International Journal of Ethical AI Application
eISSN: 3093-7736
https://ijeaa.cultechpub.com/index.php/ijeaa

Privacy-Preserving Weight Reconstruction in Federated Learning: A Legal Governance Framework Based on Algorithmic Justice

Zikang Zhang

Hebei University of Economics and Business, No. 47, Xuefu Road, Xinhua District, Shijiazhuang City, Hebei Province, China

Email: zikang131@gmail.com

Abstract

This study takes the "right to be forgotten" established by the European Union's General Data Protection Regulation (GDPR) and the compliance requirements of China's Personal Information Protection Law as its theoretical basis. It comprehensively employs a cross-disciplinary research method combining normative analysis and technical deconstruction to deeply explore the legal dilemmas inherent in the technical architecture of federated learning. The study reveals that in the operation of the federated learning system, there is a profound legal conflict between "algorithmic shadowing" and data control rights. Traditional privacy rights theories face increasing limitations in their explanatory power when confronted with this new data processing model. Based on this, this paper innovatively proposes a "dynamic responsibility allocation" legal governance paradigm. By precisely defining the rights and obligations of various participants in the federated learning ecosystem, such as data providers, model trainers, and platform operators, this paper constructs a legal governance framework that meets the requirements of algorithmic justice. The framework aims to achieve a dynamic balance between technological innovation and privacy protection, providing a solid theoretical foundation and feasible institutional guidance for the compliant and robust development of federated learning technology.

Keywords

Federated Learning, Privacy Weighting, Algorithmic Justice, Legal Governance Framework, Dynamic Responsibility Allocation

1. The Root Cause of the Problem: The Legal Deviation of Federated Learning

1.1 Rights Gaps in the Technical Architecture

Federated learning technology relies on a global model aggregation mechanism to achieve collaborative training of data from multiple parties. Its core mathematical expression is $M_{fed} = \frac{1}{n}\sum_{i=1}^{n}\Delta w_i$. From the perspective of interdisciplinary research in legal technology, this technical mechanism, while activating the flow of data value, quietly triggers a structural erosion of data subjects' rights. Traditional privacy rights theory regards personal data as a natural extension of personality rights, granting data subjects absolute control over their personal data. This right is most clearly articulated in the "right to be forgotten" provision of the EU's General Data Protection Regulation (GDPR) Article 17—data subjects have the right to request data controllers to delete their personal data under specific conditions to achieve ultimate control over their personal information [1].

However, in the technical context of federated learning, the data subject's control rights face unprecedented fundamental challenges. When user data participates in gradient calculations and is integrated into the global model, the data is not stored in its original form but is transformed into components of the model parameters. After iterative updates, these parameters irreversibly embed data traces into the algorithmic structure, forming a "data gene"-like existence. This technical characteristic creates an irreconcilable institutional contradiction with the "right to be forgotten," which requires complete data deletion: even if data subjects assert their right to deletion, it is technically impossible to completely erase the training traces already integrated into the model [2]. For example, in a medical data federated learning scenario, even if a patient requests the deletion of their health data, the disease feature parameters already absorbed by the model will continue to influence subsequent prediction results, leading to a practical dilemma where the data subject's rights are exercised in a "formally legal but substantively ineffective" manner.

1.2 The Failure of the Compliance Framework

The current personal information protection rules centered on "notification-consent" expose significant institutional adaptability defects when addressing the local gradient update mechanism of federated learning $\nabla w_i = f(D_i)$. The

local gradient update process involves complex matrix operations, backpropagation, and other mathematical logic and technical processes. These technical terms and details far exceed the cognitive scope of ordinary users. When companies fulfill their notification obligations through standardized terms, users often struggle to truly understand the specific content, risk level, and potential consequences of data processing, rendering the notification process a one-way transmission of information lacking substantive communication. The consent mechanism is thus reduced to an ineffective form of "click-to-agree."

