Proceedings of ICILLP 2025 Symposium: Property Law and Blockchain Applications in International Law and Legal Policy
DOI: 10.54254/2753-7048/2025.LD29277

Research on the Legal Policy of Data Poisoning from the Perspective of Safety Regulation in Federated Learning

Qingyue Zhang

School of Law, Central South University, Changsha, China
8103230304@csu.edu.cn

Abstract. In the context of the rapid development of artificial intelligence industry, federal learning is the key technology to crack the "data island", but it is vulnerable to data poisoning attacks, which not only causes losses to enterprises, but also causes public service and social governance risks. The existing data security legal framework in China has shortcomings in the regulation of federal learning data poisoning, lacking clear definition, exclusive provisions and a sound regulatory accountability mechanism. Based on the logic of federal learning technology and the risk characteristics of data poisoning, and with reference to the international mainstream regulatory experience, this paper proposes an optimization path from the perspective of system construction: on the one hand, build a full process regulatory mechanism of "pre filing access, in-process dual monitoring, and post traceability accountability" to clarify the filing and monitoring requirements of projects with different risk levels; on the other hand, balance technological innovation and safety prevention and control through the improvement of the legal system, clarification of behavior boundaries and responsibility attribution, formulation of supporting technical standards, and international collaborative governance, so as to provide theoretical reference and practical path for the development of federal learning compliance and the safety governance of AI industry.

Keywords: federated learning, data poisoning, security supervision, legal responsibility, policy improvement.

1. Introduction

With the advancement of artificial intelligence technology, the supply of high-quality training data has become a crucial link in determining the functionality of artificial intelligence products and services, as well as the optimization level of algorithm models [1]. As a distributed machine learning method, federated learning has been widely applied in many fields such as finance, healthcare, and politics by enabling collaborative participation among multiple participants while protecting data privacy. However, its characteristics also provide a hidden space for malicious attacks, among which digital poisoning attacks have gradually become the primary risk threatening the security of federated learning due to their low cost, high concealment, and strong destructive power. Digital poisoning interferes with the training process of global models by injecting malicious samples into local training datasets, tampering with data labels, or transmitting false model parameters, leading to

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2025.LD29277

erroneous decision-making in model outputs. This not only undermines the economic interests of enterprises but also potentially jeopardizes public safety and social order.

From a practical perspective, digital poisoning incidents have been occurring more frequently. In 2023, iFLYTEK faced a brand trust crisis after third-party poisoning introduced content that violated mainstream values, resulting in a market value loss of 12 billion yuan. In 2024, an intern at ByteDance exploited a platform vulnerability to inject destructive code into the training data of a large model, rendering the training results ineffective, and the company sought compensation of 8 million yuan [2]. In the same year, Rick Claypool, research director at a U.S. non-profit consumer rights organization, pointed out that a popular mushroom identification AI tool in North America often mistakenly identified deadly poisonous mushrooms as edible varieties, leading to dozens of people being poisoned and hospitalized [3]. These data poisoning incidents make us think that in the context of federated learning, risks are more diffuse, and digital poisoning has evolved from a technical risk to a real hazard, posing severe challenges to the existing legal policy system and regulatory framework.

Currently, digital poisoning prevention and control in federated learning faces a triple contradiction: First, the disconnection between technical defense and legal regulation. The distributed architecture of federated learning makes it difficult to trace the source of poisoning attacks, and existing laws have ambiguities in regulating such behavior, which is reflected in the conflict between technical "anonymity" and legal "accountability". Second, the conflict between data privacy protection and poisoning detection. The core advantage of federated learning lies in its ability to protect private data. However, due to its non-sharing nature, it also provides cover for poisoning behaviors. How to achieve effective detection without sacrificing privacy has become a difficult problem in institutional design. Third, the balance between innovation incentives and regulatory systems. Excessively strict regulation may inhibit the innovative application of federated learning technology, while a hands-off approach may lead to systemic harm. How to construct an appropriate regulatory framework is a key consideration in policy formulation.

