ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

Generative Artificial Intelligence: Risks of Data Acquisition, Regulatory Deficiencies and Suggestions for Countermeasures—From the Inspiration of European Union Legislation

Huang Xinru[1,a]

[1]*School of Law, Zhongnan University of Economics and Law, Wuhan, China*
[a]*IrisHuang2000@163.com*

Abstract: *With the development of ChatGPT, generative artificial intelligence has become the focus of the times. The advancement of Generative Artificial Intelligence model starts from data acquisition, traceability and training, and the data acquisition's new characteristics of openness, invisibility and falsity will lead to unpredictable risks. Because of the inadequate AI legislative system, the lack of valid relations in the existing legal mechanism and the incompleteness of the existing legal provisions, it is difficult for the existing mechanism to prevent, control and solve the problems from Generative AI data acquisition. As far as the relevant regulatory measures concerned, the European Union legislation of Generative AI provides useful reference. Only by improving the associated mechanism between AI legislation and existing laws, unifying and coordinating the classification and grading protection mechanism for data, playing the role of the data protection institution, coordinating the cross-border flow of data, and strengthening international regulatory cooperation can a full-process regulatory system for data acquisition be established, thus achieving a balance between the safety control of Generative AI and innovative development.*

Keywords: *Generative Artificial Intelligence, Data Acquisition, EU Legislation, Data Regulation*

1. Introduction

With the development of large language models like ChatGPT, Generative AI has become the focus of the times, and will certainly bring a new round of revolutionary impact to the Internet, with far-reaching effects on people's daily life and productive ways. Behind the advantages of bridging the knowledge gap to a certain extent and intelligence and interactive application, the underlying risks from Generative AI should not be underestimated. The worldwide legal system around Generative AI and its in-depth synthesis technologies has not yet taken shape, although countries and regions have taken actions, such as the EU has drafted and attempted to pass the Artificial Intelligence Bill (draft), and China's State Council has issued the Interim Measures for the Administration of Generative Artificial Intelligence Services, but the contents of its stipulation takes AI as an overall object, and did not establish a effective association with the existing data security. The governance system established an effectively associated mechanism, making it difficult to control the data acquisition behavior of major developers when they train AI models.

As a pioneer in data protection, the EU will build a regulatory mechanism with the Artificial Intelligence Act as the core of AI legislation, combined with existing data protection laws such as the General Data Protection Regulation (GDPR), to provide favourable guidelines for defusing the risks of AI data acquisition and its solution countermeasures. With the unknown future of Generative AI, the law should act as a safety valve. Combining the EU's focus on personal data protection and China's emphasis on data sovereignty, building a full-process regulatory system from the perspective of data acquisition and maintaining the limited development of Generative AI may be a more appropriate approach to deal with the early stage of the AI technology revolution.

2. New Risks of Generative AI Data Acquisition

Based on the characteristics of the training data of Generative AI models, the personal data collected

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

can be mainly divided into three categories: public personal data, personal authorised data and unauthorised personal data. Combined with the characteristics of the data acquisition behaviour of Generative AI, the types of risks arising from it are more comprehensive and diversified, which broadly include illegal collection of data sources, uncontrolled algorithmic bias, difficulty in distinguishing the authenticity of the content, the proliferation of false information, the lack of privacy protection, infringement of intellectual property rights and other issues.[1]

2.1 Data Acquisition Behaviour in Generative Artificial Intelligence

Generative Artificial Intelligence (hereinafter referred to as "Generative AI") refers to the technology of generating content such as text, pictures, sounds, videos, codes, etc. Based on algorithms, models and rules, which can achieve better content generation effects by virtue of its technical capabilities of autonomous learning and self-optimisation;[1] Generative AI services refers to the services of providing crowd-generated content to the domestic public by using Generative AI technology to provide the domestic public with services for generating content like text, images, audio, video, and so on.[2] As a generic language model for Generative AI, the training process of ChatGPT involves the use of a large amount of text data. From the public disclosure, the GPT-3 language model is trained from 175 billion parameters, and the iterative GPT-3.5 language model obviously requires a much larger amount of data as support. According to OpenAI's documents, ChatGPT's data training goes through six cycles of the whole process. The details are included in Table 1.

Table 1: Data training process of ChatGPT.

Serial Number	Point	Main Behaviour
1	Data collection	The collection of raw data from various sources relies on two main sources of textual data, namely user input content and training databases.
2	Data preprocessing	Raw data is cleaned and standardized for subsequent processing and analysis.
3	Data Annotation	The data is labelled to provide training data for machine learning.
4	Feature Extraction	Extracting features from labelled data.
5	Model Training	Analysing and learning from training data.
6	Result Generation	Output Generator.