More seriously, the algorithmic shadow effect ($M_{fed}^{t+1} \leftarrow M_{fed}^{t} + \eta \nabla w_i$) further exacerbates the dilemma of rights remedies. This effect refers to the situation where, during the iterative updating of a model, parameters formed by historical data training continue to influence subsequent model outputs. Even if data subjects legally exercise their right to delete data, the data training traces solidified in the algorithmic model persist in the form of parameter weights, continuing to participate in model optimization and predictive decision-making. For example, in the scenario of federated learning for financial credit scoring models, even if a borrower deletes some sensitive financial data, the risk assessment parameters formed through historical training will still influence the credit scores of subsequent users, rendering the legally prescribed data deletion right and remedial mechanisms unable to achieve their intended institutional functions, thereby creating a governance paradox of "rights on paper" [3].

2. Theoretical Reconstruction: The Shift of Algorithmic Justice in Privacy Rights

2.1 Expansion of the Legal Object

Traditional privacy rights theory has long confined the scope of protection to the static domain of data itself. In the dynamic data processing scenario of federated learning, these limitations become increasingly evident. In federated learning, data undergoes continuous interaction and computation, giving rise to new value beyond the original data. Research has confirmed that gradient parameters ∇w_i pose the risk of reconstructing original data through reverse engineering attacks D_i, indicating that the outputs generated by algorithmic operations also contain user privacy information. In light of this severe situation, this paper innovatively proposes the theory of "derived privacy rights," advocating that data derivatives such as algorithm models and gradient parameters be included as objects of legal protection. By expanding the regulatory boundaries of privacy rights, this paper constructs a rights protection system tailored to the characteristics of data processing in the digital economy era, thereby establishing a robust legal framework to safeguard user privacy.

2.2 New Mechanisms for Power Balancing

In the field of algorithmic power regulation, the synergistic application of privacy-enhancing technologies (PETs) and secure multi-party computation (MPC) has opened up a new technical pathway. Based on the technical logic of " $\mathrm{PETs} \cap \mathrm{MPC} \Rightarrow \mathrm{AlgorithmicTransparency}\ \tau$ ", the organic integration of the two can significantly enhance the explainability and supervisability of algorithmic decision-making [4]. Specifically, this technical combination makes the algorithmic process more transparent, enabling regulators and users to understand how algorithms process data and make decisions. This paper proposes the establishment of "technical compliance" legal presumption rules, using the Trusted Execution Environment (TEE) technical architecture as a benchmark, to mandate that data processors adopt technology combinations compliant with statutory standards, thereby transforming abstract algorithm transparency metrics into concrete legal obligations. Through this approach, effective constraints and checks on algorithmic power can be achieved, ensuring fairness, impartiality, and legality in data processing.

3. Core Innovation: Dynamic Responsibility Allocation Framework

3.1 Three-Dimensional Reconstruction of Legal Relationships

Table1. Three-Dimensional Legal Relationship Configuration in Federated Learning Ecosystem

Subject Categories	Subject of Rights	Content of Rights and Obligations	Legal Basis
Data Subject	Raw Data D_i	Right to request an explanation of the algorithm, right to data portability	GDPR Article 22, China's PIPL[5]
Data Processor	Gradient parameters ∇w_i	Privacy impact assessment obligation, algorithm security safeguards obligation	GDPR Article 35, PIPL Article 55
Supervisory authority	Global Model M_{fed}	Algorithm audit rights, compliance supervision and inspection rights	EU DSA Act, China's AIGC regulations

This study constructs a three-dimensional legal relationship model of "data subject-data processor-regulatory authority" to clarify the rights and obligations of all parties in the federated learning ecosystem(table1). Data subjects enjoy core rights over their original data based on the extension of their personality rights; data processors bear the obligations of risk prevention and compliance assurance during data processing; regulatory authorities exercise algorithm audit rights

to achieve full lifecycle supervision of the global model, forming a legal relationship system with clear responsibilities and mutual checks and balances.

