The Current Status, Problems, and Policy Recommendations of AI Data Utilization in China

Runkai Chen
School of Law, University of International Business and Economics, Beijing, China
***Corresponding author, e-mail: chenrunkaichn@foxmail.com**

Abstract
The study delves into the status quo, challenges, and policy suggestions concerning the application of artificial intelligence (AI) data in China. It emphasizes the pivotal role of data utilization rules in the progression of AI and pinpoints significant issues in China's present AI data utilization. These include the insufficiency of high-quality training data, the absence of effective data privacy protection mechanisms, and the lack of flexible data circulation mechanisms. To tackle these issues, a series of policy recommendations are proposed. These encompass the creation of scientifically rational AI data utilization rules, the construction of a dual governance system integrating "institution + technology", the development of data utilization rules that harmonize the interests of various parties across different fields and scenarios, and the acceleration of the enhancement of laws and regulations related to data utilization. These suggestions aim to foster the growth of China's AI industry while ensuring the secure and compliant use of data.

Keywords
artificial intelligence; data utilization; data security; data governance

1.Introduction

Artificial Intelligence (AI) has emerged as a transformative force in the global economy, and its progress is heavily reliant on the effective utilization of data. In China, the rapid development of AI has brought forth both opportunities and challenges, particularly in the realm of data utilization. This paper aims to critically examine the current state of AI data utilization in China, identify the key problems hindering its advancement, and propose policy recommendations to address these challenges. By doing so, it seeks to contribute to the advancement of China's AI industry, ensuring that it progresses in a manner that is both innovative and compliant with global data security and privacy standards.

2.The Importance of Data Utilization Rules for AI Development

Data stands as a primary pillar among the triad that constitutes artificial intelligence, representing an essential resource for the AI sector. The reliance of contemporary AI technologies on extensive datasets is unparalleled. Utilizing methods like deep learning, AI systems can distill valuable insights and intricate logical patterns from raw data, with the proliferation of data directly correlating to enhanced deep learning efficacy. Such advancements are most evident during the development phase of algorithms, where access to large-scale datasets can address complex issues and markedly boost algorithmic performance.

The quality and volume of data directly feed into the sophistication and innovation of AI models. In the cur-

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

rent technological era, the quest for data has become increasingly voracious. Leading technology firms, such as Google and Meta, now train their models on more than a trillion words, reflecting the scale of data required for contemporary AI operations. To secure an edge, corporations are aggressively seeking diverse and high-caliber data through varied means, including partnerships for multimedia data, the creation of synthetic data via algorithms, and user data analytics for model refinement. These efforts reflect the evolving landscape and creative approaches in data sourcing.

The last ten years have seen AI burgeon, propelled by the boom in mobile internet and the Internet of Things (IoT), both of which provide a continuous stream of real-time data that fuels deep learning algorithms with abundant training samples. The ready availability of such vast and dynamic data repositories is crucial for ongoing AI breakthroughs. Consequently, establishing fair and effective data utilization rules is a strategic imperative to underpin AI's sustained advancement (Cheng, 2022).

3.Main Problems in the Current Utilization of AI Data in China

Despite the rapid development of the AI industry in China, there are still some deficiencies and problems in data utilization rules, mainly manifested in the following aspects:

3.1Insufficient Supply of High-Quality Training Data

Firstly, the data trading market in China is nascent, and ambiguity surrounds the legal risks and compliance costs associated with data transactions. The complexity of data ownership, the relativity of its value, and the variability of content over time complicate the establishment of clear definitions and evaluation standards, hindering agreement on price and transaction terms. Secondly, the industrial distribution of data resources is strikingly lopsided. While certain sectors are data-rich, this wealth is not evenly spread. Large enterprises typically guard their data as a vital competitive asset, which sets daunting barriers for smaller entities (Sun, 2023). Thirdly, the limited openness of public data. Despite its acknowledged importance, is restricted in practice. Local governments and agencies often treat such data as state assets and impose stringent usage conditions.

