

VALUE ORIENTATIONS AND INSTITUTIONAL CONSTRUCTION OF ARTIFICIAL INTELLIGENCE SYNTHETIC DATA GOVERNANCE IN CHINA

Zhang Zesong

TABLE OF CONTENTS

I. INTRODUCTION	57
II. SYNTHETIC DATA AS	
A SECURE ALTERNATIVE TO REAL DATA	59
<i>A. Limitations of AI Training Based on Real Data</i>	59
<i>B. The Comparative Advantages of</i> <i>Synthetic Data in Driving AI Training</i>	64
III. THE IMPACT OF SYNTHETIC DATA ON	
TRADITIONAL DATA GOVERNANCE SYSTEMS	66
<i>A. Training with Synthetic Data Does Not Eliminate</i> <i>the Risk of Privacy Leakage</i>	67
<i>B. Training with Synthetic Data Exacerbates</i> <i>the Algorithmic 'Black Box' Problem</i>	68
<i>C. Synthetic Data May Become a New Tool for</i> <i>Monopoly in Digital Markets</i>	69
<i>D. Synthetic Data May Lead to</i> <i>Algorithmic Models Becoming Disconnected from Reality</i>	70
IV. FUTURE DIRECTIONS AND INSTITUTIONAL DESIGN FOR	
SYNTHETIC DATA GOVERNANCE	71
<i>A. Value Orientations of Synthetic Data Governance</i>	72
<i>B. Innovative Mechanisms for Synthetic Data Governance</i>	77
V. CONCLUSION	84

VALUE ORIENTATIONS AND INSTITUTIONAL CONSTRUCTION OF ARTIFICIAL INTELLIGENCE SYNTHETIC DATA GOVERNANCE IN CHINA

Zhang Zesong^{*}

As an innovative solution, synthetic data provides a secure and efficient alternative data source for AI training by algorithmically simulating real data's statistical characteristics and distribution patterns. It reduces data collection costs, mitigates privacy leakage risks, and resolves copyright issues, gradually becoming a key driver of AI development. However, widespread use also brings complex governance challenges. During generation and use, sensitive personal information may leak, and the opacity of the process may exacerbate the algorithmic 'black box' problem, making decisions difficult to trace and explain. Meanwhile, large technology enterprises, leveraging first-mover advantages in data and technology, may further consolidate market dominance through synthetic data, triggering new forms of data monopoly. Low-quality or biased synthetic data may also cause models to disconnect from reality, leading to algorithmic discrimination or even system collapse. Facing these challenges, China urgently needs a governance system adapted to synthetic data. Future governance should take standardization, transparency, socialization, and high quality as core values, guiding institutional design that balances innovation and risk prevention by enhancing privacy protection, breaking algorithmic 'black boxes', preventing data monopolies, and avoiding data distortion. On this basis, synthetic data can be incorporated into data protection law frameworks with more relaxed standards, while relying on compliance audits, open-source software, and quality control to build a forward-looking governance system with Chinese characteristics.

I. INTRODUCTION

As a fundamental component of the information world, data has become the lifeblood of the digital economy. The *Outline of the 15th Five-Year Plan for National Economic and Social Development of the People's Republic of*

^{*} Lecturer at Law School of Nankai University. This paper was supported by the General Project of the National Social Science Fund of China, entitled *Artificial Intelligence Modeling of Legal Argumentation* (Grant No. 21BFX033).

China (hereinafter referred to as the 15th Five-Year Plan Outline) proposes that ‘coordinated efforts should be made to advance the development of computing infrastructure, the progress of models and algorithms, and the supply of high-quality data resources, so as to consolidate the foundation for digital and intelligent development.’ Artificial intelligence (hereinafter referred to as AI) relies on massive amounts of data to build training models, and the scale and quality of data directly determine the reliability of model outputs. Currently, much data faces high barriers to access and circulation, including highly sensitive medical, financial, and biometric data. Collecting, storing, and analyzing such data not only entails high costs but may also infringe upon personal data property rights and even lead to large-scale privacy leaks. The limitations and risks of real data have given rise to an alternative solution — synthetic data. Synthetic data is ‘simulated’ data generated by algorithms; it can closely replicate the statistical characteristics and usage scenarios of real data while completely stripping away identifiable information. Because synthetic data is trained on information that does not include identifiable individuals, it alleviates thorny privacy issues. It has become a key driver of the AI industry’s future development. According to Gartner statistics, by 2030, without access to sufficient synthetic data, enterprises will be unable to build high-quality, high-value AI models.¹

The emergence of synthetic data has reshaped how data is accessed and circulated and has also altered the competitive landscape for data-scarce industries. Therefore, some scholars compare the transformative effect of synthetic data on the data industry to the disruptive change brought about by synthetic fibers to the cotton industry.² However, this technological innovation has also impacted China’s data governance system, not only challenging the legal framework centered on the collection of real data but also affecting the balance between data utility and privacy protection. Academically, research on synthetic data has largely focused on its technical risks and privacy value, failing to adequately reveal its impact on deeper

1 See CLAYTON NICHOLAS, ACCELERATING INNOVATION WITH AI USING SYNTHETIC DATA, at <https://vibronyx.com/accelerating-innovation-with-ai-using-synthetic-data> (Last visited on May 19, 2026).

2 See Michal S. Gat & Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 109 Iowa Law Review 1087, 1087 (2024).

dimensions such as market competition patterns and model security. At the same time, existing research lacks effective integration with China's data legal system, resulting in synthetic data not being systematically incorporated into the overall data governance framework. Strengthening AI governance, improving relevant legal norms, and supporting institutional measures are essential to implementing the strategic deployment of 'deeply advancing the construction of a Digital China' as outlined in the 15th Five-Year Plan Outline. In this context, there is an urgent need to re-examine the institutional design of China's data governance and to construct a new governance system adapted to synthetic data, so as to further unlock its value in AI training and promote the healthy development of the data element market.

II. SYNTHETIC DATA AS A SECURE ALTERNATIVE TO REAL DATA

AI training is based on 'large-scale, high-fidelity, and machine-readable data'. However, the collection and use of such data are often subject to multiple constraints, including privacy protection, data security, and compliance regulations. Furthermore, in certain specific scenarios (such as rare medical cases or extreme weather conditions), real data itself is inherently scarce, creating training bottlenecks for models that rely on it. In this context, synthetic data has emerged as a novel data paradigm. By using algorithms to simulate the statistical characteristics and distribution patterns of real data, it can effectively circumvent privacy and security risks while generating diverse, customized training samples on demand. This characteristic of being a 'secure alternative' makes it a key link in deepening the integration of real and digital economies. This section primarily explores how synthetic data alleviates technical and legal challenges posed by real data in AI training, thereby laying the foundation for a systematic analysis of the governance dilemmas surrounding synthetic data and potential pathways forward.

A. Limitations of AI Training Based on Real Data

AI training primarily relies on real-world data. However, with technological evolution and increasing stringency of global data protection

regulations, this training paradigm based on real data has revealed problems.