3.2 Responsibility Allocation Formula

Based on risk society theory and the technical governance paradigm, this paper constructs a dynamic responsibility allocation formula: $\mathcal{L}_{legal} = \alpha \cdot$ Data Sensitivity $+\beta \cdot$ Algorithmic Complexi ty$+\gamma \cdot$ Number of Participants . Among them, the coefficients α , β , and γ are quantified and determined through a privacy risk assessment matrix. This formula converts technical variables such as data sensitivity, algorithm complexity, and the number of participating parties into legal responsibility weights, achieving the scientific and dynamic allocation of responsibilities and providing operational quantitative standards for judicial practice [6].

3.3 Innovation Points Diagram

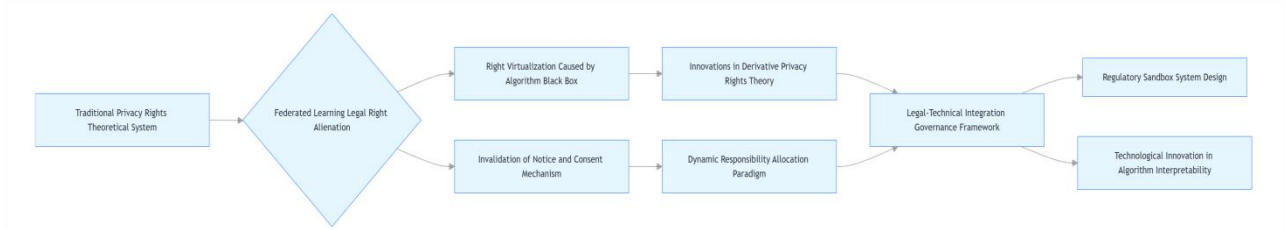

Figure 1. Innovation Points Diagram

This study profoundly reveals the adaptability crisis faced by traditional privacy rights theory in the context of federated learning. In this new data processing model of federated learning, where data flows and integrates dynamically in the form of model parameters, the traditional approach of confining the scope of protection to static data itself can no longer cope with the privacy challenges brought by the derivative value of data. To address this, the research innovatively proposes the theory of "derived privacy rights," which includes data derivatives such as algorithm models and gradient parameters into the scope of legal protection. This breaks through the boundaries of the original subject of rights and extends the reach of privacy protection to the entire process of data processing and all types of outputs. On this basis, the rights and responsibilities system is reconstructed through a dynamic responsibility allocation paradigm. By constructing a quantitative responsibility allocation formula based on factors such as data sensitivity, algorithm complexity, and the number of participants, it achieves precise definition and dynamic adjustment of the rights and obligations of all parties involved (data subjects, data processors, and regulatory authorities).Ultimately, the innovations in the above theories and paradigms together contribute to the formation of a three-in-one legal-technical integration governance framework of "prevention-regulation-remedy." In the prevention phase, measures such as establishing legal verification agreements and implementing identity anonymization are adopted to reduce privacy risks from the source. In the regulation phase, mechanisms like regulatory sandboxes and algorithm audits are used to effectively supervise the entire process of federated learning. In the remedy phase, algorithm restoration technology is introduced to provide practical remedies when the rights of data subjects are infringed. This framework offers a systematic institutional innovation path for the compliance issues of federated learning in the development of the digital economy, achieving a dynamic balance between technological innovation and privacy protection (see Figure 1).

4. Institutional Implementation: Legal-Technical Integration Pathways

4.1 Procedural Regulation via Gradient Updates

To address the challenge of implementing the "right to be forgotten" in federated learning, this study innovatively introduces "algorithm restoration" technology as an institutional innovation tool. Its core technical principle follows the formula Degree of Implementation of the Rightto Erasure $\propto \frac{1}{\tau \cdot \left|\nabla w_i\right|_2}$, where the effectiveness of the right to deletion is positively correlated with algorithm transparencyτ and inversely proportional to gradient norm $\left|\nabla w_i\right|_2$. Specifically, by systematically enhancing algorithm transparency (τ), the system discloses critical information such as algorithm execution logic and data processing workflows to data subjects and regulatory authorities, enabling them to clearly understand data usage. Concurrently, advanced technical methods are employed to precisely control the gradient norm ($\left|\nabla w_i\right|_2$), thereby limiting the possibility of data residue during model training and effectively reducing data residue risks, thereby safeguarding data subjects' "right to be forgotten" [7].