Same as general AI systems based on natural language processing, the working principle of ChatGPT can be divided into three stages: data input - machine learning - result output. As a language generation model, ChatGPT's training data usually comes from a large amount of textual data, including user input content and training database. Regarding the user input content, according to ChatGPT's Terms of Use, if the user does not agree to be used as training data, he/she can refuse the authorization by email or other means and it will not affect his/her normal use of the system.[2] With regard to the training database, the data sources can be broadly categorized into three types: the first is public domain data that is not privately owned and can be used and processed by anyone without restriction.[1] The second type is content that has been legally authorized by mutual consent, i.e. by signing a contract with the right holder and thus obtaining a valid authorization to use the relevant data and content legally. The third type is unauthorized information and content, which mainly refers to the fact that the data and content itself are objects protected by rights, but ChatGPT uses the relevant data without authorization, and the ways of obtaining network data by using "crawler" technology, illegally obtaining the content of databases, and illegally digitizing the content of non-electronic data.[3] Accordingly, ChatGPT's research, development and use of the training data without the authorization or knowledge of the users will result in a high risk of private infringement.

2.2 Risks to Personal Data from Generative AI

Generative AI creates risks at all stages of the training data process: at the input stage, Generative AI collects a large amount of publicly available data and personal data by crawling the data through the network; at the intermediate end, Generative AI acts as an embedded programme of other Apps or applications, and the process of obtaining personal data is often covert; at the output end, Generative AI's generated content is extremely malleable and causeless. Its potential use is not limited to the developer's

[1] Article 2 of the Interim Measures for the Management of Generative AI Services (Exposure Draft).
[2] Article 2 of the Interim Measures for the Management of Generative AI Services.

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

original intention, but may be applied by various ways of subjects.

2.2.1 Openness: Crawling Public Data

From ChatGPT to Baidu's Wenxin Yiyin, to various ways of Generative AI such as the latest version of GPT-4 released by OpenAI, there are more and more risks around Generative AI technology and economic benefits. In the past, the risky pattern related to personal data was mainly the misuse of personal data by data recipients like service providers, technology developers, etc., with the common ground being the unlawful and illegal processing of personal data entered by users themselves. Taking the application of the EU GDPR as an example, the necessity of personal data acquisition in the case of ICANN v. EPAG Contact Information Collection, the conditions of consent and the right to know in the case of ClearviewAI's processing of face recognition data, and the way of personal data acquisition in the case of Grindr's sharing of users' data with a third party, the focus of the disputes is on how to lawfully acquire and process the data that the individuals inputted on their own. The focus of the dispute is how to lawfully obtain and process the data entered by individuals themselves.

However, nowadays, the model training principle of Generative AI does not rely on individuals inputting data into an application, instead, personal data that is already in the public domain on major websites can be put to its use through web crawlers. In fact, in the current situation of serious personal data leakage, the public personal data on the network has even been enough for Generative AI - completely detailed restoration of personal situation and outline the user profile, while it can continue to carry out refined calculations, speculation and back to other unknown personal data. The abuse of user profiles by e-commerce companies such as Taobao and Meituan is commonplace, and we need to be vigilant that user profiles will become more accurate with the grafting of Generative AI and business models in the future, because publicly crawled data comes from the entire Internet, which means that platforms are holding more comprehensive, connected, and three-dimensional personal information, and therefore when Generative AI begins to be commercially available on a large scale, the characteristics of its openly crawled data will inevitably lead to more serious damage to personal privacy. The powerful creative ability makes Generative AI what could be called a "technological black box", and it's even more confused by the fact that developers such as OpenAI have not announced their technological principles, which will inevitably break the balance between technological innovation and security control, bringing unprecedented challenges to regulation.

2.2.2 Invisibility: Preservation and Misuse of Data

Until countries introduce strict data use regulations and data localization requirements, Generative AI will always be at serious risk of data exfiltration, which is reflected in the insidious nature of its data retention and misuse, and will be more imperceptible because it is embedded in the applications of software. The insidiousness is specifically reflected in the unawareness and unpredictability of users.[4]

On the one hand, the data preservation of Generative AI is hidden. The current problem of data exfiltration caused by Generative AI can no longer be underestimated. According to a survey by Cyberhaven, a cybersecurity company, at least 4% of employees from industry giants entered sensitive data from their own companies into ChatGPT, and sensitive data accounted for 11% of the input content. While this may seem like a small percentage, its statistics show that 0.9% of employees are responsible for 80% of data exfiltration incidents, while a user base of hundreds of millions of users means that 11% of inputs is actually an astronomical number.[5] In order to prevent the outflow of sensitive data, industry giants such as Microsoft, Verizon, Amazon, Walmart and Samsung have also restricted or banned the use of ChatGPT by their employees. Since 31 March this year, the Italian Data Protection Authority has temporarily banned ChatGPT for a month on the grounds of data privacy violations, while EU countries such as Italy, France, Germany, Spain and the EU itself have begun a flurry of investigating ChatGPT's data outflow problem. However, the issue of data outflow is difficult to trace back and impose controls because ChatGPT does not provide any source of response.[6]

