

Research on the Trust Black Box and Governance Reconstruction of Data Audit Accountability under the Privacy Computing Framework

Lai Huizhen

School of Business, Nanfang College Guangzhou, Major in Big Data
Management and Application, Guangzhou, 510970, China;

Keywords: Privacy computing; Data audit; Trust black box; Verifiable computation; Governance reconstruction Multi; Degree; Freedom;	ABSTRACT Privacy computing technology achieves data that is “usable but not visible” through approaches such as federated learning and secure multi-party computation, providing key technical support for breaking data silos and unlocking the value of data elements. This paper, however, argues that while “invisibility” dissolves traditional privacy risks, it also endogenously creates a new governance paradox: the encryption and distributed processing of data cause audit traceability, compliance verification, and accountability mechanisms based on data readability to fail comprehensively, giving rise to a “trust black box” that cannot be effectively penetrated. In the distributed architecture of federated learning, the central aggregator cannot directly verify whether each client correctly generates model updates according to the specified algorithm; this technical feature constitutes the starting point of the audit dilemma and gives rise to a self-reinforcing cycle of three intertwined dilemmas—traceability failure, compliance failure, and ambiguous accountability. Building on this, the paper proposes that the key to resolving the trust black box lies in
---	--

	driving a fundamental transformation of the audit paradigm from one “based on data visibility” to one “based on process verifiability,” and in constructing a three-dimensional governance framework centered on the embedding of verifiable computation, third-party algorithmic auditing, and institutional accountability, so as to achieve a dynamic balance between privacy protection and effective governance.
--	---

1. Introduction

Between the exponential growth of the global data volume and the release of data value lies a fundamental governance challenge: increasingly stringent requirements for privacy and security protection mean that dispersed and heterogeneous parties cannot freely share data (Li Shuyuan et al., 2022), and massive amounts of data remain scattered across mutually isolated institutions and systems, forming “data silos” that are difficult to overcome. The deep root of this dilemma lies in the non-rivalrous and easily replicable nature of data—centralized aggregation amplifies the risk of leakage and misuse, while isolation prevents the release of data value at the macro level. The classic proposition of data governance can therefore be stated as: how can the effective circulation and aggregation of data value be achieved while safeguarding data privacy and security?

Privacy computing—particularly federated learning and secure multi-party computation—offers a breakthrough technical response to this proposition. McMahan et al. (2017) point out that modern mobile devices hold large amounts of data suitable for training models that could greatly improve on-device user experience; however, this rich data is often privacy-sensitive, large in volume, or both, which can preclude logging it to a data center and training there using conventional approaches. They therefore advocate an alternative approach in which training data is distributed across mobile devices and a shared model is learned by aggregating locally computed

updates, terming this decentralized approach federated learning. As an emerging technology in the field of artificial intelligence, federated learning simultaneously addresses the problems of “data silos” and privacy protection, uniting dispersed data holders to train a global model while keeping each party’s data local (Xiao Xiong et al., 2023). By exchanging model parameters rather than raw data between clients and the parameter server, federated learning effectively protects data privacy and security (Jiang Weijin et al., 2026). It thereby achieves data that is “usable but not visible”—a technical route that precisely responds to the long-standing tension between data privacy protection and value release, and one that is widely regarded as the key infrastructure for breaking down data silos.

However, any technical solution that resolves an existing problem also tends to endogenously generate new governance challenges. As data and computing power move toward the network edge, “the development of artificial intelligence applications increasingly relies on privacy-sensitive user data,” and federated learning, owing to its emphasis on privacy protection, has gradually become a widely adopted distributed machine-learning framework. At the same time, research on horizontal federated learning spans multiple fields—machine learning, distributed systems, wireless communication, and information security—and exhibits diversity from research motivation to technical approach (Wu Wentai et al., 2025). Under a distributed training architecture, the central aggregator cannot verify whether each participant correctly generates local model updates according to the specified algorithm. While this feature dissolves the privacy risks inherent in traditional data sharing, it brings an equally important problem to the surface: when auditors cannot “see” the raw data, how can the correctness of the computation process be confirmed? How can the authenticity of privacy-protection commitments be verified? How can accountability be traced in multi-party collaboration?

The effective operation of auditing, in institutional logic, depends on a self-evident premise—“visibility”: auditors must be able to access the behavioral traces and evidentiary results of the audited party in order to make compliance

judgments. Privacy protection and transaction-data auditing are two relatively conflicting requirements among blockchain-system stakeholders (Gai Keke et al., 2025); verification requires transparent access, while privacy protection requires information concealment—a tension that is especially pronounced in encrypted-computing environments. Under the privacy-computing framework, this conflict is pushed to its extreme: data is encrypted and stored in a distributed manner, and the visibility premise of auditing is systematically dissolved at the technical level. Although combining verifiability with privacy-preserving methods offers clear benefits, certain challenges remain before this combination can be widely adopted in practice (Bontekoe et al., 2025). Although technical explorations combining the two have emerged, a notable gap remains between technical feasibility and institutional auditability. This forms the core paradox of data governance in the era of privacy computing.

Existing research on privacy computing has mostly concentrated on technical dimensions such as algorithmic security proofs and efficiency optimization. As an emerging technology that builds machine-learning models using distributed training datasets, federated learning “can effectively address the problem of local data privacy leakage caused by joint modeling among different data users”; however, “existing federated learning systems have been shown to face potential threats at the data-collection, training, and inference stages, endangering both data privacy and system robustness” (Gao Ying et al., 2023). Only a small number of studies have begun to touch on the design of verifiable computation and audit mechanisms. Bontekoe et al. (2025) analyzed existing solutions that combine verifiability with privacy-preserving computation on distributed data, classifying and comparing 41 different schemes and discussing some of the most promising approaches in this area. In other words, the academic community has already undertaken preliminary exploration of “how to technically achieve auditability,” but has yet to fully answer a more fundamental governance question: “why does auditing become so difficult under privacy computing?”