1. High Costs of Collecting and Preprocessing Real Data. — Although the total volume of data in human society is growing exponentially, the cost of data collection in specific domains is facing a significant upward trend. On the one hand, large data-holding enterprises are gradually raising the barriers to data access and strictly restricting third parties from using their data for AI training. Currently, companies represented by OpenAI and Google have comprehensively revised their terms of service, explicitly prohibiting third parties from collecting their data for AI training. While the enforceability of such clauses remains controversial, most tech giants tend to leverage their control over data access and collection to exclude or discriminate against downstream competitors. This creates a dual market-suppression effect: Small and medium-sized enterprises, constrained by prohibitively high data-collection costs, struggle to build competitive databases; simultaneously, these enterprises' willingness to invest in R&D to optimize algorithmic models diminishes. On the other hand, the scarcity of specific data drives up collection costs, further constraining the training of algorithmic models. In autonomous driving models, incorporating edge cases (such as a vehicle ahead dropping heavy objects) can enhance the driving system's emergency response capabilities. However, such data is inherently scarce — the probability of its natural occurrence in real-world road tests is extremely low. Enterprises must either invest substantial resources in building simulated scenarios for data collection or pay premiums to purchase such data from third parties. Similar situations also exist in fields such as financial fraud prevention and healthcare, constituting common challenges across industries.

Real data not only entails high collection costs but also requires preprocessing to be transformed into structured data suitable for training. Preprocessing includes cleaning, annotation, and integration. These tasks are universally characterized by high costs, with annotation being particularly expensive, time-consuming, and error-prone.³ Industry data indicates that annotating data for Waymo (an autonomous driving company) requires manual operations by professional teams, with per-frame annotation costs

3 See James Steinhoff, *Toward a Political Economy of Synthetic Data: A Data Intensive Capitalism That Is Not a Surveillance Capitalism?*, 109 *New Media & Society* 3290, 3295 (2024).

exceeding \$5, while constructing a complete training dataset requires annotating at least millions of frames. Furthermore, human annotators may introduce harmful biases during the labeling process, leading to significant error amplification effects and incurring high retraining costs.⁴ Evidently, the demand for data preprocessing further escalates the difficulty and cost of training with real data.

2. Threats to Personal Privacy Posed by the Use of Real Data. — Algorithmic model training datasets contain vast amounts of publicly available personal data scraped from the Internet, and model users can extract this data with relative ease, using it to infer even more sensitive information about individuals. For instance, when researchers analyzed the public LAION dataset, they precisely pinpointed the location where a photograph of a baby was taken.⁵ Furthermore, algorithmic models may inadvertently leak private information in their outputs due to memorization of training data, enabling third parties to access sensitive information about data subjects directly. Users might disclose their personal data on online platforms for specific purposes, yet they do not expect this data to be used for AI training. In reality, most users who provide data are entirely unaware that their information could be incorporated into training datasets and utilized in contexts extending beyond their original purposes. Although *China's Personal Information Protection Law* includes publicly available personal data within its scope of protection, it currently provides general safeguards through basic obligations and use restrictions, which leaves room for improvement in effectively preventing data scraping and privacy risks inherent in derived datasets.

Anonymizing personal data has been regarded as the optimal solution for preventing privacy breaches. The appeal of anonymization lies in its ability to strike a balance between privacy protection and data utility. It enables the full utilization of valuable yet sensitive data while reducing the linkability between such data and individuals.⁶ Data processed through anonymization

4 See Zhang Xin, *Data Risks and Governance Pathways for Generative Artificial Intelligence*, 5 Science of Law (Journal of Northwest University of Political Science and Law) 42, 44 (2023).

5 See Mindy Nunez Duffourc, Sara Gerke & Konrad Kollnig, *Privacy of Personal Data in the Generative AI Data Lifecycle*, 3 New York University Journal of Intellectual Property and Entertainment Law 219, 250 (2024).

6 See Elizabeth A. Brasher, *Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation*, 2018 Columbia Business Law Review 209, 217 (2018).

no longer falls within the scope of legally protected personal data, thereby exempting data processors from subsequent statutory obligations to safeguard personal privacy. However, anonymization has not genuinely eliminated the privacy threats posed by real data, as such data remains susceptible to re-identification. With advancements in re-identification techniques, instances of identifying specific individuals from seemingly anonymous data are increasingly common. Professor *Sweeney* effortlessly identified the medical records of a public official in her region by matching ‘anonymized’ health information released by an insurance company with publicly available voter registration records.⁷ Consequently, some scholars regard anonymization as a failed technology, emphasizing its inadequacy in addressing emerging threats such as data linkage and model memorization.⁸ The limitations of anonymization demonstrate that merely deleting or concealing identifiable information within datasets cannot provide absolute privacy protection. In this sense, regardless of whether the data has been anonymized, training with real data consistently faces the challenge of privacy leakage.

3. *Real-Data Training Prone to Copyright Infringement Risks.* — The use of real data for training also presents a significant issue — the potential for large-scale copyright infringement. This risk is particularly pronounced in applications of large AI models like ChatGPT. The training data used by such applications primarily originates from publicly available online resources, including news reports, e-books, images, videos, and artistic works. Since copyright protection only requires meeting the two basic conditions of originality and fixation, the vast majority of this training data consists of copyright-protected material.⁹ In practice, many enterprises utilize others’ works as training data through opaque processes without obtaining their consent. Such training typically involves reproducing underlying texts, sounds, and images, thereby impeding authors’ exclusive control over their works. For instance, Meta Platforms trained its AI model on the ‘Books’ dataset, which contains nearly 200,000 copyright-protected e-books, without obtaining

7 See Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 Washington Law Review 703, 711 (2016).

8 See Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA Law Review 1701, 1744 (2010).

9 See Peter Lee, *Synthetic Data and the Future of AI*, 110 Cornell Law Review 1, 19 (2025).

permission from the authors.¹⁰ The unpaid training of AI models using copyright-protected works has triggered numerous infringement lawsuits. The New York Times has filed a lawsuit against OpenAI, alleging that its language model reproduced millions of articles verbatim and utilized them for training.

Given that copyright protection serves the dual purposes of incentivizing creative activities and promoting public cultural, social, and economic welfare, scholars tend to argue that using copyright-protected data for training, while involving reproduction, falls within the scope of fair use as long as it does not actually harm the copyright holder's interests in disseminating original works to the public.¹¹ Fair use refers to the lawful act of allowing others to freely use protected works without the copyright owner's consent under specific conditions. Section 107 of Title 17 of the *U.S. Copyright Code* stipulates four factors for consideration in determining fair use: (1) the purpose and character of the use; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work. Courts must evaluate specific conduct based on these factors to determine whether it constitutes fair use.¹² In *Google LLC v. Oracle America, Inc.*, Google used Oracle's Sun Java API for training without permission or payment to assist in developing its Android system. The court held that although Google reproduced substantial portions of the Sun Java API code during AI training, it intended to create a system capable of 'stimulating creativity for the public benefit', which aligned with copyright's dual purposes. The impact on the market value of Sun Java API was minimal. Accordingly, the court ruled that Google's conduct constituted fair use.¹³

Although fair use justifies the large-scale reproduction of copyright-protected data, it has not eliminated concerns that training with real data may infringe copyright. This is because fair use itself is fraught with uncertainty in legal application; specific usage contexts constrain it and require *ex post facto* determination by courts. Until such a determination is made, AI training remains

10 See Art Neill, James Thomas & Erika Lee, *A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence*, 32 Texas Intellectual Property Law Journal 225, 227 (2024).