On the other hand, data misuse in Generative AI is insidious. In fact, OpenAI companies have been defaulting to using user inputs as training data, saving and utilizing a large amount of sensitive data. Take Amazon.com as an example, it disabled ChatGPT because it found that the content generated by ChatGPT was highly similar to its confidential data, which is obviously that ChatGPT received confidential data input by employees or confidential data obtained by calculating based on ordinary data input by employees, and the consequences of saving this kind of sensitive data are unimaginable. Even if OpenAI doesn't misuse the data, third parties can bypass the filtering mechanism and access the sensitive data by asking clever questions; even if OpenAI was forced to add the option "Prohibit AI training with user data" to ChatGPT, the risk of leakage is extremely high, and it's not a good idea to keep this kind of data. Even though OpenAI was forced to add an option to ChatGPT on 25 April to

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

"prohibit the use of user data to train AI", the risk of such data leakage is extremely high, and the consequences are endless. In the case of Prismgate, OpenAI was even coerced into providing user data.

2.2.3 Falsity: Unpredictable Use of Data

Compared to previous AI technologies, the most significant feature of Generative AI is the generation of content, which, while cost-effectively broadening the boundaries of personal knowledge, presents the risk of generating false content, and the use of the generated false content exposes everyone to risk without any obstacles. Crucially, the neutrality of generated content and the universality of generic models essentially make its use unpredictable, increasing efficiency while providing many malicious users with the tools to break the law and commit crimes.[7]

The instantaneous and unsourced nature of Generative AI's responses may lead users to believe false information that appears reasonable and authoritative, especially when there is a lack of training in media literacy, and Generative AI's algorithms can be designed to target vulnerable individuals so that the generated content resonates with their personal beliefs and sentiments, thus becoming a potent tool for destabilizing societies and spreading false narratives, which would undoubtedly put personal privacy, individual rights and personal security at risk, and even lead to serious situations such as social crime.[8] For example, in addition to taking advantage of the general sense of trust between relatives, friends, and loved ones, the recent "AI scam" is also a scam specifically targeting widows and orphans, which takes advantage of the fact that such elderly people know little about the Internet and care about their children who are working outside the home to carry out precise scams on such elderly people, through Generative AI to collect personal information and process it into "AI". AI collects personal information and processes "face-swapping" and dubbing, imitating the image, voice and language logic of children by phone or video, making the other party believe that it is true, etc., or using Generative AI to output biased ideologies to achieve the purpose of disseminating false information.

3. Deficiencies in the Existing Regulation of Generative AI Data

The new characteristics and corresponding risks of Generative AI data acquisition behaviour have not attracted special attention, and there are still many deficiencies in the existing legal mechanism and its specific provisions. From the imperfection of the AI legislative system at the macro level, to the lack of clarity of the existing legal articulation mechanism at the meso level, to the incompleteness of the specific provisions of the existing laws at the micro level, the three constitute concentric circles of risk from the outside to the inside, which makes it difficult to effectively control and constrain the risks of data acquisition.

3.1 Incomplete Legislative System for AI

At the macro level, China's legislative work in the field of AI is progressing steadily, but no special attention has been paid to the relationship between the field of AI and data security.2023 The Interim Measures for the Administration of Generative Artificial Intelligence Services was passed on 10 July and came into force in mid-August, and will form part of the legal regulatory system of AI together with the Cybersecurity Law, Data Security Law, Personal Information Protection Law, and Science and Technology Progress Law. Together, they form part of the legal regulatory system for AI. On the one hand, having a law to follow does not mean having good effectiveness. There is a competition between the Data Security Law and the Personal Information Protection Law in terms of the object of regulation, and at the same time, Generative AI acquires a wider range of personal data, which may result in the difference between "what law is applicable" and "not available" in the regulation of Generative AI's data acquisition. In regulating the data acquisition, there may be the problems of "which law to apply" and "unavailability".[9] On the other hand, the unsound legislative system makes the regulation of Generative AI slow. The Interim Measures for the Administration of Generative Artificial Intelligence Services is the first legal provision for Generative AI and the first governance framework in China in the field of AI. However, with the volume of rules in 24 articles and the level of effectiveness of departmental regulations, it is difficult to keep pace with the development of the entire field of AI, which may result in a passive situation in the regulation of AI by law.

