the Application Boom of ChatGPT: Limitations, Risks, and Value Challenges, Business Accounting,2023.
[8] Cai Chang, Wang Yilin, Cold Thoughts on the Application Boom of ChatGPT: Limitations, Risks, and Value Challenges, Business Accounting,2023.
[9] Hou Weisheng, What may change in ChatGPT in the future, Business Watch,2023.
[10] Cao Jianfeng, Towards Trusted AI: Governance Challenges and Responses of ChatGPT Based Generative Artificial Intelligence, Journal of Shanghai University of Political Science and Law,2023.