11 See Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 Berkeley Technology Law Journal 1407, 1407 (2021).

12 See Lin Xiuqin, *Reshaping the Fair Use System in Copyright Law in the AI Era*, 6 Chinese Journal of Law 170, 174 (2021).

13 See *Google LLC v. Oracle America, Inc.*, 141 S. Ct. 1183, 1209 (2021).

at risk of infringement.¹⁴ In *Fox News Network, LLC v. TVEyes, Inc.*, the court held that the media-monitoring service TVEyes provided to its users exceeded the bounds of fair use. This was because TVEyes allowed users to save and copy copyright-protected program videos — sufficiently complete to present the program content in their entirety, effectively serving as a substitute for Fox’s original works.¹⁵ This demonstrates that when copying conduct may lead to market substitution, fair use cannot legitimize the large-scale unauthorized copying undertaken by AI to collect training data. Some scholars further argue that fully subsuming the infringement issues arising from AI training under copyright control would not only further blur the boundaries of fair use but could also foster a high-risk environment that erodes the creative industries, ultimately leading to the overall atrophy of the Internet’s creative ecosystem.¹⁶

B. The Comparative Advantages of Synthetic Data in Driving AI Training

Synthetic data is a type of artificial data generated through two primary methods: first, by deep learning models; and second, by data simulators.¹⁷ Deep learning models primarily include Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs are unsupervised machine learning models consisting of an encoder and a decoder. The encoder compresses the input real data into a more compact representation and maps it to a low-dimensional latent space. The decoder extracts variables from this space and reverse-maps them back to the real data space, generating synthetic data that retains the statistical characteristics of the real data while exhibiting differentiation. GANs mimic human thinking by creating two competing neural networks: a generator and a discriminator. The generator produces synthetic data without using collected data, while the discriminator compares it with real data to identify which aspects of the synthetic data reveal its ‘falseness’. The generator then optimizes itself based on the discriminator’s feedback until the

14 See Micaela Mantegna, *Artificial: Why Copyright Is Not the Right Policy Tool to Deal with Generative AI*, 133 The Yale Law Journal Forum 1126, 1165 (2024).

15 See Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 Fordham Law Review 1887, 1905 (2024).

16 Micaela Mantegna, *supra* note 14, 1129.

17 See Ye Yingjie & Li Chuan, *Application Risks of Synthetic Data in AI Model Training and Its Governance Pathways*, 6 Information Studies: Theory & Application 47, 48 (2021).

discriminator can no longer distinguish between real and synthetic data. Data simulators can generate data directly using preset mathematical models or rules, without relying on real data, with the complexity of these models or rules depending on the intended use of the generated data. By analyzing the methods for generating synthetic data, the key distinctions between synthetic and real data can be clarified, thereby better revealing its advantages in AI training.

Firstly, training with synthetic data significantly reduces the costs of data collection and preprocessing. By transforming the physical-world data collection process into a digital parameter-modeling procedure, synthetic data enables AI training to break free from reliance on real data. By defining key parameters and statistical characteristics of the data, algorithms can automatically generate data that meets training requirements. This generation mechanism not only reduces investment in data-collection hardware but also overcomes the spatiotemporal limitations of scarce data. Moreover, generating synthetic data using GANs or data simulators does not require prior collection of large-scale real data. Particularly when generated through data simulators, there is fundamentally no need to collect and use real data directly. Additionally, synthetic data enables the integration of data collection and preprocessing by completing annotation and structuring simultaneously during generation, thereby reducing data preprocessing costs. The on-demand generation characteristic of synthetic data naturally results in well-structured data formats, reducing data redundancy and the need for data cleaning and integration. Deep learning models and data simulators can automatically add annotations to data, eliminating expensive manual annotation and precisely capturing information that manual operations cannot annotate.

Secondly, training with synthetic data effectively mitigates threats to personal privacy. The generation process of synthetic data itself inherently protects privacy. On the one hand, traditional anonymization techniques rely on obfuscating personal data, with the core approach being the removal or concealment of identifiable information. However, to preserve the statistical characteristics of data and maintain its utility, such processing often struggles to eliminate the risk of re-identification. Synthetic data effectively circumvents this issue. By using algorithms to generate ‘artificial’ data that is statistically similar to the original but lacks real-world counterparts, it severs the correspondence between data and subjects, making it difficult for attackers to restore the original

data and identify its source. On the other hand, to control the potential scope of impact from data leakage or misuse, article 5(1) of the *EU General Data Protection Regulation* (GDPR) establishes the principle of data minimization. Article 1035 of *China's Civil Code* and article 6 of the *Personal Information Protection Law* similarly emphasize processing only the minimum amount of personal data necessary for specific purposes. Since synthetic data is generated on demand, it neither creates redundant storage nor retains data beyond its intended use, thereby reducing data processors' motivation to collect data excessively, which aligns with the principle of data minimization.

Finally, training with synthetic data can mitigate the risk of copyright infringement under specific conditions. Once AI directly uses copyright-protected data for training, it may trigger large-scale copyright infringement. Although fair use can serve as a defense, the uncertainty surrounding its application still exposes enterprises to considerable legal risks. In contrast, whether based on deep learning models or employing data simulators, the data generation process does not involve direct copying of original copyrighted content, thereby technically circumventing potential infringement risks. Consequently, OpenAI's CEO has praised synthetic data as a technology that enables the development of more powerful models without relying on copyright-protected data.¹⁸ It is worth noting that the assertion that 'using synthetic data can circumvent copyright infringement risks' does not hold in all cases. Synthetic data itself may contain original expressions protected by law, and if a third party uses such data for training without authorization from the copyright holder, infringement may still occur. Overall, synthetic data has not fully resolved the issue of copyright infringement; rather, it has shifted these concerns to the data-generation stage.

III. THE IMPACT OF SYNTHETIC DATA ON TRADITIONAL DATA GOVERNANCE SYSTEMS

Synthetic data, by shifting data production from physical collection to algorithm-driven generation, effectively addresses issues such as high costs,

18 See KATHERINE LEE, A. FEDER COOPER & JAMES GRIMMELMANN, TALKING ABOUT AI GENERATION: COPYRIGHT AND THE GENERATIVE AI SUPPLY CHAIN, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551 (Last visited on May 19, 2026).

privacy leaks, and copyright infringement inherent in training with real data. However, while bringing technological breakthroughs, synthetic data also alters the underlying logic of various aspects of data governance, making it difficult to apply traditional legal systems grounded in real data collection directly to algorithmically generated synthetic data. Consequently, under the current governance framework, even though synthetic data circumvents many limitations of real-data training, it still faces unique governance challenges.