This paper aims to fill this research gap, with its core inquiry organized around three questions: how does the “trust black box” of audit accountability arise under the privacy-computing framework? What are its deep roots? And how can governance reconstruction achieve a leap from “technical trust” to “institutional trust”?

To address these questions, this paper draws on five complementary methods: normative research, model analysis, case analysis, comparative analysis, and literature research. Normative research runs throughout the paper and is used to justify the legitimacy of governance reconstruction; model analysis is used for the formal modeling in Chapter 3; case analysis draws on two representative privacy-computing deployment cases in finance and healthcare for in-depth examination; comparative analysis is used to reveal differences in audit visibility between centralized and distributed architectures; and literature research provides support for the theoretical foundation. These methods correspond respectively to the three research stages of “diagnosis—attribution—prescription.”

This paper unfolds a progressive analysis across three levels:

First, by integrating information asymmetry theory with findings on verifiable computation, it constructs an analytical framework for the mechanism by which the trust black box arises;

Second, starting from the structural contradiction between the “invisibility” of the technical architecture and the “visibility” presumption of audit institutions, it examines the formation logic of the trust black box and its institutional manifestations across three dimensions—traceability failure, compliance failure, and ambiguous accountability;

Finally, building on this analysis, it proposes a governance reconstruction scheme centered on the embedding of verifiable computation, independent algorithmic auditing, and joint accountability determination. The concluding section summarizes the findings and offers corresponding recommendations.

2. Theoretical Foundation and Analytical Framework

A fundamental challenge in data governance is the long-standing tension between releasing the value of data and protecting privacy and security. Privacy computing—represented by federated learning and secure multi-party computation—attempts to resolve this tension through the technical pathway of making data “usable but not visible.” Yet its distributed, encrypted computing architecture, while cutting off traditional privacy risks, also systematically restructures the informational foundation on which audit accountability depends. The theoretical question raised by this paradox is: when auditors cannot “see” the data, how can traditional visibility-based audit logic remain self-consistent? What blind spots exist in existing research’s response to this? What analytical tools does this paper need to address this predicament? Around these questions, this chapter reviews the contributions and limitations of existing research along three dimensions—the technical logic of privacy computing, technical solutions for verifiability, and the institutional conditions of data circulation—before distilling the theoretical foundation and analytical framework of this paper.

2.1 Information Asymmetry and Principal–Agent Theory

Principal–agent theory concerns how, under conditions of information asymmetry, a principal can effectively supervise and incentivize an agent. In privacy-computing scenarios, the data provider (principal) “entrusts” data processing to a privacy-computing platform or aggregation server (agent), but because the raw data is invisible, the principal cannot directly verify whether the agent’s behavior conforms to the agreement, giving rise to typical adverse selection and moral hazard. Federated learning systems “have been shown to face potential threats at the data-collection, training, and inference stages” (Gao Ying et al., 2023), and research on horizontal federated learning spans multiple fields such as distributed systems and information security, exhibiting diversity from research motivation to technical approach (Wu Wentai et al., 2025). These characteristics further intensify, in a distributed environment, the difficulty principals face in monitoring agents’ behavior.

Information asymmetry theory therefore provides a foundational analytical tool for understanding the erosion of trust in privacy computing.

2.2 Verifiable Computation and Zero-Knowledge Proof Theory

Verifiable computation is a frontier branch of cryptography that studies how a computing party can prove the correctness of a computation result to a verifying party without disclosing sensitive data. Bontekoe et al. (2025) trace the research progress of integrating verifiability into privacy-preserving computation. Zero-knowledge proof, as a core technique, allows one party to prove that a statement is true to another party without revealing any additional information. In the Masquerade scheme of Mouris and Tsoutsos (2024), a tailored multiplicative commitment scheme is proposed to ensure the integrity of data aggregation, with all participants' commitments published on a ledger to provide public verifiability. Zhang et al. (2025) propose a verifiable federated learning system called Zkfhed, which designs a two-stage audit mechanism based on zero-knowledge proofs to verify the provenance of training data and the correctness of the computation process; it also leverages fully homomorphic encryption for homomorphically encrypted delegated learning to achieve secure outsourcing of computation tasks, effectively identifying malicious clients while remaining efficient and scalable in terms of online time and communication overhead. This body of theory forms the technical foundation of this paper's governance-reconstruction proposal—namely, that the core of auditing should shift from “data visibility” to “process verifiability.”

2.3 The Institutional Logic of Data Circulation and Governance Paradigms

From an institutional-design perspective, data circulation is essentially a form of data-use licensing. The value of data lies in cognizing the world; big data and artificial intelligence have transformed the means of acquiring and analyzing data. Data circulation is essentially a matter of data-use licensing, encompassing three modes—one-to-one licensing, one-to-many licensing, and mutual licensing—that together constitute the pattern of socialized data utilization (Gao Fuping, 2019). However, the “invisibility” characteristic of privacy computing dissolves this

institutional premise at the technical level. Privacy computing harbors multiple risks, including data-compliance risk, data-leakage risk, discrimination-amplification risk, data-archipelago risk, and trust-disintegration risk, all of which urgently require legal regulation (Yin Huarong & Wang Huimin, 2022). Administrative-law regulation can achieve full-chain governance covering the ex-ante, interim, and ex-post stages, offering advantages of timeliness, flexibility, and systematicity (Yin Huarong & Wang Huimin, 2022). This structural contradiction constitutes the deep institutional root of what this paper terms the “trust black box.”

To answer this question, this paper introduces information asymmetry and principal–agent theory as the structural analytical tool for understanding the erosion of trust in privacy computing, introduces verifiable computation theory as the technical foundation for an alternative audit paradigm, and takes the institutional logic of data circulation as the institutional design framework for governance reconstruction. These three correspond respectively to the progressively deeper analytical stages of “how the trust black box arises” (the principal–agent dilemma under information asymmetry), “how it can be resolved” (the shift from data visibility to process verifiability), and “how it can be institutionalized” (third-party auditing and joint accountability determination), forming the theoretical skeleton of “diagnosis—attribution—prescription” that runs through the paper.