China now places a high emphasis on AI security and data governance, which has also become a major social hot topic. Laws and regulations such as the Cybersecurity Law (PRC), the Data Security Law (PRC) and the Interim AI Services Admin Measures have established a basic framework for data security protection. However, specific regulations targeting digital poisoning in federated learning remain inadequate. On the one hand, the existing laws do not clearly define data poisoning, and there is controversy over the application of criminal charges. Traditional criminal charges, such as the crime of damaging computer information systems, cannot fully encompass distributed data poisoning behaviors. On the other hand, the responsibility allocation mechanism is unclear. There are many participants in federated learning, and the responsibilities and rights of data providers, model developers, platform operators, and other entities are not clearly defined. When poisoning incidents occur, it is often difficult to assign responsibility. In addition, the current regulatory measures are not well-suited to the technical characteristics of federated learning, and a corresponding regulatory system needs to be established.

In this context, studying the legal and policy issues of digital poisoning from the perspective of federal learning security regulation has important theoretical and practical significance. Theoretically, it helps to fill the gap in data security legal research and improve the legal regulation theory in the era of artificial intelligence. In practice, it can provide specific paths for building a federal learning security regulation system, help enterprises prevent poisoning risks, and ensure the safe development of technology. This article will be based on the technical characteristics of federated learning, examine the technical mechanisms and risk transmission pathways of poisoning

DOI: 10.54254/2753-7048/2025.LD29277

attacks, sort out the regulatory dilemmas faced by existing legal policies, and analyze this behavior from an international perspective. Drawing on typical policy documents, it will ultimately propose legal improvement suggestions and regulatory strategies tailored to China's federated learning scenarios, providing legal protection for the healthy development of China's digital economy.

2. Basic theoretical explanation of federated learning and data poisoning

2.1. Technical principles and behavior types of data poisoning

Data poisoning is an attack method against machine learning models. Attackers interfere with the training process of the model by injecting malicious samples or modifying data in the training data of the model, so that it produces incorrect results in the prediction or decision-making stage [4]. Its core technical principle is to utilize the characteristic that machine learning models rely on the distribution of training data, and by disrupting the consistency of the data distribution, we can induce the model to learn incorrect patterns. So far, data poisoning can be simply divided into the following two types: one is the poisoning of training set data, where attackers tamper with the original training data; the other is the poisoning after formally entering the training process, mainly affecting the model parameters to disrupt the model's prediction outcomes, such as transmission layer interference and process vulnerability exploitation.

2.2. The concept and characteristics of federated learning

Federated learning is a fundamental machine learning technique that fully leverages data from multiple institutions or entities for joint modeling, all while ensuring privacy protection [5]. Its core principle is 'data remains stationary, while the model moves', meaning that each participant's data is kept locally, and synchronization is achieved between multiple devices and computing servers by exchanging intermediate computational results (e.g., model parameters, gradients), thereby constructing a shared, high-performance model. Federated learning, as a new type of machine learning model, is characterized by strong privacy protection, decentralization, multi-party participation, and high communication efficiency. At the same time, this also creates more concealed and convenient conditions for data poisoning.

2.3. Hazards and risk evaluation of data poisoning in federated learning

The harm of data poisoning to federated learning can be analyzed from three dimensions: technical, economic, and legal. From a technical perspective, data poisoning can directly damage model performance. More critically, when privacy protection mechanisms are targeted, it may compromise the security of encryption parameters, resulting in the indirect leakage of raw data and violating the original intention of privacy protection in federated learning. From a legal perspective, data poisoning may pose a threat to public safety and affect the public domain; At the same time, it may also infringe upon the legitimate rights and interests of users, leading to disputes over rights and responsibilities. From an economic perspective, data poisoning can cause direct losses to participants in federated learning, with incalculable maintenance costs in the later stages, and may also damage the business reputation of enterprises and other businesses.

In the "Artificial Intelligence Security Governance Framework 1.0" released by China in 2024, data poisoning is listed as an important category of endogenous data security risks in artificial intelligence [6]. With only 0.006% of malicious samples, there is a 50% probability of completing a "data poisoning" attack [7]. At the same time, judging from the current state of China's industrial

DOI: 10.54254/2753-7048/2025.LD29277

development, federated learning has deeply penetrated key fields such as finance, healthcare, and the Internet of Things in terms of its application. Moreover, at the policy level, there is a clear supportive stance, which shows its vast growth potential. Therefore, the threat posed by data poisoning to federated learning should not be taken lightly, and how to regulate and oversee it has become an urgent research topic.