A. Training with Synthetic Data Does Not Eliminate the Risk of Privacy Leakage

Anonymization prevents privacy leaks by stripping identifiable information from real data. Yet it struggles to eliminate the possibility of re-identification and is often accompanied by a loss of data utility. Some scholars point out that anonymization is a ‘subtractive’ technique that performs poorly in balancing privacy protection with data utility; leveraging the power of algorithms, synthetic data offers an ‘additive’ approach to privacy protection, creating high-fidelity artificial data to achieve the effect that ‘no real data means no risk exists’.¹⁹ Nevertheless, synthetic data is not infallible in terms of privacy protection and cannot serve as the ultimate solution for privacy safeguarding.

On the one hand, the security of synthetic data depends on the extent to which algorithms replicate the statistical characteristics of real data. If the model over-memorizes individual features of real data, unintended similarities may emerge between synthetic and real data. In such cases, attackers may deduce identity information from the original data by comparing generated samples with public databases — a phenomenon in computer science known as model overfitting.²⁰ On the other hand, although synthetic data generated by algorithms severs ties with data subjects upon sharing, parameter sharing among algorithmic models may still be a potential pathway for privacy leakage. Furthermore, algorithmic model parameters may contain more sensitive characteristics of the original data than the synthetic data itself; when attackers gain access to these parameters, they can

19 See Steven M. Bellovin, Preetam K. Dutta & Nathan Reitering, *Privacy and Synthetic Datasets*, 22 Stanford Technology Law Review 1, 35 (2019).

20 See Reuben Binns & Lilian Edwards, *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*, 376 Philosophical Transactions of the Royal Society 1, 4 (2018).

reconstruct the real data through model reverse engineering, even without direct access to the data. Therefore, even limited access to the generative models of synthetic data may lead to privacy breaches.

*B. Training with Synthetic Data Exacerbates
the Algorithmic ‘Black Box’ Problem*

There is a well-known ‘black box’ problem in algorithms, where omitting intermediate reasoning steps in autonomous learning results in a lack of transparency and interpretability in the decision-making process. During AI training, this issue may be exacerbated by the use of synthetic data. On the one hand, the generation process of synthetic data itself suffers from the ‘black box’ problem. This is because its generative models, based on complex neural networks, abstractly fit the distribution of real data through multi-layer nonlinear transformations. This process relies on the implicit functionality of the model’s autonomous learning, bypassing the explicit steps in machine learning. Although model users can comprehend the overall architecture and principles of the algorithm, capturing or understanding the specific steps from generation instructions to synthetic data output remains practically unattainable. On the other hand, training with synthetic data may lead to a superposition of ‘black box’ problems, forming a dual-layered opaque structure. When models are trained on data of unknown content and origin, their transparency further diminishes. Since synthetic data is not a direct reflection of real data, it lacks transparency and interpretability at its source and may embed biases distinct from those in real data. This introduces additional uncertainty into algorithmic models that are already inherently ‘black box’ in nature,²¹ making problem tracing significantly more difficult — rendering it impossible to distinguish whether issues originate from the synthetic data generation stage or the training stage.

China currently addresses the issue of algorithmic ‘black boxes’ primarily through laws and regulations such as the *Personal Information Protection Law* and the *Provisions on the Administration of Algorithmic Recommendation in Internet Information Services*. These regulations emphasize the fairness, transparency, and interpretability of algorithmic

21 See Zhang Tao, *Coherent Legal Governance of Synthetic Data in AI Training*, 2 Computer Science 20, 26 (2021).

decision-making and grant users rights and protections, such as the right to know and the right to choose. However, existing laws remain insufficient to address the algorithmic ‘black box’ issues arising from synthetic-data training. First, current laws are mostly principled provisions and lack specific norms tailored to the characteristics of synthetic data, leaving room for clearer normative guidance to support judicial adjudication. Second, existing laws focus on *post hoc* accountability rather than process transparency, failing to establish a traceability mechanism for the synthetic-data training process, making it difficult to identify responsible parties. Third, current laws lack effective measures to address the dual ‘black box’ problem — they neither require disclosure of the synthetic data generation process nor impose obligations on downstream models to verify the source of synthetic data, allowing the ‘black box’ problem to propagate continuously.

C. Synthetic Data May Become a New Tool for Monopoly in Digital Markets

As the fundamental raw material for generating information and knowledge, data enables more precise decision-making and more efficient training, thereby becoming a critical resource for competition among enterprises in digital markets. When barriers to data collection are high, startups often struggle to challenge incumbent enterprises that, by virtue of their data advantages, hold dominant positions, as they cannot offer users products and services derived from superior datasets. Theoretically, synthetic data could help reshape this competitive landscape. On the one hand, synthetic data can expand dataset scale, effectively compensating for deficiencies in real data volume or limitations in scenario coverage, thereby helping enterprises with relatively scarce data resources compete against those with data advantages. On the other hand, synthetic data holds promise for enhancing competitiveness by promoting data sharing. As synthetic data technology matures, the market pricing of real data will no longer be determined solely by its scarcity. Rather, it will increasingly be constrained by the costs of generating comparable synthetic data. When enterprises find it difficult to translate their data into significant competitive advantages, their willingness to share data increases, stimulating competition and fostering the

emergence of new, higher-quality data products or services.²²

The view that synthetic data promotes competition in digital markets rests on an idealized assumption that potential entrants or startups can generate synthetic data freely and at low cost. However, the capability to generate synthetic data may itself become a tool for monopolization. This is because synthetic data has not eliminated the need for real data collection; rather, it may perpetuate or even reinforce dependence on real data during generation. Enterprises with first-mover advantages, leveraging their vast data reserves and technological superiority, can generate larger quantities of higher-quality synthetic data; disadvantaged enterprises, lacking comparable real data volumes and generative capabilities, struggle to independently generate competitive synthetic data, resulting in greater dependence on the former. More importantly, the generation process of synthetic data lacks transparency and interpretability, enabling advantaged enterprises to indirectly dominate the training of downstream models by controlling the generation rules, output quality, and access conditions of synthetic data.²³ This implicit control not only consolidates their market position but also leads to further concentration of data resources in the hands of a few enterprises.

D. Synthetic Data May Lead to Algorithmic Models Becoming Disconnected from Reality

If the data used for AI training contains biases, the algorithm may make decisions that have discriminatory effects on individuals. Amazon once developed an AI recruitment tool for automatically screening job applicants' resumes. Because the training data for this tool was based on resumes received by the company over nearly a decade (in which males constituted the vast majority), the algorithm quickly learned and reinforced the gender bias present in the historical data during its operation, leading to discrimination against female job applicants.²⁴ When deep learning models generate synthetic data, if the underlying real data contains biases, the generated data will perpetuate these flaws. Coupled with the

²² Michal S. Gat & Orla Lynskey, *supra* note 2, 1112.