3. The Formation Logic and Core Dilemmas of the Trust Black Box

3.1 Technical Root: The Paradigm Shift from “Visible” to “Invisible”

Traditional centralized data governance follows the logic of “aggregate first, audit later”—once data is aggregated onto a centralized platform, auditors can directly retrieve the raw data to complete their review, and the effectiveness of this process depends on the “visibility” of data at both the physical and logical levels. Privacy computing—especially as represented by federated learning—completely overturns this premise.

This paper adopts the objective-function definition and notational system of the

Federated Averaging algorithm (FedAvg) proposed by McMahan et al. (2017). In standard machine learning, the global empirical risk is defined as the average loss over all samples:

$$\min_{w \in \mathbb{R}^d} f(w) \text{ where } f(w) := (1/n) \sum_{i=1}^n f_i(w). \quad (1)$$

Since the data is partitioned across K clients, let the k -th client hold a local index set of data P_k , with data volume $n_k = |P_k|$, and let the total global data volume be $n = \sum_{k=1}^K n_k$. Equation (1) can then be equivalently rewritten as a weighted average of each client's local loss function:

$$f(w) = \sum_{k=1}^K (n_k/n) F_k(w) \text{ where } F_k(w) = (1/n_k) \sum_{i \in P_k} f_i(w). \quad (2)$$

Equation (2) reveals the core architecture of federated learning: the optimization objective of the global model is decomposed into a weighted combination of multiple clients' local objectives. In each round of communication, each client independently computes updates on its local data P_k , and the server can only receive the local update results uploaded by each client (such as gradients or model parameters), while never being able to directly access any raw data sample within P_k . Unlike centralized storage and processing by a data collector, federated learning distributes all participants' data locally and conducts joint training; no participant may access another participant's local data or other sensitive information without authorization, which effectively safeguards user privacy and data security (Xiao Xiong et al., 2023). Yet this causes the “visibility” premise of auditing to be systematically dissolved—standing at the server side, the auditor can only access the aggregated result on the right-hand side of Equation (2), $f(w) = \sum (n_k/n) F_k(w)$, and cannot independently verify whether the computation of any given $F_k(w)$ is correct or strictly generated based on P_k .

To quantify how this “invisibility” erodes audit effectiveness, define the audit visibility index V as the proportion of the total information that the auditor can directly obtain as effective information. In a centralized scenario, data is aggregated on a single platform and the auditor can directly access the complete dataset, so $V \approx 1$;

whereas under the federated learning framework, letting the raw data space be X and the information space accessible to the auditor be A (such as the gradient or model-parameter space), since the mapping from raw data to uploaded information is a many-to-one, non-injective function, the auditor cannot uniquely infer the raw data. Hence:

$$V = H(A)/H(X) \ll 1. \quad (3)$$

Here $H(\cdot)$ denotes information entropy. Equation (3) shows that, based solely on the uploaded aggregate information, the auditor cannot reconstruct or verify the true state of the raw data—this protects privacy, but also means a substantive impoverishment of audit information. In federated learning, client-node data is usable but not visible, and it is very difficult for the server node to audit the data quality and historical behavior of client nodes (Wang Chenyi, 2025). The technical root of the trust black box lies precisely here—the paradigm shift from “visible data” to “invisible aggregated results” means that the informational foundation on which auditing depends is actively curtailed by the algorithmic architecture itself.

3.2 Institutional Dilemma: Three Manifestations of Audit Failure

3.2.1 Dilemma One: Failure of Data Traceability

The primary step in data auditing is to trace the complete lifecycle of data—from collection and processing to every stage of computation, a verifiable “data lineage” must be established. However, in an encrypted computing environment, raw data exists in encrypted form and circulates among multiple participants, rendering traditional plaintext-identifier-based traceability techniques entirely ineffective. Gai Keke et al. (2025) note that privacy protection and transaction-data auditing are two relatively conflicting requirements among blockchain-system stakeholders; verification requires transparent access while privacy protection requires information concealment, a tension especially pronounced in encrypted computation. Defining the traceability reliability ρ as the probability that an auditor can correctly trace the raw-data source on which a given computational output depends, then in a scenario

where K clients jointly train a model, this probability is no better than random guessing:

$$\rho = \Pr(\text{successful tracing} \mid \text{output jointly generated by } \{P_k\}_{k=1}^K) \leq 1/K + \varepsilon. \quad (4)$$

Here ε is an arbitrarily small positive number (Lan & Zhang, 2025). Equation (4) shows that as the number of clients K increases, the probability of successful tracing approaches zero—this is the intrinsic mathematical expression of the traceability failure caused by “invisibility.”

This failure is not merely theoretical speculation; it has already manifested as an observable practical dilemma in real-world deployments. In 2026, JPMorgan’s Kinexys blockchain business unit, together with BNY, RBC, DeepTempo, and NVIDIA, completed a proof-of-concept for federated-learning-based fraud detection, using NVIDIA FLARE as the federated-learning development kit and running tests in a multi-site federated environment provided by NVIDIA DGX Cloud. Its core design logic is that each participating institution independently trains an AI model in its local environment, uploading only encrypted model updates and insights, while raw transaction data always remains behind each institution’s own firewall. The proof-of-concept results showed that the model trained via federated learning significantly outperformed models trained independently by a single institution in fraud detection. However, when the anti-fraud model produces false positives or false negatives, auditors face precisely the dilemma described by Equation (4)—letting K denote the number of financial institutions participating in the anti-fraud training, each bank holding its own local transaction dataset P_k , the auditor, standing at the level of the central aggregator, can only access the encrypted model updates uploaded by each bank and cannot trace a specific prediction result back to which bank’s particular historical transaction data it depended on. As the number of participating institutions K increases, the probability of successful tracing ρ rapidly approaches zero, and the auditor essentially loses any reliable capacity to trace model outputs. Federated learning enables institutions to “collaboratively train AI models without

sharing sensitive transaction data”—the flip side of this privacy-protection advantage is precisely that the traceability foundation of auditing is actively removed by the algorithmic architecture itself. The localized execution of federated learning lacks third-party visibility into the training process, which increases the necessity of verifying that process (Ma et al., 2024).

