

Research on Personal Information Protection in the Context of Generative Artificial Intelligence

Tingting Yan

School of Anhui, Anhui University of Finance and Economics, Bengbu 233030, China.

Abstract

Personal information is the foundation of generative AI, and ChatGPT-like generative AI needs to process a large amount of personal information at various stages such as model training, model generation, and model optimization, which also has a certain impact on traditional personal information protection rules. During the information collection phase, generative AI may fugitive the informed consent rules and infringe on the privacy rights of information subjects. In the information utilization stage, generative AI may impact basic personal information processing rules such as the principle of purpose limitation and the principle of openness and transparency, increasing the risk of personal information leakage. At the information generation stage, generative AI can generate false and discriminatory information. Therefore, in the context of generative AI, personal information protection is faced with the problems of the notification and consent rules being hollowed out, the principle of minimum necessity being voided, and the frequent leakage of personal information. Based on this, it is necessary to promote the transformation of the "personal control center to the risk control center" of the notification and consent rules, promote the risk-based interpretation of the principle of least necessary, and improve the risk-based personal information protection compliance system to solve the problem of personal information protection in the context of generative AI.

Keywords

Generative AI; ChatGPT; Protection of Personal Information.

1. Introduction

At the beginning of 2023, ChatGPT has become an important milestone in the field of artificial intelligence with its excellent natural language processing capabilities, causing disruptive changes in social production and life and future technological development. However, generative AI based on massive corpus data has brought positive changes and impacts to society, but also raised many issues at the level of personal information protection. On March 20, 2023, ChatGPT suffered a serious information security incident, and sensitive information such as chat log fragments, user credit card information (including the last four digits and expiration date), name, email address, and payment address were leaked to some users. On March 31, Italy's data protection authority, GPDP, announced that it would immediately impose temporary restrictions on ChatGPT due to OpenAI's collection, use, and disclosure of personal information without consent, making it the first government ban on ChatGPT worldwide. There are many signs that generative AI applications are about to become massively adopted. This means that the potential risks brought by generative AI to the protection of personal information may continue to ferment with the development of AI technology. In order to promote the healthy development and standardized application of generative AI, the Cyberspace Administration of China, together with the National Development and Reform Commission, the Ministry of Education, and the Ministry of Science and Technology, promulgated the Interim Law on the Management of Generative AI Services on July 13, 2023,

which will come into force on August 15, 2023. While welcoming the development of emerging AI technologies, human society must face up to and calmly examine the legitimacy of generative AI technologies, and must think forward-looking, design, and configure a matching and compatible personal information protection system.

2. The Basic Concepts and Technical Logic of Generative AI

2.1. Basic Concepts of Generative AI

Generative AI refers to an AI software system that can autonomously generate text, pictures, audio, video, code, and other content based on algorithms, models, and other technologies. Through a large amount of data feeding and repeated training based on neural network models in the early stage, generative AI has the ability to capture data with high capacity and accuracy, and with the support of large language models (LLMs), it can discover language rules from learning materials, and recursively summarize words and sentences, so as to generate and even "create" in real time according to the user's instructions. At the end of 2022, ChatGPT, a generative AI software launched by OpenAI, was born, and the number of monthly active users exceeded 100 million in just two months, making it the fastest-growing app in history. ChatGPT, which stands for "Chat Generative Pre-trained Transformed", is a chatbot based on the GPT-3.5 language model. ChatGPT can process natural language and communicate with users, and when the user enters a Prompt, it can accurately understand the user's intent, generate long or short answers according to the instructions, and can link with other AIGC to generate pictures, videos, etc. ChatGPT can be used not only in everyday conversations, but also in activities such as language translation, business plan writing, legal advice, and coding, and the ability to question, admit mistakes, and reject unjustified requests, so that it can communicate fluently and almost indiscriminately with humans. In the 50s of the 20th century, the British mathematician Turing proposed the "Turing test", that is, the tester in a closed hut talks to two test subjects by typing, one of which is a computer, and the other is a living person, and the tester judges whether it is a person or a computer outside the hut by constantly asking questions and receiving the answers of the test subject, if the computer can imitate the human answer very well and the tester has a misjudgment, it means that the "Turing test" [1], ChatGPT Considered to be the closest AI model to the "Turing test" today. From the perspective of the logic of ChatGPT's input and output, ChatGPT is based on the underlying technology of the Generative Pre-Trained Transformer (GPT) model, and has been developed through continuous iteration. GPT-1, GPT-2, GPT-3 and ChatGPT are all models with Transformer as the core structure, GPT-1 launched by Open AI in 2018 has 117 million parameters and about 5GB of pre-trained data, while GPT-3 launched in May 2020 has reached a staggering 175 billion parameters and about 45TB of pre-trained data The "Self Attention" mechanism was used by GPT deep learning to obtain a general "pre-trained" version of the model in unsupervised training mode. On the basis of GPT-3, ChatGPT has added RLHF (Reinforcement Learning from Human Feedback), which artificially annotates and sorts multiple responses of the model, and then forms an answer that is infinitely close to human itself.