²³ See Shang Jiangang, *A Study on the Meta-Rules of Risk Governance for Generative Artificial Intelligence*, 3 *Oriental Law* 4, 7 (2023).

²⁴ See JEFFREY DASTIN, AMAZON SCRAPS SECRET AI RECRUITING TOOL THAT SHOWED BIAS AGAINST WOMEN, at <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-USKCN1MK08G> (Last visited on May 19, 2026).

large-scale generation capabilities of deep learning models, even minor biases can trigger significant discrimination. Additionally, biases can also be introduced when generating synthetic data using data simulators. Although simulators can generate synthetic data without using real data, their generation logic still relies on predefined mathematical models and rules. These settings reflect the designers' subjective cognition and assumptions, which may embed implicit biases in the synthetic data. Moreover, without the objective constraints of real data, such biases tend to be more difficult to detect and correct.

From a broader perspective, synthetic data containing biases poses a serious threat to the long-term development of AI. The term 'bias' here refers not only to the perpetuation of social inequalities but, more broadly, to the deviation of synthetic data from reality. If the proportion of low-quality synthetic data within a dataset is too high, it is likely to trigger a 'recursive risk' — downstream models continuously train on and generate large amounts of distorted data, causing biases to accumulate through multiple iterations of synthetic data generation. The data generated in this way will ultimately 'pollute' the next generation of data.²⁵ Training on such contaminated data leads to erroneous judgments about reality, which not only degrades the quality of model outputs but may also cause a continuous decline in model performance, or even collapse. Some scholars refer to this as 'AI eating its own tail'.²⁶ In this regard, article 10(2) of the *EU Artificial Intelligence Act* requires a review of synthetic data for possible biases, but how to proceed after discovering biases remains unclear. Article 4 of *China's Interim Measures for the Management of Generative Artificial Intelligence Services* states that when selecting training data, effective measures should be taken to prevent discrimination based on ethnicity, beliefs, gender, age, and other factors.

IV. FUTURE DIRECTIONS AND INSTITUTIONAL DESIGN FOR SYNTHETIC DATA GOVERNANCE

While synthetic data aims to address issues such as privacy protection and data scarcity, it also engenders new governance challenges: Not only

25 See Ilia Shumailov, Zakhar Shumaylov & Yiren Zhao et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 *Nature* 755, 755 (2024).

26 See Amy Cyphert, Samuel J. Perl & S. Sean Tu, *AI Cannibalism and the Law*, 22 *Colorado Technology Law Journal* 301, 301 (2024).

does it fail to eliminate the risk of privacy leakage, but it also exacerbates the algorithmic ‘black box’ problem due to the opacity of its generation process. The barriers to generating and using synthetic data may foster new forms of data monopolies, leading to a high concentration of data resources. More critically, the use of low-quality synthetic data may cause machine learning models to deviate from reality, leading to severe consequences, including algorithmic discrimination and system collapse. Faced with these challenges, synthetic data governance must transcend technology itself, establishing value orientations centered on standardization, transparency, socialization, and high quality. Translating these principles into specific institutional norms is essential to overcoming existing governance dilemmas and to forming a more systematic and comprehensive synthetic data governance system.

A. Value Orientations of Synthetic Data Governance

Standardization, transparency, socialization, and high quality constitute the ‘soul’ of synthetic data governance, systematically answering the pivotal question: ‘What kind of synthetic data governance system should we establish?’ Standardization enhances privacy protection through legal constraints, laying a solid legal foundation for the development and utilization of synthetic data. Transparency pursues traceability in the generation and use processes of synthetic data, aiming to strengthen algorithmic and data supervision, thereby ensuring that synthetic data is trustworthy and controllable. Socialization is dedicated to promoting the open sharing of synthetic data resources, enabling the benefits of synthetic data technologies to reach a wider range of market entities and the public. High quality aims to support the deep integration of the digital and real economies, ensuring the practical utility of synthetic data by strengthening the supply of high-level data. These elements are both independent and mutually reinforcing, collectively constructing the fundamental framework for synthetic data governance.

1. Enhancing Privacy Protection Through Standardization. — Currently, laws in various countries generally categorize synthetic data as ‘anonymized data’ and exempt it from protection obligations. For example, the French National Commission on Informatics and Liberty (CNIL) recognizes the anonymized nature of synthetic data based on criteria such

as individualization, linkability, and inferability.²⁷ These views are also reflected in the GDPR and *China's Personal Information Protection Law*. However, technical 'anonymization' of synthetic data does not equate to legal 'harmlessness'; overfitting of algorithmic models and parameter sharing may both lead to the identification of sensitive personal information from the original data. Current laws have yet to fully address the privacy concerns arising from the generation and use of synthetic data. Therefore, it is necessary to move beyond a technical understanding of synthetic data and incorporate it more explicitly within the regulatory scope of data protection laws through normative design, thereby achieving full-cycle standardized governance of synthetic data from generation to use.

To enhance privacy protection through standardization, the primary task is to clarify the legal status of synthetic data. The law needs to define its differences from anonymization, highlighting its technical characteristic of being generated through algorithms rather than mere desensitization. On this basis, synthetic data can be classified for protection according to different application scenarios. For example, different protection standards could be established for synthetic data used in scientific research versus commercial analysis, which also aligns with the 'classification and grading' protection principle in China's data governance.²⁸ It is worth noting that while standardized governance of synthetic data can fill gaps in privacy protection, it may also lead to misapplication of laws. Current laws are designed to regulate the processing of real personal data, and their normative requirements, such as informed consent, purpose limitation, and data minimization, have clear data subjects. Although most synthetic data originates from personal data, once generated algorithmically, it no longer directly corresponds to specific subjects, making it difficult to apply existing norms. For instance, the core of informed consent lies in the data subjects' control over their own data, yet synthetic data, in form, is detached from the identity of the original data subject. At this point, there is no clear subject from whom to obtain consent, rendering informed consent ineffective due to the virtualization of the subject. Therefore, current laws need to be adapted

27 See Alexander Boudewijn & Andrea F. Ferraris, *Legal and Regulatory Perspectives on Synthetic Data as an Anonymization Strategy*, 1 *Journal of Personal Data Protection Law* 17, 24 (2024).

28 Zhang Tao, *supra* note 21, 27.

to the characteristics of synthetic data, shifting the regulatory focus from ‘individual empowerment’ to ‘process control over data generation and use’.