3. Legal definition and existing regulation of data poisoning in federated learning

3.1. Analysis of the legal attributes of data poisoning behavior

Combined with China's current criminal and administrative law, data poisoning may involve multiple legal issues, determined by the attack target, method, and consequences. For example, tampering with or deleting training data that is lawfully owned by enterprises or institutions may constitute the crime of intentional property destruction under Article 275 of China's Criminal Law. Leaking personal information during such poisoning may lead to the crime of infringing on citizens' personal information. Second, if tampering or forging model parameters disrupts a federal learning system, preventing normal training, it may fall under Article 286's provisions on the crime of damaging computer information systems. Third, if data poisoning causes failures in public safety-related models (e.g., traffic, medical, emergency), resulting in casualties or public disorder, it may constitute the "crime of endangering public safety by dangerous means." Additionally, data poisoning may be punished as illegal business operations or sabotaging production and operations [8].

3.2. The current regulatory landscape of laws and policies in China

In China's current laws and policies, the regulatory provisions related to federal learning data poisoning are mainly distributed in the fields of network security, data security, artificial intelligence management and so on, but there are obvious shortcomings: for example, articles 21 and 42 of the network security law of the people's Republic of China focus on regulating "network operation security", and there are no specific regulatory measures for the data poisoning of federal learning, a new distributed architecture, and the protection scope of "data integrity" does not cover special data types such as training data and model parameters. Articles 3 and 21 of the data security law of the people's Republic of China are restricted in principle only through "data security protection obligations", lacking a specific definition of poisoning behavior and legal consequences, resulting in unclear law enforcement standards in practice.

On the whole, the deficiencies of current laws and policies are concentrated in three aspects: first, the scope of regulation is vague, the behavior boundary of data poisoning under the federal learning framework is not clear, and there are no special provisions on the legal protection of training data and model parameters; Second, the regulatory measures are static. The existing provisions require "prior filing" and "subsequent punishment". There is a lack of in-process monitoring mechanism for the dynamic training process of federal learning, which can not prevent and control the risk of poisoning in real time; Third, the punishment is insufficient, the administrative liability is mainly fines and warnings, the deterrent force to large enterprises is limited, and the civil compensation standard is not clear, so it is difficult to fully compensate for the losses of victims [9].

DOI: 10.54254/2753-7048/2025.LD29277

4. Legal definition and existing regulation of data poisoning in federated learning

4.1. Defect analysis of current regulatory mechanism

At present, China's regulatory mechanism for the poisoning of federal learning data has not yet formed a complete system that adapts to the technical characteristics. The core defects are reflected in three aspects: first, the regulatory bodies are scattered, and the rights and responsibilities overlap. Federal learning involves many fields, and the regulatory responsibilities are scattered in the competent departments of cyberspace, industry and information technology, public security and industry, lacking a unified plan; Second, the supervision means are lagging behind, relying on the punishment after the event, and lack of pre prevention and in-process monitoring; Third, the lack of technology adaptation makes it difficult to effectively monitor the distributed training process of federal learning [10].

4.2. Drawing on international regulatory experience

The European Union, the United States, Singapore and other countries or regions have introduced regulatory measures for AI Model Security (including federal learning), which holds significant reference value for experiences in preventing data poisoning: The EU takes "risk prevention" as the core, requiring high-risk AI systems to conduct training data review and risk assessment; The United States has adopted the mode of "regulatory by Sector+accountability after the event", formulated special poison prevention standards in the fields of medical treatment, finance and so on, and put forward the detection indicators of the risk of poisoning by federal learning, providing technical reference for regulation; Singapore has adopted a 'technology-oriented, guidance-first' approach and released a federal learning safety technology guide, distilling actionable insights such as dynamic supervision, a list of main responsibilities, and prioritizing technical standards [11, 12].