In contrast, the EU began to embark on the path of regulating AI as early as 7 years ago. In 2016, the EU Legal Affairs Committee issued the EU Civil Law Rules on Robotics, which put forward regulatory requirements for AI-based controlled robots; in 2018, the European Commission issued the EU Artificial Intelligence, which put forward a "human-centred " development path for AI; from 2018 to 2023, the EU

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

and others issued a total of 31 validity documents for AI regulation, which were basically updated at a rate of 4-5 documents.[10] In 2023, the EU's Artificial Intelligence Bill and its draft is an even bigger step towards regulating AI, exploring the balance between innovation and safety in the form of comprehensive legislation balance between innovation and safety. The Artificial Intelligence Bill has a number of innovative highlights, including provisions on risk classification and hierarchical regulation and hierarchical AI lists, provisions on general-purpose AI, algorithmic transparency requirements and safety assessments, and an innovative regulatory sandbox system, presenting a legislative trend towards enhanced personal data protection. Although the AI Bill has been criticised for flaws in the liability mechanism and other criticisms, the EU has been at the forefront of the world in the field of digital legislation, and still provides a certain reference for other countries.

Meanwhile, the General Data Protection Regulation (hereinafter referred to as GDPR) introduced by the EU in 2018 is known as the world's most stringent personal data protection legislation, which is famous for its broad scope of protection for personal data, extremely high implementation standards and harsh regulatory penalties, and provides a legislative model for other countries to follow. It provides a model for personal data protection legislation for other countries. In recent years, the EU has issued the Digital Service Act, the Digital Market Act and the Digital Governance Act on data governance. These Acts, together with the GDPR and the soon-to-be-adopted Artificial Intelligence Act, will constitute important rules for the regulation of AI under the framework of the EU's data strategy. [11] They not only aim to improve the system, strengthen the security protection of data, promote the flow of data in Europe, and prevent the potential risks of automated algorithmic decision-making; on the other hand, but also safeguard the rights of individuals, establish the relevant ethical and value standards, and build a balanced mechanism between regulation and innovative development.

3.2 Lack of Relations in the Existing Legal Mechanisms

At the meso level, China's AI regulatory framework adopts a subject-based governance paradigm, with the implementation of the main body's responsibility as the starting point when targeting different Internet information services, but there is no discernment of the similarities and differences in the concepts of the main body of the law and the establishment of a clear system of allocating the responsibility among the current laws. Regarding the main body of regulatory responsibility, laws such as the Network Security Law, the Data Security Law and the Personal Information Protection Law stipulate different concepts of the main body and its rights and obligations for their respective scopes of application.[12] However, in the context of comprehensive governance of AI, the incompatibility of the concepts of the main body of regulatory responsibility may lead to the uncertainty of the main body of responsibility, and it is easier to shift the blame to each other, and to unclearly assign rights and responsibilities. For the specific obligations of each subject, if the definition of responsible parties is too simplified, all relevant parties in the AI supply chain may be required to assume the same obligations, causing internal difficulties in allocating responsibilities. Although the Interim Measures for the Administration of Generative Artificial Intelligence Services, which will soon come into effect, clearly stipulates the obligations of Generative AI service providers to a certain extent, it does not completely solve the problem of simplistic and incomplete definition of the subject of responsibility. At present, it is only generally stipulated that the "relevant competent authorities in accordance with their duties" or the "national net information department" shall deal with those who violate the management measures, but it has not yet required the establishment of a special AI management organization or the coordination of the division of powers and responsibilities with the existing management organizations.[13]

Comparatively speaking, the EU AI Bill first establishes a risk-based AI governance framework based on the governance paradigm of risk. By assessing AI systems, AI systems will be classified into four tiers of unacceptable risk, high risk, limited risk and minimal risk, matched with different liability measures and differentiated regulation, which is conducive to the subsequent concretisation of regulation.[14] Under the AI systems with different risks, the draft AI Bill 2023 further defines what kind of obligations should be undertaken by various types of operators such as providers, authorised representatives, distributors, importers, deployers, etc. specifically. For one thing, the risk governance framework is conducive to AI system developers to self-suppress some of the risks, and then take different regulatory measures according to different levels of risk, including pre-market "regulatory sandbox" risk assessment, CE marking, assessment by regulators, market supervision, establishment of databases, and stringent enforcement and punishment, etc., so as to achieve all-round control of the entire chain of AI.[15] For another thing, for some AI systems across application scenarios, the AI Draft Law 2023 refers to AI models trained on a large scale on a wide range of data as "basic models", and specifically stipulates the relevant obligations for such general-purpose AI models, which is different from the obligations for AI

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

systems that use risk as a classification criterion, and effectively addresses the issue of generating AI systems that can be classified into different categories. This effectively solves the problem that generic models of Generative AI cannot be included in the risk governance framework. Distilling the commonality of AI governance as the content of the provisions of the general AI legislation, meanwhile providing separate provisions for special types of AI, the EU's idea of legislative frameworks is both horizontal and vertical, and the idea of governance of commonalities and characteristics, will help to solve this problem.