2. *Breaking the Algorithmic ‘Black Box’ Through Transparency.* — Due to the high opacity of algorithmic models and training processes, biases and privacy risks latent within synthetic data are difficult to identify and trace effectively. This not only undermines the reliability of the data itself but also makes it challenging to determine responsible parties. In this context, ensuring transparency in synthetic data governance becomes particularly important. The drafters of the GDPR incorporated transparency into data protection legislation, emphasizing that information regarding how personal data is collected, used, or otherwise processed should be transparent and easily accessible.²⁹ The emphasis on transparency in the GDPR aligns with the provisions of the *EU Artificial Intelligence Act*, which also requires the creation and maintenance of a publicly accessible database of information on numerous high-risk automated decision-making systems to enhance transparency in AI training.³⁰ Although these regulations are directed at the transparency of personal data, some scholars suggest extending their scope to synthetic data. This would entail data controllers disclosing to data subjects whether their personal data has been used to generate synthetic data and the corresponding protection measures, including communication through privacy notices and, where necessary, publishing the conclusions of Data Protection Impact Assessments.³¹

This article posits that, in addition to integrating the principle of transparency into data protection legislation, the transparency of synthetic data governance can be enhanced from two further aspects. On the one hand, it is necessary to improve model interpretability at a technical level. Fully understanding algorithmic models with diverse parameters is impractical; the core of interpretability lies in seeking abstract explanation methods that match the model’s characteristics. Professor *Lipton* distinguishes between different types of explanations based on model features. For highly opaque models such as VAEs and GANs, *Lipton* advocates

29 Recital 39 of the GDPR.

30 See W. Gregory Voss, *AI Act: The European Union’s Proposed Framework Regulation for Artificial Intelligence Governance*, 25 *Journal of Internet Law* 1, 7 (2021).

31 Alexander Boudewijn & Andrea F. Ferraris, *supra* note 27, 28 (2024).

abandoning the dismantling of their internal structures and instead adopting *post-hoc* explanation methods, such as local and example-based explanations, to analyze the relationship between model inputs and outputs, thereby inferring their behavioral logic and decision-making basis in reverse.³² Local explanations focus on a specific generated output, analyzing which parts of the input led to that outcome; example-based explanations provide intuitive support for model interpretation by displaying generated results most similar to the model's current output and employing analogical logic.³³ It is important to note that an excessive pursuit of interpretability may simplify model structures or restrict data features, thereby sacrificing the accuracy of synthetic data. Therefore, a trade-off between accuracy and interpretability is necessary, with treatments differentiated by risk levels across application scenarios.³⁴ On the other hand, information disclosure concerning key stages from the generation to the use of synthetic data is required, including the sources of original data, the basis for model selection, and core parameter settings, thereby forming a traceable technical chain. Achieving transparency also requires ensuring timely disclosure and tailoring it to the stakeholders' level of professional knowledge. For example, data producers should disclose information such as the sources of personal data in training corpora. They can provide detailed technical documentation to regulatory authorities for audit purposes. At the same time, data controllers should use clear, understandable language to explain the purpose of data use and the potential impact on personal privacy.³⁵

3. *Preventing Market Monopoly Through Socialization.* — Synthetic data was originally expected to optimize the competitive landscape of digital markets. However, large technology companies, leveraging their first-mover advantages in accumulating original data and technological capabilities, can generate higher-quality, larger-scale synthetic data, thereby consolidating their market dominance. The scale of data held by any single entity is insufficient to construct datasets that meet diverse needs. Only by sharing

32 See Philipp Hacker, Ralf Krestel & Stefan Grundmann et al., *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, 28 *Artificial Intelligence and Law* 415, 430 (2020).

33 See Ding Xiaodong, *Trust-Based Automated Decision-Making: Reflections on the Principles and Institutional Reconstruction of the Right to Algorithmic Explanation*, 1 *China Legal Science* 99, 103 (2022).

34 *Id.*

35 See Zhang Xinbao, *Research on Personal Information Protection in Training Data for Generative Artificial Intelligence*, 6 *China Legal Science* 86, 100 (2024).

data resources with other entities — shifting from private, exclusive use to open, shared, and socialized use — can the competitive vitality of digital markets be truly stimulated.³⁶ Therefore, the governance of synthetic data should promote the formation of a more socialized governance ecosystem, enabling small and medium-sized enterprises to independently generate or conveniently collect synthetic data. To achieve this goal, efforts can be made on both the supply and access sides: increasing the sources of synthetic data while also broadening the channels for accessing it.³⁷ Increasing supply sources requires cultivating diverse data-generating entities, encouraging enterprises and even the public sector to independently generate high-quality synthetic data, thereby breaking dominant companies' control over data supply. Broadening access channels means establishing convenient mechanisms for data circulation and sharing, thereby offsetting market concentration effects by reducing data collection costs and enabling a wider range of user groups to utilize synthetic data.

Synthetic data, like other data assets, possesses the economic characteristics of public goods: non-rivalry and non-excludability.³⁸ In a competitive market environment, public goods such as synthetic data may face under-provision or access barriers.³⁹ The reason lies in the relatively high costs of the initial generation phase of synthetic data (such as collecting original data and designing generative models). In contrast, the marginal cost of replication and distribution is extremely low. 'Free-riders' can acquire vast amounts of synthetic data at nearly zero cost with just a click. Therefore, to promote the socialization of synthetic data governance on both the supply and access sides, it is necessary to provide economic incentives to data developers and to construct an open ecosystem for data production and circulation, ensuring they receive appropriate returns for creating and sharing synthetic data.

4. Avoiding Data Distortion Through High Quality. — Synthetic data plays a crucial role in addressing the scarcity of real data and the challenges of privacy protection, and the core foundation of its functionality lies in ensuring

36 See Gao Fuping, *The Theory of Data Production — A Foundational Theory for the Allocation of Data Resource Rights*, 4 SJTU Law Review 5, 6 (2019).

37 See Zhao Jingwu, *On the Institutional Construction of High-Quality Supply of Training Data for Artificial Intelligence*, 1 China Law Review 92, 96 (2025).

38 See Ji Hailong, *The Positioning and Protection of Data in Private Law*, 6 Chinese Journal of Law 72, 82 (2018).

39 See Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 Texas Law Review 1031, 1051 (2005).

authenticity. It must be clarified that fidelity does not mean pursuing complete consistency with the superficial form of real data, as excessive similarity may instead increase the risk of privacy leakage. Rather, it primarily requires the accurate capture and reproduction of the intrinsic statistical characteristics, distribution patterns, and structural relationships of real data. If synthetic data is severely distorted, it will bring significant harm. Taking IBM's Watson Health as an example, this model was trained on artificially generated, biased medical records, ultimately producing inappropriate tumor treatment recommendations. The authenticity of synthetic data directly determines the reliability of model outputs. Thus, avoiding data distortion becomes an important goal of synthetic data governance.