In addition, to further strengthen data privacy and security, this paper recommends establishing a legal verification agreement prior to gradient sharing. Drawing on the mature technical architecture of the FATE framework, a comprehensive regulatory mechanism covering data legality review, compliance verification, and risk assessment should be established. During the data legality review stage, the legality of data sources is strictly verified to ensure the legitimacy of the data acquisition process. During the compliance verification stage, the compliance of data processing

activities is comprehensively reviewed against relevant data protection regulations and industry standards. The risk assessment stage involves scientific prediction and quantitative analysis of risks such as privacy leaks that may arise from data sharing, providing a solid procedural guarantee for data privacy and security [8].

4.2 Regulatory Sandbox System Design

Based on the theory of privacy impact assessment (PIA), this paper has carefully designed a four-stage review regulatory sandbox system aimed at constructing a full-cycle, multi-level risk control system.

First tier: Identity anonymization protection: By applying k-anonymization technology, this ensures that each record in the data set is indistinguishable from at least k-1 other records, i.e., k-anonymity ≥ legal standards. This technology uses data processing methods such as generalization and suppression to effectively conceal the identity information of data subjects, ensuring data usability while preventing re-identification of data subjects and reducing privacy leakage risks from the source.

Second Tier: Algorithm Discrimination Prevention: Utilize the bias detection formula from the " $\left|\nabla w_i - \mathbb{E}(\nabla w\right| < \delta$ " to monitor gradients generated during the training of federated learning models in real time. This formula calculates the difference between each participant's gradient (∇w_i) and the global average gradient ($\mathbb{E}(\nabla w$), and compares it with a pre-set threshold (δ). If the difference exceeds the threshold, it indicates a potential risk of algorithmic discrimination, triggering an immediate warning mechanism to enable timely intervention measures and ensure the fairness and impartiality of data processing.

Third stage: Exit rights protection: Establish a comprehensive participant exit protection mechanism, clearly stipulating that when a participant requests to exit, all parties involved must strictly follow the established procedures to properly handle the data related to that participant, including data deletion and anonymization, to ensure that the data privacy of the exiting participant is not disclosed. At the same time, provide effective remedies for exiting participants, such as establishing dedicated complaint handling channels and compensation mechanisms, to protect their legitimate rights and interests.

Fourth stage: Risk simulation and pre-control: Conduct damage simulation assessments, use advanced risk assessment models and simulation technologies to simulate and analyze various privacy risks that may arise during federated learning. By pre-setting different risk scenarios, analyze the likelihood of risk occurrence and the potential consequences of damage, and develop targeted emergency response plans in advance. Once actual risks occur, emergency response plans can be quickly activated to minimize losses and achieve dynamic and forward-looking control of privacy risks [9].

5. Rights Scenario-Based: Typical Application Verification

5.1 Medical Joint Research Scenario

In the medical data federated learning scenario, patient diagnostic data D_{health} includes highly sensitive information such as medical records, $\mathcal{L}_{legal} = \alpha \cdot$ Data Sensitivity $+\beta \cdot$ Algorithmic Complexi ty$+\gamma \cdot$ Number of Participants cause detection, and diagnostic images, which are directly linked to personal identity and health privacy. Based on the responsibility allocation formula, the sensitivity coefficient α is set to 0.9 as a high weight, reflecting the severe ethical and legal consequences that may result from the leakage of such data. In practice, a three-tier informed consent mechanism should be established: during the data collection phase, a dynamic risk disclosure statement should be used to clearly explain the purpose of data use, storage periods, and potential risks; during the model training phase, selective authorization should be introduced, allowing patients to specify the scope of data participation in training and the types of algorithms used; and at the application stage, patients must provide secondary confirmation of their authorization for data use scenarios (e.g., academic publication, commercial collaboration). By differentiating the content of disclosures and consent processes, this approach can meet the needs of medical research for integrated analysis of multi-source data while effectively safeguarding patients' privacy and self-determination rights.