From a comprehensive perspective, the core insights from international experience are as follows: First, adopting a risk-oriented approach, differentiated regulatory requirements should be formulated to address the poisoning risks in different areas of federated learning. Second, emphasis should be placed on in-process monitoring, utilizing technical tools to monitor the training process in real time. Third, it is crucial to clarify the main responsibilities and establish a clear accountability and traceability mechanism. Fourth, a balance should be struck between innovation and safety, encouraging enterprises to comply through means such as regulation and certification programs.

4.3. Suggestions on improving China's regulatory mechanism

A whole-process supervision system covering "before, during, and after" can be established.

4.3.1. "Pre-event": establish a filing and access mechanism

First, the classified filing system can be filed with the corresponding regulatory authority according to the risk level. Some high-risk projects should be double filed with the industry competent department and the Internet Information Department, and the poison risk assessment report and prevention plan should be submitted; Low and medium risk projects only need to be filed with the Internet Information Department to simplify the filing process. The second is to set access standards, clarify the requirements of the participants in federal learning, and prohibit unqualified subjects from participating in high-risk projects; At the same time, the training data quality standard shall be formulated, and the data that does not meet the standard shall not be accessed to the training.

DOI: 10.54254/2753-7048/2025.LD29277

4.3.2. "During-event": promote "technology+manual" dual monitoring

The first is to build a federal learning supervision platform, led by the Internet Information Department and jointly developed with the competent industry departments to develop a unified supervision platform. Federal learning projects need to be connected to the platform and upload training data quality reports and other relevant contents for supervision. At the same time, the platform has a built-in poison risk detection algorithm, which can automatically identify abnormal data and abnormal parameters, and give real-time warning to regulators and project participants. The second is the regular safety assessment. The regulatory authority carries out manual assessment on high-risk federal learning projects every quarter, and invites technical experts to test and verify; Low and medium risk projects are evaluated once a year and can be entrusted to a third party.

4.3.3. "Post-event": improve the accountability and traceability mechanism

The first is to establish a traceability platform for data poisoning events. After the poisoning events, the regulatory authorities can quickly trace the toxic data injection nodes and Parameter Tampering links through the platform to clarify the responsible parties. The second is to refine the punishment measures and formulate differentiated punishment for different responsible subjects: data providers who intentionally poison can be imposed corresponding fines, and the data processing qualification will be revoked if the circumstances are serious; The external attackers shall be investigated for criminal responsibility according to law, confiscate their illegal gains and impose a high fine. The third is to establish a risk notification mechanism. After the handling of the poisoning incident, the regulatory authority will issue a risk notification to the industry, disclose the poisoning means and preventive measures, and avoid the recurrence of similar incidents.

5. Legal and policy improvement path of data poisoning in federated learning

The first is the optimization of the legal system. Now there are two different views in the academic community. One is that there is no need to set up a new crime, and it should be incorporated into the existing criminal law system and strictly prosecuted; The second is to improve the legal system from two aspects of legislative revision and special legislation in view of the blank of the current law on data poisoning regulation. For example, the "data poisoning" clause was added in the revision of the data security law of the People's Republic of China to clarify its definition and legal consequences; The federal learning safety management regulations can also be formulated to refine the safety requirements for the whole process of data collection, training and use. At the same time, it should be noted that legislation should only intervene when necessary, and justice should avoid consequentialism and abuse of charges [13].

The second is the refinement of supporting policies. Under the legal framework, supporting policies such as technical standards, certification systems and financial support need to be adopted: The first is to issue technical standards and unify prevention norms; The second is to establish a safety certification system and strengthen incentives and constraints. A federal learning safety certification system can be established and implemented by a third-party organization. The certification content includes poison prevention measures, emergency response ability, etc. at the same time, policy support will be given to the federal learning projects that have passed the certification. The third is to set up a special fund to support technology research and development.

Finally, international collaborative governance needs to improve the rules of cross-border data flow to prevent cross-border risks caused by data poisoning; Participate in global AI security

DOI: 10.54254/2753-7048/2025.LD29277

governance, and promote the establishment of data poisoning risk information sharing mechanism. Learn from the General Data Protection Regulation (GDPR) of the European Union, and clarify the responsibility identification rules for data poisoning in cross-border federal learning.