To achieve this goal, the concept of 'high quality' must be implemented, which requires systematic quality control throughout the entire lifecycle of synthetic data. Generating high-quality synthetic data requires clarifying which statistical characteristics and key variables of the original data should be preserved. Its technical realization depends on the selection and continuous optimization of algorithmic models. For already-generated synthetic data, clear quality evaluation standards need to be established and systematic testing procedures implemented. Only synthetic data that meets established quality standards should be used for training. In contrast, anomalous data that deviates from expectations and fails to meet standards should be traced back to the corresponding earlier processes for targeted optimization or regeneration.⁴⁰

B. Innovative Mechanisms for Synthetic Data Governance

Based on an exploration of the future directions of synthetic data governance, this article has preliminarily clarified the type of synthetic data governance system to be established. Guided by the values of standardization, transparency, socialization, and high quality, it is necessary to incorporate synthetic data into the normative framework of China's data protection law. Furthermore, various institutional challenges in synthetic data governance may be addressed through the establishment of mechanisms for

⁴⁰ See Huo Jingyu, *Risks and Regulations of Empowering the Construction of a Digital Law-Based Government with Synthetic Data*, 5 Administrative Law Review 63, 75 (2025).

auditing, open source, and quality control.

1. Incorporating Synthetic Data into the Scope of Protection Under Current Laws. — From a functional perspective, synthetic data does not exceed the scope of anonymized data; both sever the direct correspondence between data and specific individuals. The difference lies in their implementation pathways: anonymized data relies on deleting or masking identifiable information in real data. In contrast, synthetic data is generated through algorithms to replace identifiable information in real data. Precisely because of this, as long as it is generated properly, synthetic data technology can be regarded as a novel and efficient anonymization tool. Therefore, treating synthetic data as a type of anonymized data is theoretically feasible. However, *China's Personal Information Protection Law* excludes from its scope of protection data that, after anonymization processing, cannot identify specific individuals and cannot be restored. Synthetic data falls into this category, so it automatically obtains legally exempt status upon generation. The *Personal Information Protection Law's* current approach to anonymized data poses a theoretical challenge for the regulation of synthetic data under current laws.

Under the GDPR, although anonymized data is exempt from certain requirements, the concept of 'pseudonymization' encompasses data that poses lower privacy risks but is not fully anonymized and thus remains within the scope of protected personal data. Recital 26 of the GDPR states that 'information which has undergone pseudonymization ... should be considered as information on an identifiable natural person.'⁴¹ Clearly, pseudonymization in the GDPR represents a process of reducing the linkability between data and specific individuals, rather than completely preventing the identification of specific individuals. Therefore, pseudonymous data remains protected under the GDPR, albeit with more lenient protection standards than general personal data. Article 73 of *China's Personal Information Protection Law* clarifies the concept of de-identification, which is similar to pseudonymization in the GDPR, emphasizing that de-identified data cannot, by itself, identify specific individuals. However, their identity may still be restored with additional information. In this regard, synthetic data is closer

41 Recital 26 of the GDPR.

to de-identified data under Chinese law, except that, as a de-identification method, it replaces real identifiers with virtual ones, and there is no obstacle to regulating such data within the legal framework. If the generation phase of synthetic data involves the collection and use of original data, data protection norms such as informed consent, purpose limitation, and minimization principles are certainly applicable. However, for the use of synthetic data, the law still needs to create a more relaxed protection environment. For instance, data processors could be exempted from their obligations to assist data subjects in exercising rights such as access, rectification, and erasure when directed at synthetic data.⁴²

2. Establishing a Compliance Audit System to Enhance Governance Transparency. — The governance dilemma posed by the algorithmic ‘black box’ to synthetic data significantly constrains its application in high-risk fields such as finance and healthcare. In this context, establishing a comprehensive compliance audit system becomes key to breaking the deadlock. Article 54 of *China’s Personal Information Protection Law* stipulates that ‘Personal information processors shall regularly conduct compliance audits of their personal information processing activities in accordance with laws and administrative regulations.’ This provision provides the basis for constructing a compliance audit system. Based on the need for governance transparency, the compliance audit system should aim for ‘recordability, interpretability, and traceability’. By providing detailed technical documentation, reasonable explanations, and a complete traceability chain, it transforms opaque algorithmic models and data-generation logic into verifiable audit evidence, thereby creating conditions for regulatory accountability and trustworthy application of synthetic data.

To establish a compliance audit system for synthetic data governance, data controllers should first provide regulatory authorities with detailed technical documentation for auditing purposes. According to article 21(3) of *China’s Cybersecurity Law*, the retention period for technical documentation shall be at least six months. The content of technical documentation should at least include the basic parameters of synthetic data, the selected model architecture, known model limitations, and corresponding countermeasures adopted. Second, to

⁴² See Zhang Zesong, *Value Trade-Offs and Normative Reconstruction in Personal-Information Anonymization*, 6 Journal of Nankai University (Philosophy, Literature, and Social) 66, 77 (2025).

achieve the goal of making synthetic data ‘interpretable’ for downstream users, data controllers should be required to provide interpretability reports, explaining to downstream users the purpose of data generation, which core features dominated the data generation, as well as the applicable scenarios and usage contraindications of the data. Furthermore, data controllers need to fulfill the obligations to add identifiable watermarks or effective warning content to synthetic data, ensuring its identifiability and interpretability at the source.⁴³ Finally, under specific circumstances, accredited third-party institutions or individuals should be permitted to audit the compliance and security of synthetic data generation and usage processes, based on technical documentation or interpretability reports, enabling precise identification of problematic links and the determination of responsible parties when security incidents such as privacy leaks or algorithmic discrimination occur.

3. Exploring an Open Source Software System to Promote Governance Socialization. — Open source is a system that makes the source code of technological achievements (represented by software, and extendable to hardware design, etc.) publicly available through specific licensing agreements. Open source enables comprehensive disclosure of model source code, creating a highly transparent collaborative environment. Therefore, many government departments, academic organizations, and non-profit institutions advocate for developing synthetic data algorithms based on open-source models. For example, a non-profit organization in the United States, together with academic institutions, created the open-source health data model Synthea, which has generated over one million synthetic medical records for virtual patients and made them available to public and private users.⁴⁴ Open source is an important pillar for building an open, shared, and secure national integrated data market. In this regard, the government should play a leading role. It may adopt the following measures: First, the government can increase funding to support the development of synthetic data, similar to how it subsidizes other public goods. Second, the government can take the lead in building and maintaining open-source infrastructure

43 See Zhang Linghan, *Logic Update and System Iteration of Deep Synthetic Governance — ChatGPT and Other Generative AI Governance in China*, 3 Science of Law (Journal of Northwest University of Political Science and Law) 38, 48 (2023).

44 See Jason Walonoski, Mark Kramer & Joseph Nichols et al., *Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record*, 25 Journal of the American Medical Informatics Association 230, 232 (2018).

(such as establishing public databases or project-hosting platforms), thereby facilitating deep participation by all relevant parties in the data ecosystem. This promotes data sharing and reuse, providing technical support for open-source collaboration.⁴⁵ Third, the government can support enterprises in launching open-source synthetic data products, allowing them to charge reasonable fees for services related to implementing open-source software, while also granting tax incentives to enterprises that proactively adopt open-source practices.⁴⁶

In terms of synthetic data governance, open source brings numerous advantages. First, open source relies on open collaboration mechanisms, alleviating the supply shortage and access barriers faced by synthetic data as a public good by lowering technical thresholds and reusing development outcomes. Second, open source makes the internal structure of models and the data generation process transparent and traceable by disclosing model source code, further eliminating the ‘black box’ characteristics of synthetic data. Third, open source means users can learn from and improve models, facilitating continuous model optimization and significantly reducing the risk of model distortion or collapse. Although open source provides innovative impetus for generating synthetic data, its legal openness simultaneously poses challenges to the core principles of the intellectual property system, which incentivizes innovation by granting creators exclusive legal rights. In other words, open source may conflict with intellectual property protection, necessitating a balance between open-source technology and rights protection.