3.3 Inadequacy of Specific Provisions of Existing Laws

At the micro level, personal data protection under Generative AI is currently dominated by generalized general provisions. The Interim Measures for the Administration of Generative AI Services need to be implemented in conjunction with the provisions of the Cyber-security Law, the Data Security Law, the Personal Information Protection Law, and other laws and administrative regulations, however, there are a number of as yet unclarified issues between these three laws.

On the one hand, in response to the mixed use of the concepts of "information" and "data" in China's existing legislation, there is an overlap in the scope of the relevant laws regulating Generative AI, which can lead to conflicts between different laws in terms of subject-matter jurisdiction, both positive and negative. For example, "data" appears 16 times and "information" 105 times in the Cyber-security Law, and although the Personal Information Protection Law and the Data Security Law do not mix "information" and "data", they do not use "information" and "data" in the same way. Although the Personal Information Protection Law and the Data Security Law do not mix the terms "information" and "data," they also fail to analyse the relationship between their respective categories and establish a grading mechanism around their core concepts: the Personal Information Protection Law establishes a grading system for information, categorising personal information into general information and sensitive information, while the Data Security Law categorises data into core data, important data, and general data, and provides for the establishment of a data classification and grading system by the state and the formulation of an important data catalogue by the competent authorities.[16]

On the other hand, China lacks a complete regulatory mechanism for personal data protection. Compared to China, where law enforcement agencies at all levels can apply the Data Security Law and the Personal Information Protection Law, the EU has not only established the European Data Protection Commission under the GDPR and established a connectivity mechanism with national data regulatory agencies, but also made more detailed provisions on setting penalty levels and circumstances, which provide the EU and national administrative and law enforcement agencies with clearer and more precise application standards in the performance of their functions. Clearer and more precise application standards for the EU and national administrative enforcement agencies to fulfil their functions. Meanwhile, China's legislation has more often strengthened the regulation of large enterprises, and has not yet formulated special provisions for small and medium-sized service providers. Approaches to algorithmic recommendation, deep synthesis and Generative AI in the field of artificial intelligence have also failed to set out specific regulations for small and medium-sized service providers in order to control their compliance costs. In contrast, EU data legislation has always been aware of the huge compliance costs faced by SMEs, and from the GDPR to the Digital Services Act to the draft Artificial Functions Bill, measures such as regulatory sandboxes have been adopted to moderately safeguard SMEs' space for technological innovation through specific Article 28(a) and Article 53(a). In conclusion, the draft AI Law takes into full consideration the needs of SMEs in formulating the relevant guidelines in order to safeguard the fundamental interests and economic viability of SMEs.

4. Regulatory System for Generative AI's Data Acquisition

In order to effectively control the development of Generative AI and make it do something and not do something, the only way is to control the training of large-scale language models from the source of acquiring data, to prevent Generative AI from being abused by full-coverage control, and to establish a full-process regulatory system by improving the articulation mechanism between AI legislation and the existing laws, coordinating the mechanism of hierarchical classification management of personal data, playing the role of the data protection management agency, and coordinating the cross-border flow of data and international cooperation, and other measures.

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

4.1 Improvement of the Relevant Mechanism between AI Legislation and Existing Laws

The articulation of AI legislation with existing data protection laws, especially the personal data protection law regime, should be done properly. Firstly, to carry out unified AI legislation, it is always impossible to avoid the issue of the legality of personal data processing.[17] The research and development and deployment of AI for use cannot be separated from the acquisition of personal data and personal information, and the provisions of the Personal Information Protection Law may create certain compliance barriers to the above activities. Among the legality of the processing of various types of personal information as stipulated in Article 13 of the Personal Information Protection Law, only "individual consent" and "necessary for the performance of a contract" can be used as the basis for the legality of the acquisition of personal information using AI; however, according to the Sixth Measures for the Administration of Generative Artificial Intelligence, the legality of the acquisition of personal information may be hindered by the provisions of the Personal Information Protection Law. Article 6 of the Measures for the Administration of Generative Artificial Intelligence "promotes the construction of Generative AI infrastructure and public training resource platforms ..... to expand high-quality public training data resources", and if a public training data resource base is to be established, it is required to establish a public training data resource base as soon as possible. Individual consent specification process for AI access to personal data.

For the consideration of legislative cost and feasibility, the applicability of existing laws such as the Personal Information Protection Act can be stipulated in the AI legislation, and Article 13, paragraph 1, item 7 of the Personal Information Protection Act, "Other cases stipulated by laws and administrative regulations", provides the legitimacy basis for the handling of personal information by AI systems. At the same time, the revision process of the Personal Information Protection Law should be promoted to specifically regulate the special cases of AI acquiring personal data, and to clarify the relationship between personal data in AI training data and personal information in the Personal Information Protection Law.[18] The EU Artificial Intelligence Act does not exclude the application of the GDPR in general, and the GDPR has additional bases of lawfulness that can be asserted by processors of personal data, such as "the legitimate interests of the controller and other third parties" and "processing in the public interest, scientific or historical research or statistical purposes". The draft Artificial Intelligence Bill 2023, without explicitly affecting the implementation of the GDPR, explains and articulates the processing of personal data in specific provisions.[19]