3.2.2 Dilemma Two: Failure of Compliance Verification

Ensuring the correctness and integrity of aggregated results is a basic prerequisite for privacy-computing systems to gain institutional trust, and ensuring the integrity of aggregated results remains a key research focus for protecting users' legitimate interests (Li et al., 2025). Privacy-computing systems typically introduce mathematical tools such as differential privacy (DP) to quantify the degree of privacy protection. In practical federated learning, the objective function after incorporating differential privacy is typically written as:

$$\min_w f(w) + \lambda \cdot \Pi(w). (5)$$

Here $f(w)$ is the global empirical loss defined in Equations (1) and (2), $\Pi(w)$ is the privacy-leakage risk function (such as an upper bound on the L2 sensitivity of the gradient, $\max_{P, P'} \|\nabla F(w; P) - \nabla F(w; P')\|_2$, or an L2 regularization term on model parameters), and $\lambda \in (0, \infty)$ is the trade-off coefficient. Equation (5) shows that the model's optimization objective shifts from “pursuing accuracy alone” to “a compromise between accuracy and privacy.” However, the setting of λ is essentially a governance decision—too large and model utility is severely impaired, too small and privacy protection becomes nominal. Currently, the vast majority of privacy-computing systems leave this parameter to be decided internally by the development team, lacking institutionalized audit oversight. The key to the failure of compliance verification is that, even when the mathematical proof of differential privacy is complete, errors in algorithmic implementation, programming bias, and even the non-transparency of parameter selection can cause the actual level of protection to fall far short of the theoretical value. If auditors cannot access the basis

on which λ was set and its real impact on the final model, compliance assessment loses any operable foothold.

The Health-FedNet privacy-preserving federated-learning framework proposed by Ali et al. (2025) provides a precise illustration of this failure. The framework is designed to allow multiple healthcare institutions to collaboratively train chronic-disease prediction models without transmitting raw clinical information; Health-FedNet integrates differential privacy (DP), homomorphic encryption (HE), and an adaptive node-weighting mechanism to enhance privacy, scalability, and robustness against heterogeneous data distribution, and experiments on the MIMIC-III clinical database show that Health-FedNet improves diagnostic accuracy by 12% compared with a centralized model. However, an independent third-party auditor would face precisely the fundamental obstacle revealed by Equation (5)—the value of λ , the implementation details of the differential-privacy noise mechanism, and the specific parameters of the secure aggregation protocol are all decided internally by the development team, and the auditor cannot independently verify the basis for these settings or their true impact on the final model. Even where the mathematical proof of differential privacy is theoretically complete, programming bias in implementation, non-transparency in parameter selection, and even differences in data formats across hospitals could all cause the actual level of privacy protection to fall far short of the theoretically claimed level. In the Health-FedNet case, auditors are essentially excluded from the key decision-making stages of compliance verification—this is a typical manifestation of the failure of compliance verification.

3.2.3 Dilemma Three: Ambiguous Accountability

Under a distributed computing framework, when the final model exhibits bias, discrimination, or data leakage, should responsibility fall on the data provider, the algorithm designer, or the platform operator? The traditional logic of “whoever produces is responsible” fails entirely in multi-party privacy-computing scenarios. Horizontal federated learning spans multiple academic fields (Wu Wentai et al., 2025), and the complexity of the responsibility chain grows exponentially. In an encrypted

environment, auditors cannot obtain the raw data or complete algorithmic logs of each participant, so the attribution of responsibility can only rest on ambiguous inference.

The two basic federated-learning deployment cases discussed above jointly reveal the real-world root of ambiguous accountability under multi-party collaboration. In the Kinexys financial project, which was not equipped with audit-enhancing modules such as verifiable traceability or on-chain evidence recording, the failure of traceability meant that auditors could not determine which participant's local training bias caused a given anti-fraud false positive, and fine-grained accountability determination therefore lacked a reliable data foundation—was it the data provider introducing systematic bias during local labeling, a flaw in the algorithm designer's aggregation strategy, or improper parameter configuration by the platform operator? Auditors could not make a precise, actionable traceability judgment on this. In the Health-FedNet healthcare scenario, built on a comparable underlying architecture, insufficient compliance-audit capability meant that auditors could not fully verify whether the final model truly honored its privacy-protection commitments (Ali et al., 2025); once a membership-inference attack or data-leakage risk occurs, responsibility—whether it lies with the development team (improper choice of privacy parameters), the medical institution (non-compliant local data handling), or the framework designer (vulnerabilities in the underlying security protocol)—likewise falls into an ambiguous state with no evidence to support fine-grained accountability. What is particularly concerning is that this ambiguity of rights and responsibilities arising from the absence of fine-grained traceability may be strategically exploited by participants—"the decentralization of responsibility evolves, in substance, into responsibility that is nowhere to be found," as all parties can shift blame behind the algorithmic black box of encrypted computation, further reducing their willingness to actively cooperate with in-depth auditing.

3.3 The Formation Mechanism of the Trust Black Box: A Self-Reinforcing Triple-Dilemma Cycle

The three dilemmas above do not exist independently but instead form a self-reinforcing vicious cycle: the failure of traceability strips compliance verification of its data foundation; the failure of compliance verification leaves responsibility determination without any basis; and ambiguous accountability reduces participants' willingness to cooperate with audits, further aggravating the difficulty of traceability. These three links form a positive-feedback loop, which can be approximately described using a system of difference equations:

$$\begin{aligned}\rho_{t+1} &= \rho_t - \mu_1(1 - C_t), \\ C_{t+1} &= C_t - \mu_2(1 - \rho_t) - \mu_3\psi_t, \\ \psi_{t+1} &= \psi_t + \mu_4(1 - C_t). \quad (6)\end{aligned}$$

Here ρ_t is the traceability reliability in audit cycle t , C_t is the effectiveness of compliance verification, ψ_t is the uncertainty of responsibility attribution, and $\mu_1, \mu_2, \mu_3, \mu_4 > 0$ are feedback-intensity coefficients. Equation (6) shows that, absent external intervention, the system will automatically tend toward the extreme states of $\rho \rightarrow 0$, $C \rightarrow 0$, and $\psi \rightarrow \infty$ —that is, the total collapse of the audit system. As this cycle approaches its critical point, privacy computing degenerates from a “trust solution” into a “trust black box”—technically, data circulation remains achievable, yet data subjects find it difficult to understand how privacy computing processes and operates their data, casting doubt on the accuracy and reliability of the output results (Lu Anwen, 2025), while at the institutional level no effective trust in the circulation process can be established at all.