2.2. The technical logic of generative AI

Although generative AI is new, its technical theories have a long history. From the perspective of the history of technological development, generative AI originated from "cybernetics" in the 40s of the 20th century [2], and the chatbot Eliza appeared in the 60s of the 20th century is the earliest prototype of generative AI, marking the possibility of natural language dialogue between humans and computers. At the legislative level, China's Interim Measures clarify the concept of generative AI for the first time, and according to Article 22 of the Interim Measures, generative AI technology refers to models and related technologies that have the ability to generate content such as text, images, audio, and video. It can be seen that generative AI is

multimodal, and it can not only process text tasks, but also generate other forms of content such as pictures, audio and video, etc.

At this stage, generative AI such as Dall-E2, Stable Diffusion, ERNIE Bot, and ChatGPT are emerging, and this paper uses ChatGPT, which has a large user base, as an analysis model, to clarify the technical logic of generative AI. Compared with previous chatbots such as Siri and Xiaoice with single functions and blunt conversations, ChatGPT can "remember" previous chats with users, thus achieving consecutive rounds of conversations. In addition, ChatGPT can also refuse to answer inappropriate requests from users and proactively admit mistakes. According to the relevant content published on OpenAI's official website, the operation mode of ChatGPT can generally be summarized into the following three stages, that is, user input instructions → algorithm integration and analysis → generate corresponding content. Specifically, the user enters an instruction into a dialog box specified by OpenAI, and then the system quickly analyzes the instruction and refuses to answer the user's request if the analysis determines that the user's instruction violates its built-in ethics review mechanism. On the contrary, if the instructions of the censored user do not violate its censorship mechanism, it will search for related words in a large-scale corpus, predict the frequency of the next word, and then integrate and optimize the collected words, and finally output text, images, videos and other content that conform to human natural language expression habits.

ChatGPT is a simplified version of Chat Generative Pre-trained Transformer, in which the Transformer architecture is the key technology of ChatGPT. The Transformer model was born in 2017 in the article "Attention Is All You Need" published by a Google research team, which pointed out that Transformer is a sequence transformation model based on a self-attention mechanism, which uses multi-head self-attention to replace the most commonly used recurrent layer in encoder-decoder architectures, so that Transformer can be trained significantly faster than recurrent or convolutional layer-based architectures. The use of the Transformer model enables ChatGPT to quickly learn the relationship between data, and through a large number of debugging of the model's input and output by AI trainers, the model can initially master the grammar rules of natural language. In order to optimize the expression of the model and make the generated content more in line with the cognitive concept of human society, ChatGPT adopts reinforcement learning from human feedback on the basis of the GPT-3.5 model [3].

3. Challenges of generative AI applications to personal information protection

3.1. The challenges posed by generative AI in gathering information

As the basic principle of personal information collection, the principle of informed consent is recognized by the legislation of various countries (regions). However, according to the privacy policy published on OpenAI's official website, it is clear that it does not strictly implement the rules of informed consent when collecting personal information. At the notification stage, the information handler shall clearly inform the information of the purpose and method of information processing, and other specific matters, to ensure that the information subject is fully informed. However, OpenAI's claim that it may use "the personal information it collects for the purpose of improving services, conducting research, preventing criminal activities, etc.", is vague and uncertain, and cannot provide clear behavioral expectations for information subjects. In the consent stage, the information processor can only collect personal information with the independent and genuine consent of the information subject, and if it changes the information processing method during the processing of personal information, it shall obtain the consent of the information subject again. Although OpenAI claims that the data it uses to train the underlying model comes from publicly available data and licensed data, it has not disclosed the specific source of the data to the public, and the legitimacy of the data source is

questionable. In addition, OpenAI states in its privacy policy that "when a user creates an account to use ChatGPT, OpenAI has the right to collect personal information such as the user's name, contact information, and account credentials at any time", which seriously violates the basic connotation of the principle of informed consent.