4.2 Harmonization of Classification and Protection Mechanisms for Coordinated Data

In order to effectively control the acquisition of personal data by AI, it is necessary to establish a unified data classification and grading protection system. Data classification and grading is a prerequisite for controlling the training of Generative AI and achieving data security governance. Article 21 of China's Data Security Law stipulates that "the state establishes a data classification and grading protection system ..... The National Coordination Mechanism for Data Security Work coordinates the formulation of important data records by relevant departments". Article 6 of the Measures for the Management of Generative AI, which will soon come into effect, "Promote the orderly opening of public data classification and grading, and expand high-quality public training data resources" also illustrates the necessity and urgency of data classification and grading. However, the specific data classification and grading protection system and important data catalogue have not yet taken shape, and practices vary greatly from place to place, so the data classification and grading methods stipulated in the Data Security Law, the Personal Information Protection Law and the Measures for the Administration of Generative AI should be coordinated, and a coherent and consistent important data catalogue should be introduced as soon as possible.[20] Data can be classified from the value, sensitivity and impact of the data, and data can be classified from the source, content and use of the data, and corresponding laws can be introduced to regulate the classification, labelling, evaluation and protection of government data. At the same time, in the process of data classification, a data classification engine is established to achieve real-time, automatic and accurate classification.[21]

On the one hand, as far as data classification is concerned, China can establish three sensitive data levels according to the degree of sensitivity: low-sensitive data, medium-sensitive data and high-sensitive data, and restrict the access of Generative AI to personal data for data of different sensitivity levels, for example, access to low-sensitive data is only required to be registered for the record, access to medium-sensitive data should be subjected to a general security test and filed for record, and access to requests for highly sensitive data will For requests for highly sensitive data, detailed and careful approval procedures are required. Other characteristics, such as the value of the data or the scope of the impact,

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

can also be used to classify the data. Inspired by the GDPR, the EU GDPR provides a mechanism for the processing of special types of personal data that are distinct from ordinary data processing. Currently, personal sensitive data includes specific personal information, genetic data, biometric data and health data, and other private data.[3] Personal sensitive data under the GDPR can correspond to the tier of highly sensitive data. On the other hand, as far as the classification of data is concerned, China can classify different types of data according to the relationship between data source, data content, and data use, and include different types of data in different levels of sensitivity, such as low-sensitive, moderately-sensitive, and highly-sensitive, etc. For example, data related to critical infrastructure, remote biometric systems, civic education, the safety components of products, civic employment, and public services are included in high-risk data. Inspired by the categorization of AI systems in the EU's (draft) Artificial Intelligence Act, data can be classified into four classes of unacceptable risk, high risk, limited risk and minimal risk, according to the level of risk posed to fundamental rights.[22]

4.3 Establishment and Functioning of a Data Protection Authority

If we are to establish a national coordination mechanism for data security and strengthen the acquisition, analysis, research and early warning of data security risks, we must integrate the governance of data acquisition by Generative AI into the same organization, or establish a coordination mechanism between two different organizations, with a clear delineation of powers and responsibilities. At present, unclear powers and responsibilities, intersecting scopes and lack of supervision are more prominent in China's data management organizations, which should integrate data management organizations, clarify the scope of powers and responsibilities, and implement existing regulatory matters. The EU GDPR devotes the entire Chapter 6 to detailing independent supervisory authorities and their related obligations, stating that "each Member State shall arrange for one or more independent public authorities to be responsible for overseeing the implementation of this Regulation, for protecting the fundamental rights and freedoms of natural persons involved in the processing of personal data and for promoting the free movement of personal data within the Union", and national data supervisory authorities are not clear on the main regulatory authorities mentioned in the Law on Data Security. ", national data supervisory authorities have also proven in practice to play an effective enforcement role[4] by investigating, issuing reports and fining companies for GDPR violations. Currently, the draft EU Artificial Intelligence Bill 2023 seeks to establish a unified regulator to achieve harmonization of data governance and numeracy laws and regulations under the convergence with the GDPR.