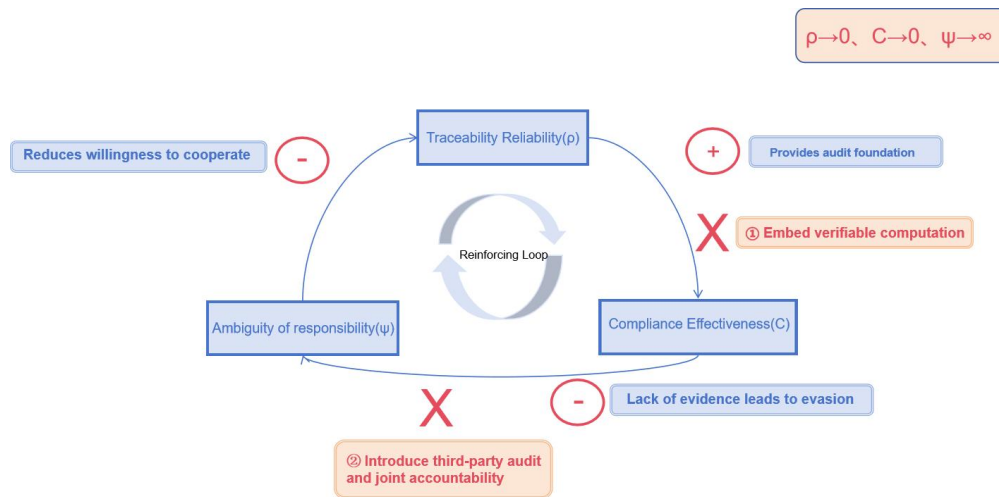


Figure 1: Positive feedback causal loop of the privacy computing audit dilemma

In summary, the trust black box is not merely a technical flaw, but a systemic governance crisis arising from a mismatch between the technical architecture and institutional arrangements. Its formation logic can be summarized as follows: privacy computing protects privacy through “invisibility,” but in doing so unexpectedly severs the “visibility” information chain on which auditing depends, thereby triggering a cascading reaction across the three failures of traceability, compliance, and accountability, and ultimately generating, at the institutional level, an impenetrable black box. This understanding lays the problem foundation for the governance-reconstruction proposal that follows.

4. Governance Reconstruction: From Technical Trust to Institutional Trust

The analysis above shows that the trust black box arises from the structural rupture between the “invisible” architecture of privacy computing and the “visibility” premise of auditing, and that its triple dilemma continuously reinforces itself through a positive-feedback loop, ultimately driving the audit system toward collapse. This raises the core question: can this governance crisis be effectively resolved? This chapter argues that the key to resolving the trust black box does not lie in abandoning the technical route of privacy computing, but in driving a fundamental transformation of the audit paradigm from “based on data visibility” to “based on process

verifiability.” This transformation requires simultaneous progress along three dimensions: technical embedding, institutional design, and regulatory coordination.

4.1 Audit Paradigm Transformation: From “Data Visibility” to “Process Verifiability”

The essence of the trust black box is that “what cannot be seen cannot be audited.” The logical starting point for resolving it is therefore that auditing need not require “making data visible,” but should instead move toward “process verifiability”—achieving independent verification of the correctness, integrity, and compliance of the computation process without sacrificing privacy.

This paradigm shift already has feasible technical support. Bontekoe et al. (2025) systematically review research progress on combining verifiability with privacy-preserving computation in order to preserve confidentiality while maintaining correctness; applications that are data-intensive and require strong privacy protection especially need verifiable correctness guarantees in outsourcing scenarios. The Masquerade scheme of Mouris and Tsoutsos designs a multiplicative commitment mechanism that ensures the integrity of multi-party data aggregation and publishes all participants’ commitments on a ledger to provide public verifiability. The Zkfhed system proposed by Zhang et al. (2025) goes further, designing a two-stage audit mechanism based on zero-knowledge proofs that can simultaneously verify the provenance of training data and the correctness of the computation process. In addition, the development of decentralized trusted federated-learning aggregation schemes that incorporate zero-knowledge proofs, together with verifiable confidential cloud-computing protocols, further demonstrates that “privacy” and “auditability” are not irreconcilable contradictions but can, at the technical level, coexist synergistically.

However, technical solutions only provide the possibility of transformation; without corresponding institutional arrangements, technical capability will struggle to translate into actual audit effectiveness.

4.2 Institutional Reconstruction: Third-Party Auditing and Accountability Mechanisms

At the institutional level, the transformation of the audit paradigm requires the construction of three core mechanisms.

4.2.1 Independent Third-Party Algorithmic Auditing

The algorithmic design, parameter settings, and computation processes of privacy-computing platforms should be subject to periodic auditing by third-party institutions independent of the developers and operators. The independence of third-party auditors can effectively alleviate the structural difficulty, inherent in the principal–agent framework, of monitoring agent behavior—when the principal cannot directly verify whether the agent’s behavior conforms to the agreement, an independent third-party review mechanism is the only feasible institutional alternative. Existing research has proposed a lightweight, blockchain-based privacy-preserving audit framework that builds a secure and efficient computing model, supporting data integrity, traceability, and confidentiality in practical application scenarios such as medical-record systems and financial audit platforms.