3.2Lack of Effective Data Privacy Protection Mechanisms, Leading to Increased Risks of Personal Information Leakage and Misuse

Personal data encompasses information that, individually or in conjunction, can identify a person or reflect their activities. Despite the rapid growth of data-centric industries, China faces significant challenges in establishing a robust personal information protection framework. The repercussions are substantial, with increased vulnerability to privacy breaches and data misuse.Firstly, the lag and dispersion of laws and regulations: The existing personal information protection measures show obvious lag in the formulation of laws and regulations. China's regulations on personal information protection are not perfect, and there are overlaps or contradictions between various laws and regulations, lacking unity and coordination. The principled nature of laws and regulations and universal requirements have not been effectively applied to all emerging data processing scenarios, making it difficult to form effective legal constraints in actual operations. Secondly, the risks in sensitive data handling: AI's evolution has spurred the aggregation of vast amounts of sensitive personal data, such as identity details, consumer behaviors, and location tracking. The lack of precise regulations and oversight mechanisms exacerbates the risks associated with data misuse and expo-

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

sure.Thirdly, data security weaknesses: The integrity of personal information is compromised at various stages—collection, storage, transfer, and processing—due to the absence of uniform security standards and enforcement measures. This deficiency leaves data susceptible to unauthorized breaches and cyber threats (Chen & Dong, 2023). Fourthly, the lack of awareness of personal privacy rights: In China, the public's understanding of personal data rights is relatively weak, which makes users often lack the necessary right to know and choice when data is collected and used. In addition, when enterprises and institutions collect user data, they usually fail to provide sufficiently transparent information processing procedures and sufficiently clear data usage explanations, further exacerbating the lack of personal information protection. Fifthly, the discrepancy between regulation and technology: While strides have been made towards better data protection, the alignment of regulatory mechanisms with technological advancements remains elusive. The pace at which data security technology evolves is not matched by corresponding updates in policy, leaving existing regulations inadequate for addressing new and emerging risks.

3.3Lack of Flexible Data Circulation Mechanism, Leading to Low Efficiency of Data Resource Development and Utilization

The flexibility of data circulation is crucial for promoting the full development and efficient utilization of data resources. In China, the mechanism of data circulation is still insufficient, which to a certain extent hinders the full exploitation of data resource potential and the rapid innovation of the AI industry. Specifically: Firstly, the circulation barriers under regulatory restrictions: The current legal and regulatory restrictions on data circulation are mainly reflected in the strict requirements for data processing principles, such as data minimization, informed consent, and de-identification processing principles. These principles play a key role in protecting personal privacy and data security, but in actual implementation, they may become overly conservative in interpretation and application, making the process of data sharing and exchange exceptionally complex and stringent. For example, overemphasis on the data minimization principle may limit the demand for data integrity, thereby affecting the accuracy and depth of data analysis. Secondly, the technical obstacles to circulation efficiency: Technical challenges are also an important factor leading to low efficiency of data circulation. The lack of uniform data formats, standards, and the incompatibility of data processing systems all become obstacles in the process of data flow. At the same time, the lack of efficient data processing and analysis tools means that a large amount of data cannot be quickly converted into usable information after collection, delaying the actual application and innovation cycle of data. Thirdly, the lack of market mechanisms and social foundations: In terms of incentives and support for data circulation, China's current policies have not yet formed an effective market-driven mechanism. The imperfect data value evaluation system makes it difficult to accurately evaluate and recognize the potential value of data, which in turn affects the trading of data rights. In addition, the lack of data trading platforms and data credit systems also makes the trust foundation of data circulation unstable, limiting the scale and depth of data exchange and cooperation (Dong & Zhang, 2023). Fourthly, the disconnect between regulatory policies and practices: There is a disconnect between regulatory policies and data circulation practices, leading to differences in policy formulation and implementation effects. The understanding and support of data circulation at the regulatory level often lags behind the development of the market and technology, lacking forward-looking policy design and flexible implementation strategies, which hinders the efficient use of data resources (Xiang, 2023).

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

4.Policy Recommendations for Establishing Scientific and Reasonable AI Data Utilization Rules

Perfecting the rules for AI data utilization is a systematic project, involving considerations of technological innovation, data security, and social development. Based on the common contribution to technology and social welfare, this project requires policymakers to find a balance between innovation and governance, promoting the coordination of technological development and social regulation through flexible policy mechanisms. Policy recommendations should be based on in-depth analysis of technological trends and practical needs assessment, achieving dynamic adjustment and continuous optimization, forming a living policy environment that adapts to technological evolution.