5.2 Financial Risk Control Scenarios

M_{credit} ∇w_{bank} s directly impact consumers' credit opportunities and costs, and the transparency of their decision-making process is crucial for market fairness and consumer rights. Therefore, the algorithm transparency metric (τ) must be set to a strict standard of ≥0.75. Regulatory authorities should establish a penetrating audit system to monitor gradient data generated during federated learning in real time. Specific measures include: requiring financial institutions to regularly submit gradient update logs and conduct traceable audits of parameter changes during model training; deploying smart contracts to automatically alert for abnormal data fluctuations; and establishing algorithm sandbox environments to simulate model output results under different data inputs to verify the absence of algorithmic discrimination based on gender, race, or other dimensions. Through these mechanisms, risks such as data misuse and algorithmic black boxes can be systematically mitigated, thereby maintaining fair competition in financial markets and protecting consumers' legitimate rights and interests.

6. Conclusion

This study achieves multiple breakthroughs in legal theory, institutional innovation, and governance paradigms, providing a comprehensive response to the challenges of privacy protection and legal governance in federated learning:

In terms of legal theory, the traditional privacy protection paradigm is inherently static, with its focus primarily on the initial collection and storage of personal data. This approach, however, is increasingly inadequate in addressing the dynamic processing and value derivation of data by algorithms in federated learning, where data continuously interacts, transforms, and generates new forms of value throughout the model training lifecycle. To bridge this gap, this study innovatively proposes the concept of "algorithm-derived privacy rights," which expands the scope of protected objects beyond raw data to include algorithm outputs, intermediate calculation results, and new types of rights generated through data aggregation. By introducing a dual analytical perspective of "data life cycle-algorithm action chain," the study further constructs a new set of rights, including data input rights (enabling data subjects to control how their data enters the algorithmic process), algorithm intervention rights (allowing timely adjustments to algorithmic operations when privacy risks arise), and result control rights (empowering subjects to oversee how algorithmic outputs impact their interests). This theoretical framework systematically enhances privacy rights theory for the digital age, effectively filling the gap in rights protection within the context of algorithmic black boxes.

At the institutional level, the study addresses the long-standing dilemma of blurred responsibility boundaries among multiple stakeholders in federated learning—including data subjects, algorithm developers, and data aggregators—by constructing a dynamic responsibility allocation mechanism. This mechanism quantifies key technical factors to determine responsibility shares: data sensitivity (encompassing data categories and their inherent sensitivity levels), algorithm complexity (such as the number of model layers and parameter scales), and data contribution (including the proportion of data provided by each participant and the importance of their data features in model training). Based on these factors, the study establishes a practical calculation model for a "responsibility coefficient," which integrates data sensitivity weight, algorithm complexity coefficient, and data contribution index. By converting technical parameters into clear, operational legal responsibility standards, this model achieves precise alignment between technical compliance and legal requirements, offering a scientific and quantitative basis for liability determination in judicial practice and ensuring that each participant’s responsibilities are proportionate to their role in the federated learning process.

In terms of governance paradigms, the study breaks away from the traditional single "post-event accountability" model and develops a dual-track governance mechanism of "regulatory sandbox prevention-algorithm restoration relief"[10]. In the prevention phase, a dedicated regulatory sandbox for federated learning is established to allow enterprises to conduct technological innovation within a controlled environment. This sandbox embeds critical compliance modules, such as rigorous data compliance reviews (to verify the legality of data sources and processing) and algorithm explainability tests (to ensure transparency in decision-making logic), thereby identifying and mitigating risks before they materialize. In the remedy phase, algorithm reverse engineering tools are developed to trace and restore model training data and parameter weights, providing concrete technical evidence to support dispute resolution in cases of data infringement. This mechanism deeply integrates technical governance and legal regulation, forming a closed-loop governance system that spans the entire lifecycle from incentivizing innovation to preventing and controlling risks.