4.2.2 Joint Accountability Determination Mechanism

The traditional one-way logic of “whoever produces is responsible” fails entirely in multi-party privacy-computing scenarios. Responsibility should be extended from the single entity of the “data provider” to a shared, multi-party arrangement, establishing a joint-responsibility framework based on “algorithmic contribution plus degree of fault.” Zhu Jianming et al. (2021) note that federated learning is a distributed machine-learning framework that stores data locally at participating nodes and can effectively protect the data privacy of intelligent edge nodes; existing federated-learning approaches typically upload intermediate model-training parameters to a parameter server for model aggregation, a process with two problems: first, the privacy leakage of intermediate parameters—existing privacy-protection schemes typically add noise to intermediate parameters via differential privacy, but excessive noise degrades the quality of the aggregated model; second, the self-interested and fully autonomous nature of node training may lead malicious

nodes to upload false parameters or low-quality models, affecting the aggregation process and model quality (Zhu Jianming et al., 2021). To address these dilemmas, that study reconstructs the centralized parameter server of federated learning into a decentralized parameter-aggregation chain, using blockchain to record the intermediate parameters of the model-training process as evidence, and incentivizing collaborating nodes to verify model parameters while penalizing participants who upload false parameters or low-quality models, thereby constraining their self-interested behavior. At the same time, it uses model quality as an evaluation basis to achieve dynamic adjustment of the privacy noise applied to intermediate parameters and adaptive model aggregation. Its prototype experiments confirm that the model not only enhances mutual trust among federated-learning participants but also prevents the leakage of intermediate-parameter privacy, thereby achieving a trusted federated-learning model with enhanced privacy protection. This blockchain-based auditable framework, through a consensus algorithm based on proof of node contribution and an asynchronous parameter-audit mechanism, provides a solid technical reference for the quantified apportionment of joint responsibility. Specifically, when the final model exhibits bias, discrimination, or data leakage, responsibility should be apportioned proportionally according to each party's actual contribution to and degree of fault in model generation—contribution can be measured by the information content of gradient updates or the sensitivity of model parameters, while the degree of fault must be determined in conjunction with specific violations identified through third-party auditing.

4.2.3 Systematic Improvement of Laws and Regulations

The legal limitations of privacy protection in federated learning (Liu Zegang, 2025) reflect the fact that regulatory development lags behind technological progress. Coordinated progress on five measures—improving laws and regulations, building unified standards, consolidating the trust system, strengthening regulatory enforcement, and fostering compliance awareness—should be pursued to provide a clear compliance framework and institutional safeguard for the development of

privacy computing.

4.3 Regulatory Innovation: Standardization and Cross-Domain Collaboration

The effective implementation of institutional reconstruction cannot be achieved without standardization at the regulatory level and international coordination.

At the domestic level, dedicated standards for privacy-computing audits should be developed. Existing research has proposed an integrated protocol combining verifiable confidential cloud computing with data-integrity auditing across both cloud-storage and cloud-computing scenarios, achieving coordinated functionality across three tasks under ciphertext: verification of cloud-computing results, integrity auditing of cloud-stored data, and privacy protection of cloud-stored data. Building on such technical explorations, dedicated audit standards should be further developed to clarify the qualifications of audit subjects, the scope of audit objects, the required frequency of audits, and the norms for audit methodology, so that third-party auditing has a clear and standardized basis to follow.

At the international level, data circulation supported by privacy computing is inherently cross-border in nature, requiring active international cooperation to establish common rules for cross-border privacy-computing audits, avoiding new “compliance silos” caused by fragmented regulatory standards across countries.

To comprehensively assess the implementation effects of the governance-reconstruction scheme above, a data-circulation governance index G can be constructed to quantitatively measure the governance level of a privacy-computing system. It is defined as:

$$G = \alpha \cdot A + \beta \cdot T + \gamma \cdot E. \quad (7)$$

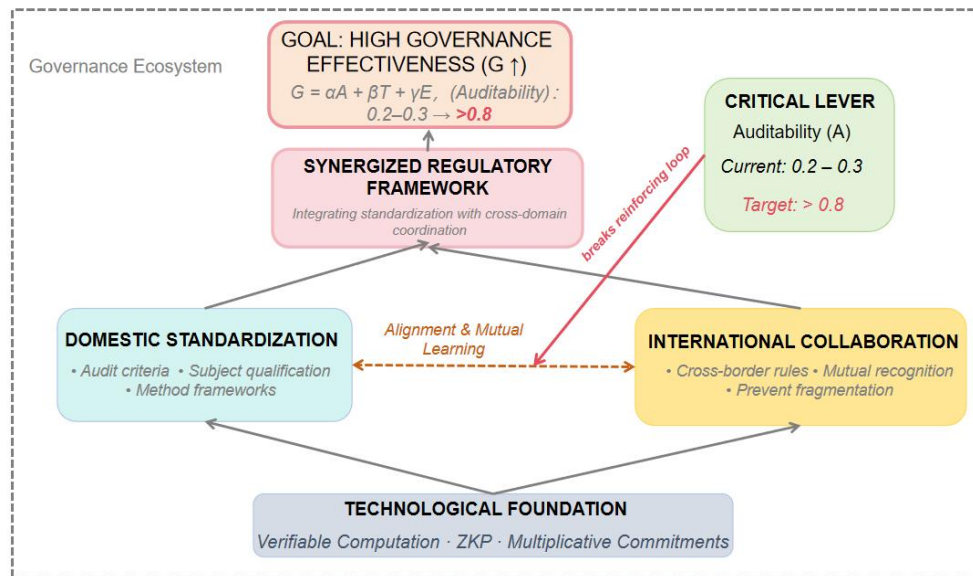


Figure 2. Synergized governance architecture for standardization and cross-domain collaboration.