In response to the specific obligations of data management agencies, on the one hand, the formation of a security monitoring platform for Generative AI crawling data. Using technologies such as the Internet of Things, cloud computing and cloud storage, it accesses all kinds of facilities and equipment, data assets as well as resources, adopts automation, visualization and real-time tracking and management of data flow, and realizes joint prevention and control of security risks in the flow of data in the storage, use, transmission, sharing and other flow links, with a view to dynamic monitoring, situational assessment and rapid response. On the other hand, the risk control measures of data management organizations should focus on the whole process management before, during and after the event. At present, there are still institutional gaps in China's AI risk control measures, and consideration can be given to establishing a "before-the-fact-after-the-fact" full-chain regulatory mechanism, and formulating AI risk control frameworks applicable to China: in the before-the-fact stage, according to the different application scenarios and data risk levels, the data management organizations should be able to control and prevent AI risks in a timely manner. In the ex-ante stage, according to the different application scenarios and data risk levels, corresponding regulatory measures will be formulated. During the interim stage, try to learn from the EU's regulatory sandbox mechanism to help regulators achieve an understanding of the full picture of the AI system and collect relevant information, and to help providers assess the real-life operation of the AI system.[5] The regulatory sandbox system provided for in Article 54 of the EU's draft Artificial Intelligence Bill of 2023 provides an additional basis of legitimacy for the processing of personal data. In the AI regulatory sandbox, personal data lawfully collected for other purposes may be processed solely for the purpose of developing and testing certain AI systems in the sandbox". In the ex-post phase, providers of high-risk AI systems should have a monitoring plan and

[3] See article 9(1) of the GDPR.

[4] The ICANN v. EPAG contact information collection case, ClearviewAI processing of face recognition data, and Grindr sharing of user data with third parties.etc.

[5] Regulatory sandbox is an important measure used in the Artificial Intelligence Bill to ensure compliance of AI systems, which is controlled environments established by public bodies to facilitate the safe development, testing and validation of innovative AI systems for a limited period of time before they are brought to market or put into use under a specific programme overseen by regulators.

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

promptly report breaches or seriously malfunctioning systems to the regulator.

4.4 Coordination of Cross-border Data Flows and AI International Regulatory Cooperation

Against the backdrop of the severe cyberspace and data sovereignty dispute[23], China's top priority in the field of digital governance is to establish a complete data security governance system. Based on the characteristics of access to massive data and high-speed computing, Generative AI will further promote the cross-border flow of data, and how to govern cross-border data flow and safeguard China's data sovereignty has become the focus of regulating Generative AI.

On the one hand, reference can be made to the cross-border data flow system of the European Union, which can be used to strengthen management internally and coordinate and cooperate externally. Internally, from the standpoint of national data sovereignty and protection of personal data, we should optimize the relevant regulations on personal data and non-personal data, put the right to know, the right to consent and other personal data rights into practice, improve the governance system of cross-border data flow, carry out real-time monitoring and policy guidance for multinational enterprises, especially include the data acquisition of Generative AI in the scope of regulation, and stipulate the corresponding authorization and restriction measures, so as to make it open and transparent. Externally, China should increase its voice in cross-border data circulation, advocate the concept of "Digital community of shared future", promote the integration of the domestic system into the international data framework, and enter into bilateral or multilateral cooperation with other countries, so that the international data standards formulated can better represent the sovereignty and interests of most countries and regions, and form a reasonable, fair and balanced multi-party sovereignty system.[24]

On the other hand, with the EU's international regulatory cooperation programme on AI, there is more room for cooperation between the EU, China and the US on AI governance. For example, the transatlantic cooperation blueprint proposed by the EU to formulate cooperation agreements on AI, seek convergence of relevant systems, and then allow multinational high-tech companies to conduct conformity assessment of AI systems under the joint supervision of national governments, while a mutual recognition mechanism for conformity assessment results can be established. When formulating rules for the AI regulation, specialized agencies of various countries can refer to each other's regulations and promote the formation of a consensus through intergovernmental dialogue and consultation, thus saving the cost of cross-border AI regulation and improving its effectiveness.

5. Conclusions

Just as the steam-engine-powered automobile first developed, no one thought that the bulky, dangerous, and costly vehicle would one day fully replace the manpower-based horse-drawn carriage, so too it is difficult to predict the future impact, good or bad, of Generative AI in the midst of change. People are divided between those who believe that Generative AI will greatly help improve human productivity and those who believe that its impact on human subjectivity will pose a global security risk. In the compound context of massive data, complex algorithms and powerful arithmetic, AI is developing at an unprecedented pace, with a wide range of applications and governance risks. Compared to the advantages of private developers in mastering algorithms and computing power, a more effective and fundamental way to effectively control AI is for the state to control access to massive data, and to balance the development needs of AI with the governance of security risks through data acquisition. Only by combining the EU's position on data protection and China's data sovereignty, drawing on the EU's useful experience in data protection, and improving the whole-process regulatory system of Generative AI Data Acquisition, can we realize the development of Generative AI in a stable and far-reaching way.