4.1Strengthen the Supply of High-Quality Training Data to Improve the Level of AI Technology

High-quality training data is the core resource for promoting the progress of AI technology. Therefore, to optimize the performance and accuracy of AI algorithms, it is crucial to strengthen the supply of high-quality training data from multiple aspects. Specific strategies can be developed from the following directions:
Firstly, improve the data trading ecosystem: By establishing and improving the data trading mechanism, the efficiency of data supply and demand matching can be improved. To this end, it is necessary to accelerate the construction of a comprehensive data element market, including but not limited to the development of on- and off-exchange trading platforms, and promoting the standardized use of data through data services and collective authorization models. In addition, by building a unified data trading platform, the trading process can be standardized, trading standards can be set, and trading friction can be reduced. Introducing third-party evaluation agencies and establishing a scientific and reasonable data value evaluation system can provide market participants with reasonable judgments of price and value, ensuring the fairness and transparency of transactions (Wang & Ma, 2022).

Secondly, promote data openness and sharing: Governments and public institutions should promote the graded classification openness of data resources to improve the efficiency of data resource management and utilization. By establishing a complete public data resource catalog and list system, public data can be effectively classified and managed, promoting the openness and sharing of more high-value data resources. At the same time, the personal information security and data compliance system should be improved, detailed data processing rules and anonymization standards should be formulated, and clear compliance guidance should be provided to ensure that personal privacy protection is not violated.
Thirdly, innovate data acquisition methods: To enrich the training dataset and improve its diversity, new data source channels can be explored, such as developing synthetic data technology, using simulators to create scenes approximating the real world, or using advanced generative adversarial networks (GANs) and other algorithms to generate data. In addition, data augmentation techniques can enhance the generalization ability and robustness of models by transforming and expanding existing datasets.

4.2Construct a Dual Governance System Combining "Institution and Technology"

Balancing data utilization and protection is an important issue in the development of the AI industry. Traditional legal systems can play an important role in this regard, but they also have limitations. For example,

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

legal paths such as granting data subjects related rights/interests, limiting the scope of data collection, and raising the compliance threshold for data collection and processing often fail to effectively protect data security and privacy, and are not conducive to the innovative development of AI. Therefore, it is suggested to follow the stages and laws of AI technology development and construct a dual governance system combining "institution + technology":

At the institutional level, it is necessary to clarify governance objectives and principles, formulate adaptable, operable, and executable governance rules, and provide institutional guarantees for data utilization and protection (Wang, 2023). Specifically, improve the personal information protection mechanism and protect personal privacy rights and interests. This includes the following three aspects: Firstly, formulate a unified personal information protection law to regulate the definition, classification, and protection principles of personal information, and clarify the legal responsibilities and relief systems for activities such as the collection, storage, transmission, and processing of personal information. Secondly, strengthen the supervision of personal information protection and crack down on illegal acts of personal information. A dedicated personal information protection supervision department and credit system can be established to supervise and publicize the use of personal information and illegal records, forming an incentive and punishment mechanism. Thirdly, strengthen the technical means of personal information protection and improve the security and controllability of personal information. Encryption, desensitization, anonymization, and other technologies can be used to prevent personal information from being leaked or tampered with during the flow. At the same time, technologies such as blockchain and digital identity can be used to enable individuals to control and choose their own information, enhancing the autonomy of personal information.

At the technical level, encourage and support technological innovation, try to let technology solve the problems brought by technology, and provide technical support for data utilization and protection. Specifically, promote the development of privacy computing, AI interpretability, and other technologies to provide feasible technical solutions for the safe, orderly, and compliant circulation of data (Zhou, 2023). Privacy computing technology refers to a class of technologies that encrypt data for processing without revealing data content, thereby protecting data privacy. Privacy computing technology can realize the safe sharing and collaborative analysis of data, provide more and higher quality training data for AI, and also protect the interests of data owners and users. Privacy computing technology mainly includes homomorphic encryption, secure multi-party computation, differential privacy, etc. AI interpretability technology refers to a class of technologies that can interpret and explain the internal mechanisms, decision-making processes, and results of AI systems. AI interpretability technology can enhance the transparency and credibility of AI systems, help users understand and supervise the behavior of AI systems, avoid or correct potential biases, errors, and risks. AI interpretability technology mainly includes model interpretability, process interpretability, result interpretability, etc. In summary, the governance model at the technical level is an effective AI data governance method, which can use technological innovation to solve the problems brought by technology, achieve a balance and coordination between data utilization and data protection. However, relying solely on the governance model at the technical level is not enough, and it also needs to be combined with other levels of governance methods such as ethical norms and legal systems to build a diversified, coordinated, and dynamic AI data governance system.