Here A is auditability, measuring the system's capacity for independent verification of the computation process; T is trustworthiness, reflecting participants' confidence that the system operates as expected; E is efficiency, a composite indicator encompassing computational efficiency, communication overhead, and audit cost; and $\alpha, \beta, \gamma \in (0,1)$ are the weighting coefficients for each dimension, satisfying $\alpha + \beta + \gamma = 1$. The construction logic of this index echoes the dynamic-systems analysis shown in Equation (6)—auditability A is the core lever variable for breaking the positive-feedback loop of the triple dilemma. The core objective of governance reconstruction is to raise A from its currently low level (estimated at roughly 0.2–0.3) to above 0.8, moving G into the high-governance-effectiveness range (≥ 0.7). This quantitative framework offers a practicable, unified tool for evaluating and comparing the governance levels of different privacy-computing systems.

Equation (7) sets auditability A as the core explanatory variable of the governance index G . Drawing on discussions of threshold effects in the field of technology governance, improvements in governance effectiveness typically display S-shaped growth—initially constrained by insufficient coupling between institutions

and technology and thus growing slowly, then entering a phase of rapid release after crossing a critical threshold, and finally tending toward saturation. Based on this theoretical expectation, this paper uses a logistic growth model to numerically simulate how G changes with A , with the functional form:

$$G(A) = G_{\max} / (1 + e^{-k(A-A_0)}) \quad (8)$$

Here $G_{\max}$ is the theoretical upper bound of the governance index, k is the growth-rate coefficient, and A_0 is the inflection point. Parameter values are calibrated according to the marginal contribution rate of the αA term in Equation (7), holding T and E at their baseline levels so as to isolate the net effect of A on G . Figure 3 presents the numerical simulation results under this model.

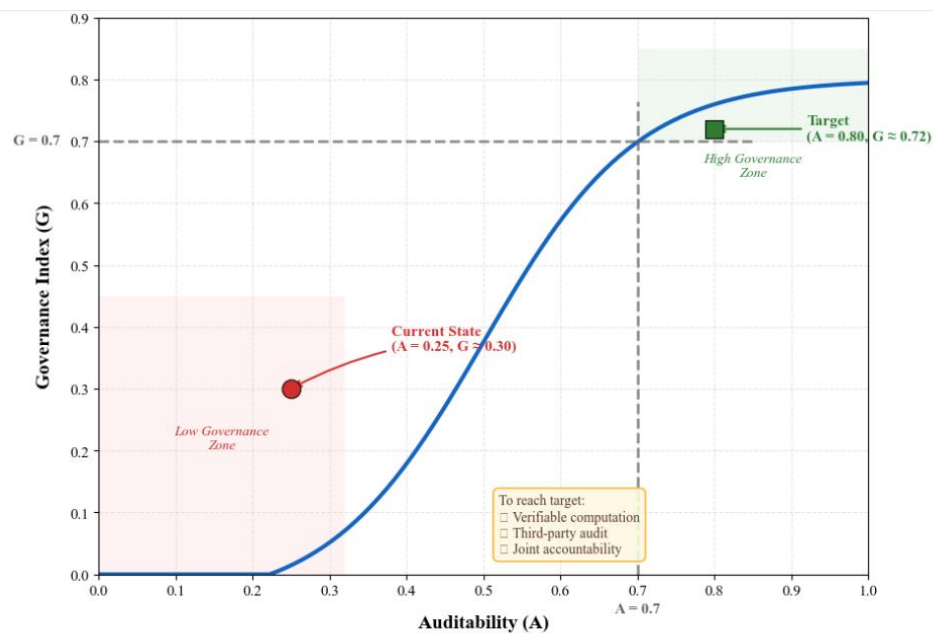


Figure X. The impact of auditability (A) on the governance index (G).

Note: The S-curve is derived from the logistic growth model based on Eq. (7). The red dot marks the current industry average ($A \approx 0.25$, $G \approx 0.30$); the green square marks the governance reform target ($A = 0.80$, $G \approx 0.72$). Dashed lines indicate the threshold ($A \geq 0.7$, $G \geq 0.7$).

Figure 3. The impact of auditability (A) on the data circulation governance index (G).

Note: The curve is derived from the logistic growth model (Equation (8)), with parameters calibrated according to the marginal contribution rate of the αA term in Equation (7); the red dot marks the current-state anchor ($A \approx 0.25$, $G \approx 0.30$),

corresponding to the “locked-in state” described in Section 3.3; the green dot marks the governance-reconstruction target anchor ($A = 0.80$, $G \approx 0.72$), corresponding to the expected effects of the three institutional innovations in Section 4.2; the grey dashed lines mark the threshold ($A \geq 0.7$, $G \geq 0.7$). This figure is a theoretical simulation intended to reveal the threshold effect of A as the core lever variable, rather than an empirical measurement.

Figure 3 intuitively illustrates the above simulation results: when auditability A is below 0.3, the governance index G is locked into a low-effectiveness range below 0.4—precisely the “low-level equilibrium” state corresponding to the self-reinforcing triple-dilemma cycle described in Section 3.3; whereas once A crosses the 0.7 threshold, G rapidly rises into the high-effectiveness range above 0.7. The gap between the current-state point ($A \approx 0.25$, $G \approx 0.30$) and the target point ($A = 0.80$, $G \approx 0.72$) shows that marginal improvement alone cannot break the governance impasse; a structural leap must be achieved through the three institutional innovations proposed in Section 4.2—the embedding of verifiable computation, independent third-party algorithmic auditing, and joint accountability determination. This quantitative analysis provides a clear institutional-effectiveness target for this paper’s governance-reconstruction proposal and lays an operable evaluative benchmark for the policy recommendations that follow.

4. Conclusion and Recommendations

Privacy computing offers a revolutionary pathway for breaking data silos and releasing the value of data elements. However, while “usable but not visible” dissolves privacy risks, it also creates a trust black box in which “what cannot be seen cannot be audited.” This paper has revealed the formation logic of this trust black box—the structural contradiction between privacy protection and audit visibility—and has shown that the “black-box problem,” whereby the central aggregator in federated learning’s distributed architecture cannot verify client model updates, is a typical technical manifestation of this contradiction. On this basis, the paper has

demonstrated, across the three dimensions of traceability failure, compliance failure, and ambiguous accountability, the self-reinforcing cycle underlying the audit dilemma.