6. Conclusion

As the key technology architecture in the era of artificial intelligence, federal learning cannot develop safely without the protection of laws and policies. Data poisoning, as the core risk in federal learning, is not only facing the defense challenge at the technical level, but also the lack of legal regulation and the lag of regulatory mechanism. Based on this, the solutions proposed in this paper clarify the legal attributes of data poisoning, build a dynamic regulatory model, and fill the gaps in the current regulation. On the one hand, it is necessary to refine the scope of legal protection through legislation and refine the current legal framework for data poisoning; On the other hand, we need to rely on technical means to achieve regulatory upgrading, use technology to establish a whole process risk monitoring mechanism, promote cross regional and cross industry regulatory coordination, and strengthen the mandatory binding force of industry standards. In the future, with the continuous iteration of federal learning technology, data poisoning methods may take on more complex forms, which requires that laws and policies should be paid attention to at all times and adjusted in time: In fostering the development of federated learning, it is imperative to reserve reasonable space for technological innovation and prevent excessive regulation from stifling industrial vitality. Concurrently, the security defense line for federated learning must be strengthened through the dual safeguards of "technology and governance." Only by striking a balance between innovation and safety, and by clearly defining rights, responsibilities, and remedial measures, can the healthy development of federated learning be promoted within a rule-of-law framework. This approach provides solid institutional support for the high-quality development of China's artificial intelligence industry.

References

[1] Zhao Jingwu. (2025). On the institutional construction of high-quality supply of AI training data. China Law Review, (1), 92–104.
[2] Li Ling, Huang Liling, & Yang Liu. (2024, November 27). Former intern alters code to attack large model training, ByteDance sues for 8 million in damages. SECRSS. Retrieved August 20, 2025, from https: //www.secrss.com/articals/72821
[3] Rick Claypool, Mushrooming Risk: Unreliable A.I. Tools Generate Mushroon Misinformation, public citizen, March 18, 2024
[4] B. Biggio, G. Fumera, F. Roli, (2008) Adversarial Pattern Classification Using Multiple Classifiers and Randomisation , "Joint IAPR International Workshop, Orlando, USA, Proceedings , Springer Berlin Heidelberg.
[5] Liu Zegang. (2025). The limitations and solutions of privacy protection in federated learning in the age of artificial intelligence. Peking University Law Journal, 37(1), 65–84.
[6] National Cybersecurity Standardization Technical Committee. (2024). Artificial intelligence security governance framework 1.0 [EB/OL]. Retrieved October 11, 2024, from https: //www.tc260.org.cn/front/postDetail.html? id=20240909102807
[7] Xu C, Wang J, Tang Y Q. et al. A targeted attack on blackbox neural machine translation with parallel data poisoning [J/ OL]. Computation and Language, 3638-3650 (2021-02-15) [2023-12-14]. https: //arxiv.org/abs/2011.00675.
[8] Sun Benxiong, & Lin Yujie. (2025).The dilemma and solution of criminal law regulation on data poisoning behavior. Journal of Kunming University of Science and Technology (Social Sciences).
[9] Ma Runyi. (2025). Risks and regulatory pathways of "data poisoning". Science and Technology Management Research, 45(8), 156–169.

DOI: 10.54254/2753-7048/2025.LD29277

[10] Sun Qingbai. (2024). On the normative construction of risk governance for AI large model training data. E-Government, (12), 41–52.
[11] Custers B, Sears A M, Dechesne F, et al. (2019) EU Personal Data Protection in Policy and Practice [M]. Cham: Springer.
[12] Li Hong, & Zhao Mengke. (2024, June 6). U.S. NIST releases Cybersecurity Framework (CSF) 2.0. Retrieved August 11, 2024, from https: //casisd.cas.cn/zkcg/ydkb/kjzcyzxkb/2024/zczxkb202404/202406/t20240606_7185729.html [in Chinese]
[13] Xu Yunjie. (2025). The application of charges and criminal liability for data poisoning. Tianfu New Review, (4), 38–48+152.