In addition to fugizing the informed consent rules, generative AI may also violate privacy in the process of collecting personal information. ChatGPT uses a generative pre-trained transformation model to actively capture Internet data without human supervision, coupled with the lack of necessary screening and filtering mechanisms, resulting in the collection of data may contain private information, and due to the intrinsic collusion of personal information, even if ChatGPT collects non-private information, these massive data may become private information after aggregation and reorganization, thus threatening the privacy and security of the public [4].

3.2. The challenges posed by generative AI harnessing information

In the process of generative AI using personal information, due to the existence of "algorithm black box" and other uncertain factors, it is difficult to implement the principle of purpose limitation and the principle of openness and transparency, and the massive data processing increases the risk of personal information leakage. The purpose limitation principle is a basic principle that should be followed in the use of personal information, which aims to prevent information processors from arbitrarily processing personal information [5]. According to Open AI's privacy policy, it may "use the collected information to analyze user behavior and characteristics" and "develop new services", which exceeds the purpose of processing at the time of information collection and may frustrate the reasonable expectations of the information subject. The principle of openness and transparency requires information processors to publicly disclose the specific details of the processing of personal information, but the deep learning technology on which generative AI relies is a "black box", and there is a "hidden layer" between the input data and the output results of the generative AI system [6].

3.3. The challenges posed by generative AI to generate information

In terms of generating content, ChatGPT may generate seemingly reasonable but false content because it cannot achieve true reasoning, and it is difficult for humans to distinguish the authenticity of the information under the logical packaging of artificial intelligence [7]. Unlike traditional search engines, ChatGPT outputs unique content and does not provide multi-source information references, making false or misleading information difficult for users to detect. For example, a user used ChatGPT to edit a news item that "Hangzhou will cancel the restriction on the tail number of motor vehicles on March 1", but a large number of the public did not recognize it as false news, causing a negative social impact. In addition to disinformation, generative AI may also output discriminatory information that challenges legal equality principles. ChatGPT's initial model is supported by massive amounts of data, which inevitably suffer from human cognitive biases, and because generative AI is not capable of learning, it extracts relevant information from large databases and uses certain permutations and combinations to form specific texts. Therefore, if there is a bias in the data in the database, then there must be a bias in the content of the output. In addition, generative AI systems may embed explicit or implicit biases of designers, which are difficult to remove from algorithms, resulting in value biases in generated information.

4. Recommendations for personal information protection in the context of generative AI

4.1. Promote the transformation of the "personal control center to the risk control center" that informs the consent rules

The notification and consent rules include two levels: "notification" and "consent", the former requires the information processor to clearly indicate to the personal information subject the scope, method, purpose of information processing, and other content that affects the personal information subject's consent to make a decision, while the latter requires the information processor to obtain the personal information subject's explicit or even separate consent for the processing behavior. For the collection of personal information in generative AI, in principle, the same rules need to be observed. The Measures for the Administration of Generative AI Services also explicitly regard authorization and consent as one of the basic criteria for the development and operation of generative AI. However, as mentioned earlier, the mandatory and intangible nature of generative AI corpus data collection makes the rules of consent largely non-existent. If the strict principle of notification and consent is followed, generative AI technology service providers must inform all information subjects one by one and obtain explicit consent, and in the case of the collection of sensitive personal information, they must obtain separate consent from the information subjects. Obviously, the adoption of strong consent rules may result in generative AI not working properly, and individuals will therefore bear the burden of lengthy privacy protocol interpretation. Therefore, in order to ensure the effective protection of personal information rights and interests, and to promote the sustainable and healthy development of generative AI technology, it is necessary to interpret and debug the notification and consent rules in a risky manner, and promote the transformation of the notification and consent rules from "personal control" to "risk control" as the center.