References

[1] Sun Qi, Research on the Legal Issues of Regulating the Providers of Generative Artificial Intelligence Products[J]. Politics and Law, 2023, Vol. 7: 162-176.
[2] OpenAI, Introducing ChatGPT. https://www.openai.com/blog/chatgpt. Accessed 22 Apr. 2023.
[3] Bi Wenxuan, The Risk Regulation Dilemma of Generative Artificial Intelligence and Its Resolution: Taking the Regulation of ChatGPT as a Perspective[J]. Comparative Law Studies, 2023, Vol. 3: 155-172.
[4] Chen Yuefeng, Beyond Data Boundary Rights: The Dual Public Law Construction of Data Processing[J]. Journal of East China University of Political Science and Law, 2022, Vol. 1: pp. 18-31.
[5] Xie Yongjiang and Yang Yongxing, ChatGPT Legal Risks and Their Regulation[J]. Journal of

ISSN 2706-6827 Vol. 6, Issue 6: 61-70, DOI: 10.25236/IJFS.2024.060611

Nanjing University of Posts and Telecommunications (Social Science Edition), 2023, Vol. 5: pp. 1-9.
[6] Liu Shuang and Zhang Xiaoyue,"Research on the Legal Protection and Regulation of Generative Artificial Intelligence Data Risks - Taking ChatGPT's Potential Data Risks as an Example[J]. Journal of Guizhou University (Social Science Edition), 2023, Vol. 5: 87-97.
[7] Zhu Jiajun, Challenges and Responses to the Regulation of False and Harmful Information of Generative Artificial Intelligence - Cited by the Application of ChatGPT[J]. Comparative Law Studies, 2023, Vol. 5: 34-54.
[8] Liu Yanhong, Three Major Security Risks and Legal Regulations of Generative Artificial Intelligence - Taking ChatGPT as an Example[J]. Oriental Jurisprudence, 2023, Vol. 4: 29-43.
[9] Guo Xiaodong, The Risks of Generative Artificial Intelligence and Its Inclusive Legal Governance[J]. Journal of Beijing Institute of Technology, 2023, Vol. 6: 93-105.
[10] Guo Jiannan, A Study on the Policy, Ethical Guidelines and Regulatory Path of Artificial Intelligence in the European Union[J]. Internet World, 2023, Vol. 1: 26-32.
[11] Chen Bing, Rule of Law Considerations and Practical Framework for Promoting the Normative Development of Generative Artificial Intelligence--An Appraisal of the Relevant Provisions of the Interim Measures for the Administration of Generative Artificial Intelligence Services[J]. China Applied Law, 2023, Vol. 4: 108-125.
[12] Zhai Zhiyong, Data Security Law[J]. Journal of Suzhou University (Philosophy and Social Science Edition), 2021, Vol. 1: 73-83.
[13] Xu Jimin, Generative Artificial Intelligence Governance Principles and Legal Strategies[J]. Theory and Reform, 2023, Vol. 5: 72-83.
[14] Xin Zhang, Algorithmic Governance Challenges of Generative Artificial Intelligence and Governance-Based Regulation[J]. Modern Law, 2023, Vol. 3: 108-123.
[15] Zhang Linghan, Logic Update and System Iteration of Deep Synthesis Governance - The Chinese Path of Generative Artificial Intelligence Governance such as ChatGPT[J]. Legal Science, 2023, Vol. 3: 38-51.
[16] Xia Han, The Market Regulation Shift in EU Legislation Regulating Cross-Border Data Flows and Implications for China[J]. Hebei Law, 2023, Vol. 8: 169-182.
[17] Rongrong Zhu, The Challenge of ChatGPT-like Generative Artificial Intelligence on Personal Information Protection and Response[J]. Journal of Chongqing University (Social Science Edition), 21 September 2023, pp. 1-14.
[18] Dazhi Wang and Ting Zhang, Risks, Dilemmas and Countermeasures: Personal Information Security Challenges and Legal Regulation Brought about by Generative Artificial Intelligence[J]. Journal of Kunming University of Science and Technology (Social Science Edition), 2023, Vol. 5: 8-17.
[19] Cheng Le, Legal Regulation of Generative Artificial Intelligence - Taking ChatGPT as a Perspective[J]. Political Law Series, 2023, Vol. 4: 69-80.
[20] Cheng Xiao, On the Obligation of Data Security Protection[J]. Comparative Law Studies, 2023, Vol. 2: 60-73.
[21] Haiyan Peng and Zhen He, Research on the Realistic Dilemma and Response Strategy of Government Data Security Governance in the Context of Artificial Intelligence[J]. Yunnan Social Science, 2022, Vol 3: 29-37.
[22] Chen Bing and Dong Siyan, Algorithmic Risks of Generative Artificial Intelligence and the Basis of Governance[J]. Learning and Practice, 2023, Vol. 10: 22-31.
[23] Gong Yunmu, The Return of the Concept of Sovereignty and EU Digital Governance in the Digital Age[J]. European Studies, 2022, Vol. 3: 18-48.
[24] Xu Jimin, Generative Artificial Intelligence Governance Principles and Legal Strategies[J]. Theory and Reform, 2023, Vol.5: 72-83.