Under China’s current laws, synthetic data itself cannot yet be protected as intellectual property. However, models related to synthetic data and their generation processes can fall within the scope of intellectual property protection. For example, models used to generate synthetic data may obtain patent protection. The source code of models, software architecture, and databases aggregating synthetic data are typically protected as objects of copyright or trade secrets. Open source and these intellectual property rights are not irreconcilably opposed; they can complement each other, achieving compatibility between sharing incentives and proprietary incentives. First, open source is primarily led by governments or non-profit institutions, aiming

45 See Cao Yang, *How to Regulate Data Monopoly?*, 1 *Oriental Law* 46, 56 (2025).

46 Peter Lee, *supra* note 9, 49.

to provide enterprises that have limited resources with access to generate and utilize synthetic data for developing and training their own algorithmic models, thereby promoting the socialization of synthetic data governance. Intellectual property, on the other hand, is mostly asserted by for-profit institutions, aiming to eliminate ‘free-rider’ phenomena through the grant of exclusive rights, thereby sustaining the commercial returns and innovative drive of rights holders. Second, the realization of open source relies on the tool of licenses. The effectiveness of a license stems from party consensus. At the same time, granting users the freedom to use, modify, and distribute software also imposes specific obligations, such as retaining attribution and ensuring license compatibility. Intellectual property rights are compulsorily protected by national laws, with their effectiveness derived from legal provisions rather than mutual agreement. The exclusive protection afforded by intellectual property is a necessary prerequisite for open source; without this initial legal protection, any subsequent open-source licensing would be meaningless.⁴⁷ Finally, intellectual property can also play a positive role in promoting the socialization of synthetic data governance. For instance, when applying for a patent, inventors must fully disclose information about the patented achievement to obtain legal protection. This trend was reinforced in *Amgen v. Sanofi*. In its ruling, the court indicated that patent disclosure must go beyond abstract ‘research tasks’ and provide specific, implementable technical details. This will encourage patent applicants to disclose more of the patented technologies used to generate synthetic data.⁴⁸

4. Implementing a Quality Control System to Enhance Governance Robustness. — The widespread application of synthetic data in the AI field and its critical impact on algorithmic models make quality assurance an urgent issue facing the industry. If this problem is not effectively addressed, the inherent privacy risks and quality biases of synthetic data may lead to data leaks and even trigger algorithmic discrimination or model collapse, ultimately undermining the AI industry’s foundation for development. High-quality synthetic data should be complete, accurate, timely, and representative. Any deficiencies in data coverage, accuracy, or update

47 See Zhang Ping & Ma Xiao, *Critique and Compatibility of Open Source Software with Intellectual Property Grey Areas (Part II)* — *A Comparative Study of Open Source Software Licenses*, 2 Science Technology and Law 46, 48 (2004).

48 *Amgen Inc. v. Sanofi*, 598 U.S. 594, 614 (2023).

frequency may directly translate into systemic biases during training. Therefore, it is necessary to implement a quality control system throughout the synthetic data lifecycle. This system defines the characteristics of high-quality data by establishing unified quality standards, introduces a data access mechanism that translates these standards into operational quality inspection processes, and relies on a preset quality traceability mechanism to return unverified synthetic data for correction.

Establishing a quality control system for synthetic data should begin with unified quality standards. National data management authorities can issue relevant quality standards for synthetic data based on specific application scenarios, clarifying indicators for high-quality synthetic data across contexts related to coverage, representativeness, update cycles, and bias control. For example, when generating data for autonomous driving training, quality standards should stipulate coverage of various road conditions (urban, highway, rural), weather conditions, and obstacles, with trajectory distributions conforming to the laws of motion in the physical world, while also incorporating sufficient edge cases (such as pedestrians suddenly crossing the road). A data access mechanism then incorporates quality standards into automated inspection processes, typically designed by model developers who utilize algorithms to create various verification modules for screening indicators. In the aforementioned example, if relevant modules in the automated inspection process detect that synthetic data lacks rainy or snowy weather conditions, or has insufficient proportions of rural road scenes or edge cases, the data will be flagged as ‘unqualified’ and prevented from entering the database. At this point, a quality traceability mechanism centered on ‘precise positioning and closed-loop correction’ is activated. For anomalous data, the quality traceability mechanism identifies key factors that cause quality deviations based on the model version used for generation, the current parameters, and the application scenarios. For instance, incorrect parameter settings in the simulation environment may cause data samples to be overly biased toward urban roads. Subsequently, it automatically generates a quality report with suggested corrections, helping developers optimize algorithms and adjust corresponding parameters. The corrected data will then undergo re-inspection by the data access and traceability mechanisms, ultimately completing the closed loop of quality control.

V. CONCLUSION

As a key technology for deepening the development and utilization of data resources, synthetic data overcomes the limitations of collecting and using real data by simulating real-world scenarios and generating virtual content, driving more efficient data supply and promoting social digital transformation. However, this technology also increases the complexity of data governance, posing challenges to data protection systems built on real data collection models. Although some challenges have already emerged in traditional data governance, synthetic data propels them to new dimensions. Faced with this technological transformation, it is imperative to accelerate the conceptual reconstruction and institutional innovation of the synthetic data governance system, providing clear legal expectations for all parties involved in data generation and use. On this basis, this article adopts standardization, transparency, socialization, and high quality as value orientations, incorporating synthetic data into the scope of protection under current laws and establishing innovative mechanisms encompassing compliance auditing, open-source software, and full-process quality control, ultimately ensuring these value concepts can be transformed into actionable, practical solutions.

(Edited by Zhai Yujia)

中国人工智能合成数据治理的价值取向与制度建构

内容提要 合成数据作为真实数据的安全替代品，在降低 AI 训练成本、规避隐私与版权侵权方面具有显著的优势。尽管如此，在合成数据训练过程中，依然存在泄露个人隐私、加剧算法“黑箱”、助长市场垄断以及导致算法决策与现实脱节等多重风险。合成数据在人工智能领域所引发的技术变革，对我国数据治理法律体系的完善提出了新的要求。未来合成数据的治理应当以规范化、透明化、社会化与高质量化为价值取向，推动形成兼顾技术创新与风险防范的治理框架。在制度建构上，可以将合成数据纳入现行法的规制范围，并为其提供适度宽松的隐私保护，同时依托合规审计、开源软件及质量管控等创新机制，逐步构建既面向未来又具有中国特色的合成数据治理体系。