The findings suggest that the key to resolving the trust black box does not lie in abandoning privacy computing, but in driving a fundamental transformation of the audit paradigm from “based on data visibility” to “based on process verifiability.” The maturation of technologies such as zero-knowledge proofs and verifiable computation provides feasible support for this transformation, while institutionalized third-party auditing and accountability mechanisms are the necessary safeguard for putting it into practice.

Accordingly, this paper offers the following recommendations:

1. Technical standards first: incorporate auditability as a core design principle of privacy-computing systems, mandating the configuration of verifiable-computation modules and audit interfaces;
2. Establishment of institutional bodies: set up independent algorithmic-audit seats within data exchanges, and establish third-party auditing and joint-accountability-determination mechanisms;
3. Refinement of regulatory rules: formulate dedicated standards for privacy-computing audits, and promote compliance certification of auditability and cross-domain coordination.

Ultimately, audit-based trust in data circulation is not merely a technical or legal issue, but a fundamental institutional question bearing on whether the data-element market can develop in a healthy manner. Only through the coordinated reconstruction of technology and institutions—namely, the organic unification of embedding verifiable computation, third-party algorithmic auditing, and institutional accountability—can the benefits of privacy computing be enjoyed while safeguarding the trust foundation of data governance.

References

Li, S., Ji, Y., Shi, D., Liao, W., Zhang, L., Tong, Y., & Xu, K. (2022). A data federation system for multi-party security. *Journal of Software*, 33(3), 1111–1127. <https://doi.org/10.13328/j.cnki.jos.006458>.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Vol. 54, pp. 1273–1282). PMLR. DOI:10.48550/arXiv.1602.05629.

Xiao, X., Tang, Z., Xiao, B., & Li, K. (2023). A survey of privacy protection and security defense research in federated learning. *Chinese Journal of Computers*, 46(5), 1019–1044.

Jiang, W., Cui, X., Liu, Z., Chen, S., & Hu, J. (2026). An analytic federated learning method based on learnable aggregation weights. *Chinese Journal of Computers*, 49(1), 84–108.

Wu, W., Wu, Y., Lin, W., & Zuo, W. (2025). Horizontal federated learning: research status, system applications and challenges. *Chinese Journal of Computers*, 48(1), 35–67.

Gai, K., Chen, S., & Zhu, L. (2025). Blockchain-based auditable privacy-preserving confidential transactions. *Acta Electronica Sinica*, 53(2), 460–473.

Gao, Y., Chen, X., Zhang, Y., Wang, W., Deng, H., Duan, P., & Chen, P. (2023). A survey of attack and defense techniques for federated learning systems. *Chinese Journal of Computers*, 46(9), 1781–1805.

Bontekoe, T., Karastoyanova, D., & Turkmen, F. (2025). Verifiability for privacy-preserving computing on distributed data — a survey. *International Journal of Information Security*, 24, Article 141. <https://doi.org/10.1007/s10207-025-01047-7>

Mouris, D., & Tsoutsos, N.G. (2024). Masquerade: Verifiable multi-party aggregation with secure multiplicative commitments. *ACM Transactions on Internet Technology*, 25, 1–31. <https://doi.org/10.1145/3705315>

Zhang, B., Lu, G., Wu, Y., Ren, K., & Zhu, F. (2025). Zkfhed: A verifiable and scalable blockchain-enhanced federated learning system. *IEEE Transactions on Knowledge and Data Engineering*, 37(6), 3841–3854. <https://doi.org/10.1109/TKDE.2025.3550546>

Gao, F. (2019). Data circulation theory: the foundation of data resource rights allocation. *Chinese and Foreign Legal Studies*, 31(6), 1405–1424.

Yin, H., & Wang, H. (2022). Administrative-law regulation of privacy computing. *Journal of Hunan University of Science and Technology (Social Science Edition)*, 25(6), 93–101. <https://doi.org/10.13582/j.cnki.1672-7835.2022.06.013>.

Wang, C. (2025). Research on secure aggregation and aggregation-node selection algorithms for blockchain-based federated learning (Master's thesis, Guangxi Minzu University). <https://doi.org/10.27035/d.cnki.ggxmc.2025.000607>.

Lan, G., & Zhang, C. (2025). A lightweight privacy-preserving audit framework based on blockchain and hybrid encryption. *Security and Privacy*, 8(6), e70102. <https://doi.org/10.1002/spy2.70102>

Ma, J., Liu, H., Zhang, M., & Liu, Z. (2024). VPFL: Enabling verifiability and privacy in federated learning with zero-knowledge proofs. *Knowledge-Based Systems*, 299, 112115. <https://doi.org/10.1016/j.knosys.2024.112115>

Li, S., Wei, X., & Wang, H. (2025). OVP-FL: Outsourced verifiable privacy-preserving federated learning. *IEEE Transactions on Network Science and Engineering*, 12(3), 2057–2068. <https://doi.org/10.1109/TNSE.2025.3543601>

Ali, A., Snášel, V., & Platoš, J. (2025). Health-FedNet: A privacy-preserving federated learning framework for scalable and secure healthcare analytics. *Results in Engineering*, 27, 106484. <https://doi.org/10.1016/j.rineng.2025.106484>

Lu, A. (2025). Generative artificial intelligence: risks, regulation, and governance models. *Journal of Chongqing University of Posts and Telecommunications (Social Science Edition)*, 37(3), 113–121.

Zhu, J., Zhang, Q., Gao, S., Ding, Q., & Yuan, L. (2021). A blockchain-based privacy-preserving trusted federated learning model. *Chinese Journal of Computers*, 44(12), 2464–2484.

Liu, Z. (2025). The limitations of and ways to overcome privacy protection in federated learning in the age of artificial intelligence. *Chinese and Foreign Legal Studies*, 37(1), 65–84.