The risky interpretation and adjustment of the notification and consent rule refers to the risk that may be raised as the criterion for judging the reasonableness of personal information processing. Specifically, if the information subject does not explicitly consent or explicitly refuse to the information collection behavior of the provider of generative AI technology services, but it can be confirmed through the personal information protection impact assessment that the information processing does not pose an unreasonable risk, then it can be presumed that the information subject has implied consent. This approach is in line with what the United States has adopted! The "optional informed consent mechanism" is similar, that is, enterprises do not have the obligation to ask for consent when collecting users' personal information, and only need to disclose information to users before or when the collection occurs [8]. From federal legislation to state legislation in the United States, the opt-in informed consent mechanism has always been dominant, and it is more conducive to improving the efficiency of the commercial use of user behavior information. It should be noted that the so-called "opt-out consent" means that even if the information subject is deemed to have implied consent, he or she still has the right to opt-out during the processing of the information, and the previous implied consent can be overturned through a new expression of intent. The same is true for generative AI applications, where users still have the right to withdraw (implicitly) consent in subsequent corpus data storage, language model training, etc. To this end, generative AI technology service providers should be required to provide a convenient way to withdraw consent in order to assist users in exercising this right.

On the contrary, if the assessment confirms that the risk level is high, in addition to obtaining the explicit consent of the information subject, the generative AI technology service provider should also initiate an enhanced notification procedure to give full play to the role of the notification rules in the protection of personal information. This requires generative AI technology service providers to explain the relevance of personal information to model training,

including the fact that "human-machine dialogue is information collection" and important matters such as how the accuracy of personal information will affect the quality of generated content, in order to provide guidance for users to make risk judgments. Secondly, the name, purpose, duration, and method of the sub-processor. Generative AI technology service providers such as OpenAI entrust all or part of their processing activities to sub-processors, and users do not enter into any information processing agreements with these third parties, which may bring more risks to the protection of personal information than expected [9]. Therefore, in the case of a high risk level confirmed by the assessment, the generative AI technology service provider should disclose the information processing of the sub-processor in detail to the information subject, so that the information subject can fully understand the risk and consequences and make decisions.

4.2. Advance the risk-based interpretation of the principle of least necessary

Unlike the notification and consent rule, which focuses on prior protection, the principle of least necessary mainly emphasizes the protection of the legitimate rights and interests of the information subject during and after the event, and it regulates the information processing behavior through the requirements of relevance, minimization, and proportionality, and plays an indispensable and important role in the personal information protection system. However, there is a huge tension between the requirements of the minimum necessity principle and the realization of the main processing purpose of generative AI: the operating mechanism of generative AI determines that it requires a large amount of data-driven model training, and this process can easily lead to problems such as the collection of personal information beyond the necessary scope, the unclear information processing period, and the unclear use of information, and the conflict with the minimum necessity principle, so that the application of the minimum necessity principle is limited in practice. In order to resolve this dilemma, the principle of minimum necessity can be reinterpreted in the way of risk-based transformation, that is, the principle of minimum necessity can be promoted to "reasonable and necessary" based on the risk control under scenario analysis. In this way, it is possible to avoid the rigid application of the principle of minimum necessity that limits the development of generative AI technologies, while preserving and recognizing the principle of minimum necessity [10].

In this case, the principle of least necessary will be regarded as a risk prevention and early warning mechanism, requiring generative AI technology service providers to control potential risks within a reasonable level when processing personal information. This will lead to the flexible use of the principle of minimum necessity to better adapt it to the practical use cases of generative AI. In contrast to China, the relevant principles under the EU legal system are more flexible. For example, the EU GDPR provides for a balancing test mechanism, requiring data processors to consider factors such as balance of interests and risk assessment when weighing the necessity and legality of personal information protection and processing, so as to ensure a dynamic balance between the reasonable use and protection of personal information [11]. In fact, the rigid application of the principle of minimum necessity in China's practice has been questioned by many scholars. Some scholars are of the view that the "prohibition of excessive damage" should be an important consideration in the application of the minimum necessity principle. Some scholars have also suggested that the principle of minimization should be replaced by the "principle of limitation" under certain circumstances, and these propositions actually imply the thinking of risk control [10]. It can be seen that in the application of generative AI, it is necessary to consider the risky interpretation of the principle of least necessary as a necessary measure that should arise from time to time.