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

4.3 Formulate Coordinated Multi-Party Interest Data Utilization Rules for Different Fields and Scenarios

The application of AI has penetrated all aspects of social life, and the risk levels of each field and scenario are different. Therefore, adopting a "risk-based regulation" is the key to achieving scientific and refined governance. Taking the "Artificial Intelligence Act" released by the EU in April 2021 as an example, the Act clearly classifies the risk levels of AI—unacceptable risks, high risks, limited risks, and minimal risks, and implements corresponding regulatory measures for different risk levels. This graded supervision model provides us with important references and inspirations. This article believes that based on the "risk governance" principle, data utilization rules that can balance the interests of all parties should be formulated for different fields and scenarios, and considering the rapid changes in AI technology, preventive and regulatory measures should also be made corresponding adjustments with the changes in risk levels and types. For this, we propose the following specific suggestions:

Firstly, a comprehensive analysis and assessment of different data types, the potential impact on individual rights and interests, and the degree of data related to national security should be carried out, and then risk levels should be divided in multiple fields and scenarios, and differentiated data utilization rules should be customized according to these risk levels (Yuan, 2023). In the division of risk levels, it is suggested to start from the following three aspects for detailed planning: Firstly, data sensitivity: considering the sensitivity of the data type, such as personal privacy data, financial transaction data, etc., its risk level and regulatory requirements should be higher than publicly available information. Secondly, the impact on individual rights and interests: analyze the direct or indirect impact that data processing and application may have on individuals, for example, personal credit scoring systems may have a long-term impact on a person's economic activities. Lastly, relevance to national security: for data utilization scenarios involving national security, such as cross-border data flow, infrastructure information, etc., the strictest monitoring and regulation should be implemented.

Secondly, distinguish between "technology research and development scenarios" and "technology application scenarios", and create more convenient and reasonable data utilization rules for the low-risk research and development stage. When formulating data utilization rules, it is necessary to clearly distinguish between different stages of technology implementation in order to more accurately apply risk-corresponding governance measures (Zhao, 2023). Here are rule design suggestions for different stages of technology:
In the stage of technology research and development: In the early stage of technology research and development, researchers usually collect and experiment with data in a closed or semi-closed environment, such as laboratory tests, data analysis in simulated environments, etc. At this stage, the use of data is limited to research purposes, and the impact on the outside world is small, so the risk is relatively low. To encourage innovation, it is recommended to implement a set of more relaxed management measures, such as implementing a data utilization filing system, researchers signing data confidentiality and ethical use commitments, etc., to simplify administrative procedures in the research and development process and promote technological progress.

In the stage of technology application: When technology transitions from the laboratory to the market, its influence extends to the public domain, involving a wider range of users and business scenarios. The pro-

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

ductization of technology and data applications may directly affect the rights and interests of users, such as privacy protection, data security, and avoiding algorithmic discrimination. At this time, the risks brought by technology applications increase significantly, requiring regulators to take more stringent regulatory measures. It is recommended that before the product is brought to market, a thorough privacy impact assessment should be conducted, ensure the explicit acquisition of user consent, and strictly implement the data minimization principle, as well as establish an ethical review mechanism for technology applications.

4.4Accelerate the Improvement of Laws and Regulations Related to Data Utilization

Data utilization rules are an important institutional guarantee for the development of the AI industry and an important means to deal with data security risks. At present, there are some important and urgent problems in China's data utilization rules, and it is necessary to accelerate the formulation and improvement of related laws, regulations, and standard norms, eliminate uncertainties and obstacles, and provide a clear and definite institutional environment for the development of the AI industry. Specific suggestions are as follows:

Firstly, build a basic institutional norm for public data openness as soon as possible and expand the scope of public data openness. Public data refers to data related to public interests that are collected, generated, or held by government departments or other public institutions. Its large volume and wide coverage have important value for the wide application of AI technology and industrial development. On February 11, 2019, then-US President Trump signed the "American AI Initiative", which clearly stated that the federal government's data, algorithms, and computer processing resources will be open to American AI researchers. This shows the importance of data openness to the development of AI. But at present, China's public data openness is insufficient and lacks unified legal norms and standards. This article suggests that a basic legal norm for public data openness should be introduced as soon as possible, clarifying the fields, content, and standards of data openness, especially prioritizing the openness of public data related to people's livelihoods that are of high value, low sensitivity, and large in volume. Establish a dynamic open list of public data in the fields of economic development, people's livelihood services, urban governance, etc., and promote the classified and graded orderly openness of public data in the fields of life and health, autonomous driving, etc., to expand the scope of public data supply for the AI industry. Support related entities to apply for public data openness alone or jointly, and ensure fair use of open data by small and medium-sized enterprises, individual developers, etc.