The innovations outlined above not only provide a theoretical paradigm for constructing rights systems in emerging technology scenarios but also offer precise institutional tools for applying legal principles in federated learning contexts. Moreover, they contribute practical examples to the advancement of digital economy governance paradigms, balancing technological progress with privacy protection.

This study has established a solid groundwork for privacy-preserving legal governance in federated learning. However, to further strengthen its resilience and broaden its scope of application, several promising future research directions emerge. Given that federated learning often involves multi-party collaboration across national borders, yet privacy laws and regulatory frameworks vary significantly between jurisdictions (e.g., the EU’s GDPR, China’s PIPL, and other regional regulations), future research could explore how to harmonize dynamic responsibility allocation mechanisms with cross-border data flow rules, developing interoperable governance standards that respect jurisdictional differences while ensuring consistent privacy protection. Although this study focuses on legal and technical governance, federated learning also raises ethical questions—such as ensuring fair representation of marginalized groups in model training or avoiding discriminatory outcomes embedded in gradient parameters—so future work could integrate ethical impact assessments into the regulatory sandbox, developing metrics to quantify and mitigate ethical risks alongside legal compliance. As federated learning evolves (e.g., with the rise of federated learning on edge devices or integration with blockchain), new privacy risks and governance challenges will emerge, meaning research should track these technological advancements, updating the dynamic responsibility allocation mechanism and governance tools to adapt to novel scenarios, such as decentralized model aggregation or real-time data processing at the edge. While this study defines rights and obligations for data subjects, processors, and regulators, the practical implementation of these roles could be strengthened through more inclusive participation, so future research might explore participatory governance models—such as involving civil society organizations or technical experts in algorithm audits—to enhance transparency and accountability. By pursuing these directions, research can further refine the legal-technical integration framework,

ensuring that federated learning continues to drive innovation while upholding the principles of algorithmic justice and privacy protection.

References

[1] European Union. (2018). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5. 2016, pp. 1–88).

[2] Gu Yuhao, Bai Yuebin. Research Progress on Security and Privacy in Federated Learning Models [J]. Journal of Software, 2023, 34(6): 2833-2864.

[3] Bai Jinlong, Cao Lifeng, Wan Jiling, et al. Research Progress on Blockchain Privacy Protection Technologies [J]. Computer Engineering and Applications, 2025, 61(02): 19-36.

[4] Wu Hong, Zhao Chang. From Empowerment to Governance: China's Response to Digital Privacy Protection [J]. Seeking Truth, 2025(2): 68-80. Chen Lei, Liu Wenmao. Frontiers and Applications of Data Security Technology from a Compliance Perspective [J]. Frontiers of Data and Computing Development, 2021, 3(3): 19-31.

[5] National People's Congress of the People's Republic of China. (2021). Personal Information Protection Law of the People's Republic of China (PIPL).

[6] He Wen, Bai Hanru, Li Chao. Exploring Enterprise Data Sharing Based on Federated Learning [J]. Information and Computers, 2020(8):173-176.

[7] Qin Peng, Shanguan Lili. Federated Learning for New Applications in Privacy Computing [J]. China Telecommunications Industry, 2024(2):77-80.

[8] Liu Yixuan, Chen Hong, Liu Yuhan, et al. Privacy Protection Techniques in Federated Learning [J]. Journal of Software, 2022, 33(3): 1057-1092.

[9] Zheng Zhifeng. Privacy Protection in the Era of Artificial Intelligence [J]. Legal Science (Journal of Northwest University of Political Science and Law), 2019(2):51-60.

[10] Deng Jianpeng, Zhao Zhison. The Breakthrough and Transformation of DeepSeek: On the Regulatory Direction of Generative Artificial Intelligence [J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition), 2025, 46(04): 99-108. DOI: 10.14100/j.cnki.65-1039/g4.20250214.001.