At the same time, it is also necessary to clarify some specific circumstances as constraints to balance the uncertainty caused by the principle of least necessary interpretation by risk. Specifically, it includes the following aspects: First, the collection of personal information is still

limited to the purpose of achieving the operation of generative AI, and personal information that is not necessary to achieve that purpose should not be collected. For example, the collection of personal information, such as search logs, equipment information, and communication information, is not relevant to the purpose for which generative AI operates, and the collection of such information is not "reasonably necessary" and increases security risks. Secondly, the time limit for the processing of personal information must also comply with the principle of minimum necessity. According to Article 47 of the PIPL, when the information stored by generative AI expires or is no longer relevant to its training and application, the storage institution is obliged to delete or de-identify it within a reasonable time. The deletion mechanism is a concrete embodiment of the principle of least necessary and the principle of purpose limitation [12], and should be implemented in order to reduce the risk of information accumulation in situations where the storage time of generative AI information is difficult to determine. Finally, the collection and processing of sensitive personal information should be strictly controlled. Sensitive personal information is subject to "a high risk of damage to basic rights". The global personal information protection system pays special attention to the collection and processing of sensitive personal information, especially the European Union's GDPR directly establishes the principle of "general prohibition and exception permitting" for the processing of sensitive personal information. Therefore, generative AI technology service providers must strictly control the collection and processing of sensitive personal information. Even if it is difficult to avoid the collection of sensitive personal information under certain circumstances, it should be avoided as much as possible for model training to reduce the risk of sensitive personal information being leaked.

4.3. Complete risk-based compliance systems for the protection of personal information

With the formal implementation of the Personal Information Protection Law, more and more enterprises have begun to establish personal information protection compliance management systems to standardize the processing of personal information within enterprises and prevent the occurrence of security incidents such as personal information leakage. Especially for large personal information processors, the establishment of a personal information protection compliance management system is necessary to achieve personal information security. Article 58 of the PIPL clearly stipulates the obligation to "establish a sound regulatory system for the protection of personal information" for "super Internet platforms". Therefore, as large-scale personal information processors, AI service providers should also undertake the obligation to improve the personal information protection compliance system. In the future, efforts should be made to embed personal information protection compliance content in the generative AI information security assurance system, focusing on the following two aspects:

First, establish a generative AI personal information protection impact assessment system. Risk assessment is a core component of risk control in all areas and forces information handlers to "identify, assess and ultimately manage the high risks of the processing of personal information to rights and freedoms" [13], helping to reduce the risk of personal information leakage at the source. In the application of generative AI, it involves sensitive personal information processing, the use of personal information for automated decision-making, the entrustment of personal information processing and other information processing high-risk situations, which are more likely to cause personal information disclosure, and personal information protection impact assessment should be regarded as a core task of information protection compliance. In its 2022 publication *How to Use AI and Personal Information Appropriately and Lawfully*, the UK's Information Commissioner's Office (ICO) made it clear that data controllers should conduct a personal information protection impact assessment before using AI systems, and take measures appropriate to their risks to mitigate possible damage. The Measures for the

Administration of Generative AI Services promulgated by the Cyberspace Administration of China and other departments on 13 July 2023 do not cover the assessment of the impact of generative AI on personal information protection, and the relevant content still needs to refer to Articles 55 and 56 of the Personal Information Protection Law. However, on the whole, the personal information protection impact assessment system in China's Personal Information Protection Law is still relatively sparse, and the application of generative AI needs to be further improved. Critical of these is the lack of mechanisms to enforce disclosure to the public, and the lack of public oversight will make it difficult to ensure that personal information protection impact assessment risk prevention tools are developed in the public interest [14]. In response to this issue, the Guidelines for Personal Information Security Impact Assessment of Information Security Technology have clearly defined a series of measures. In the personal information protection impact assessment of generative AI, the guidelines can be used as a reference to make the personal information protection impact assessment report public or form a public representative committee to ensure that the assessment process is subject to external supervision.