Secondly, unify the identification rules and catalog of "important data" as soon as possible, and further clarify its specific scope. "Important data" refers to specific fields, specific groups, specific regions, or data of a certain accuracy and scale that, once leaked or tampered with, destroyed, may directly endanger national security, economic operation, social stability, public health and safety, etc. At present, the scope of important data is still unclear, and its identification rules and catalog have a significant impact on the data utilization of AI. It is suggested that supporting rules should be introduced as soon as possible to further clarify the specific scope of "important data". At the same time, considering that different industry regulatory departments and regions all have certain powers to formulate relevant rules or catalogs in their respective fields or regions, there are likely to be inconsistencies, so it is recommended to strengthen overall coordination and unify identification rules and catalogs (Xu, 2023).

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

Thirdly, clarify the anonymization standards of data information as soon as possible, and provide clear and clear compliance guidelines for enterprises. According to Article 4 of the "Personal Information Protection Law", personal information does not include information after anonymization. Therefore, the "anonymization standard" has a significant impact on the collection and use of data information, and it is related to whether AI companies need to obtain authorization and whether they need to conduct safety assessments when collecting and using them. But at present, the anonymization standards for personal information and data in China are not clear, and companies find it difficult to judge what kind of processing constitutes compliant anonymization processing. It is recommended to formulate unified data anonymization standards from the aspects of technology, rules, etc. as soon as possible, to provide clearer basis for companies to do a good job in data classification and grading and formulate business compliance management norms (Tuo, 2023).

5.Conclusion

This study provides a comprehensive analysis of the challenges and current state of AI data utilization in China. It highlights critical issues like the scarcity of high-quality training data, inadequate data privacy protection, and inflexible data circulation mechanisms. To address these challenges, it suggests implementing scientifically grounded AI data usage rules, establishing a dual governance system, and strengthening related laws and regulations. These measures are crucial for leveraging AI's full potential, ensuring China's AI industry grows robustly and ethically, and contributing significantly to the global discourse on AI development.

References

[1] Cheng, XG. (2022). Global artificial intelligence development status, challenges and suggestions for China. Global Technology Economy Outlook, 37(1), 260-270.

[2] Sun, Y. (2023). Research on enterprise data rights and authorization mechanism. Comparative Law Research, 3, 56-73.

[3] Chen, B, & Dong, SY. (2023). Algorithm risk and governance base point of generative artificial intelligence. Learning and Practice, 10, 22-31. DOI:10.19624/j.cnki.cn42-1005/c.2023.10.012

[4] Dong, YT., & Zhang, HZ. (2023). Development concerns and optimization path of big data legal supervision. Journal of Hunan University of Science and Technology (Social Science Edition), 26, 135-142. DOI:10.13582/j.cnki.1672-7835.2023.05.017

[5] Xiang, YQ. (2023). Generative artificial intelligence and policy failure: Avoidance tool or hidden cause? Exploration, 5, 39-50. DOI:10.16501/j.cnki.50-1019/d.2023.05.003

[6] Wang, WC., & Ma, MX. (2022). On the capitalization of big data as a factor of production and its dual impact. Journal of Shanghai University of Finance and Economics, 24, 108-122. DOI:10.16538/j.cnki.jsufe.2022.06.008

[7] Wang, LM. (2023). Legal response to infringement of generative artificial intelligence. Chinese Applied Jurisprudence, 5, 27-38.

[8] Zhou, XF. (2023). Analysis of infringement liability of generative artificial intelligence. Comparative Law Research, 4, 117-131.

[9] Yuan, Z. (2023). Legal regulation of data scale use. Local Legislation Research, 8, 71-84.

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[10] Zhao, ZZ., & Zhuang, XY. (2023). High-quality development of artificial intelligence in China: Current situation, problems and strategies. Reform, 9, 11-20.
[11] Xu, JM. (2023). Principles and legal strategies of generative artificial intelligence governance. Theory and Reform, 5, 72-83. DOI:10.13553/j.cnki.llygg.2023.05.006
[12] Tuo, XD. (2023). On the data security risk of generative artificial intelligence and response governance. Eastern Methodology, 5, 106-116. DOI:10.19404/j.cnki.dffx.2023.05.003

© 2023 by the author(s); licensee Mason Publish Group (MPG), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)