Second, set up an independent supervisory body for the protection of generative AI personal information. According to Article 58 of the PIPL, large internet platform enterprises have a special obligation to establish an independent supervisory authority to ensure that their personal information processing meets compliance requirements. The applicable standards for this obligation include three aspects: Personal information processors that provide important Internet platform services, have a large number of users, and have complex business types. Based on this standard, the Guidelines for the Classification and Grading of Internet Platforms (Draft for Comments) and the Guidelines for the Implementation of Entity Responsibility of Internet Platforms (Draft for Comments) further elaborate on them. Generative AI, represented by ChatGPT, is a new type of AI technology that gained 100 million users in just two months. The amount of personal information processed is similar to that of large Internet platform enterprises, and is widely used in important fields such as promoting economic development, protecting people's livelihood, and safeguarding national security, and is an information processor that "provides important Internet platform services". Therefore, generative AI should establish an independent oversight body to monitor the protection of personal information and fully fulfill its social responsibilities.

Given that the processing of personal information in generative AI applications is often accompanied by highly complex algorithms and technologies, the staff of the independent oversight body should include professionals with the appropriate expertise and technical capabilities. In terms of the scope of their duties, independent supervisory bodies should supervise the compliance of personal information security in generative AI applications from three aspects: institutional compliance, technical compliance, and organizational compliance. This includes the establishment of a person in charge of personal information protection, the assessment of personal information cross-border security, the emergency plan for personal information security incidents, compliance audits, and the publication of reports on social responsibility for personal information protection. In general, imposing special obligations on generative AI technology service providers to protect personal information is an important measure for risk control.

References

- [1] Z, W. Feng, D, K. Zhang and G, Q. Rao: From the Turing Test to ChatGPT-Human-Computer Conversational Milestones and Revelations, *Language Strategy Studies*, Vol. 7(2023)No. 8, P. 20-24.
- [2] Q, P. PU, W. Xiang: Generative AI: ChatGPT's transformative impact, risks and challenges, and coping strategies, *Journal of Chongqing University (Social Sciences)*, Vol. 28(2023)No. 3, P. 102-114.

- [3] G, H. Zhu, X, W. Wang: ChatGPT's operating model, key technologies, and future landscape, Journal of Xinjiang Normal University (Philosophy and Social Science), Vol. 43(2023)No. 4, P. 113-122.
- [4] C, Z. Guo: Coherent Legal Governance for Generative AI: A Case Study of Generative Pretrained Models (GPT), Modern jurisprudence, Vol. 44(2023)No. 3, P. 88-107.
- [5] R, R. Zhu: Reflection and Reconstruction of the "Purpose Limitation Principle" of Personal Information Protection: Focusing on Article 6 of the Personal Information Protection Law, Financial Law, Vol. 7(2022)No. 1, P. 18-31.
- [6] F. Xu: Legal regulation of the black box of artificial intelligence algorithms: a case study of robo-advisors, Oriental Jurisprudence, Vol. 11(2019)No. 6, P. 78-86.
- [7] Y, H. Zheng, Y, N. Ding and Y. Zheng: Multi-field Changes and Challenges Arising from ChatGPT-like Artificial Intelligence (Written Talk), Journal of Tianjin Normal University (Social Sciences), Vol. 49(2023)No. 3, P. 49-63.
- [8] J, N. Zheng: Application of the Principle of Consent in Information Gathering and Rule Construction, Oriental Jurisprudence, Vol. 12(2020), No. 2, P. 198-208.
- [9] M, D. Huang, F. Zhao: Private laws and regulations on entrusted handling of personal information, Journal of Chongqing University (Social Sciences), Vol. 27(2022)No. 4, P. 203-215.
- [10] T. Wu: The minimum necessity principle is applicable in the practice of handling personal information on the platform, Legal studies, Vol. 67(2021), No. 6, P. 71-89.
- [11] L. Xie: Exemption from the legitimate interest in the use of personal information in the era of big data, Political and Legal Forum, Vol. 40(2019), No. 1, P. 74-84.
- [12] X. Cheng: On the right of erasure in the Personal Information Protection Act, Social Science Series, Vol. 49(2022), No. 1, P. 103-113.
- [13] Q. Liu: On the impact assessment of personal information protection – focusing on Articles 55 and 56 of the Personal Information Protection Law, Journal of Shanghai Jiao Tong University (Philosophy and Social Science), Vol. 43(2022), No.5, P. 39-50.
- [14] J, Y. Shi, J. Zeng: Personal Information Protection Impact Assessment: Institutional Connotation and Improvement Path, Journal of Northwestern Polytechnical University (Social Sciences), Vol. 23 (2022), No. 4, P. 90-102.