Coding Justice: Architectural Shift from Ex-Post Punishment to Ex-Ante Design in Privacy-Preserving AI

Linyu Zhang
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
99 Yanxiang Road, Yanta District
Xi'an City, Shaanxi Province, 710054, China
zhanglinyu@stu.xjtu.edu.cn

Abstract

The development of Artificial Intelligence relies on large-scale data, which creates significant privacy and compliance challenges. This article examines how privacy-preserving technologies like federated learning and differential privacy reshape legal relationships and responsibility allocation through their technical architectures. We argue that these technologies do more than protect privacy; they enable a legal paradigm shift from ex post punishment to ex ante design. This shift clearly redefines the responsibilities of data controllers and processors, laying the groundwork for a new, technology-enabled legal governance framework.

Introduction

The rapid development of artificial intelligence, especially large-scale models, has expanded its use in areas like healthcare, finance, and public services. However, a clear tension has emerged between data-driven AI systems and strict data privacy laws such as the EU's GDPR and China's Personal Information Protection Law (PIPL). This conflict is particularly evident in fields like healthcare and education, which handle highly sensitive information. Here, organizations face the challenge of using AI to improve services while also ensuring data privacy and security. Traditional legal compliance methods rely mainly on contracts and after-the-fact accountability. These approaches often fall short when dealing with the complex and non-transparent data processing practices of modern AI systems.

In response, a range of privacy-preserving AI technologies—such as split learning and differential privacy—has emerged. These methods aim to achieve "using data without exposing it" and represent a key research direction today. This paper argues that the value of these technologies goes beyond providing privacy protection tools. More importantly, they reshape legal relationships and responsibility allocation in data processing through architectural redesign. In doing so, they offer a new compliance paradigm for the AI era.

Development and Characteristics of Privacy-Preserving Technologies

The initial concept of privacy-preserving technologies stemmed from a straightforward need: to transform or distort sensitive data before sharing and analysis, thereby reducing its sensitivity and preventing direct exposure of personal identity or private information.

One of the earliest and most typical Privacy-Preserving Technologies is k-anonymity (Sweeney 2002). Its core idea is to process quasi-identifiers in a dataset (a set of attributes like zip code, age, and gender that can uniquely or nearly uniquely identify an individual) through generalization and suppression. However, k-anonymity has notable limitations. It cannot defend against homogeneity attacks or background knowledge attacks . To address these issues, enhanced models like l-diversity (Machanavajjhala et al. 2007) and t-closeness (Li, Li, and Venkatasubramanian 2006) emerged. Still, they often come at the cost of higher information loss and can be difficult to implement effectively with complex, high-dimensional data.

As data environments grew more complex, traditional techniques faced severe challenges in balancing privacy protection and data utility. Entering the 21st century, a privacy model built on strict mathematical definitions—Differential Privacy (DP) (Dwork 2006)—came to the forefront and gradually became the gold standard in the field . The differential privacy framework provides a quantifiable and provable security guarantee. Its core principle is that, regardless of an attacker's background information, adding carefully calibrated random noise to query results ensures that the presence or absence of any single data record does not significantly affect the algorithm's output. The proposal of differential privacy was a milestone because it transformed privacy protection from a vague concept into a mathematical quantity that can be precisely controlled and theoretically proven . In recent years, DP research has moved from theory to practice, with its application scenarios expanding from traditional statistical queries to complex tasks like machine learning and natural language processing (Abadi et al. 2016).

Although differential privacy provides a powerful privacy definition, it often relies on a trusted data collection center to add noise. To completely avoid raw data leaving local de-

vices, a distributed machine learning paradigm called Federated Learning (FL) (Li et al. 2020) emerged . The core idea of federated learning is that multiple participants (e.g., mobile devices, hospitals) hold data locally and compute model updates (like gradients). Only these updates are uploaded to a central server for aggregation to generate a global model, while the raw data always remains local . Federated learning is essentially an architectural privacy-preserving solution. By design, it establishes a "technical boundary" between the data source and the central server, physically restricting data flow.

As a variant of federated learning, Split Learning (Thapa et al. 2022) offers another architectural approach. It splits a deep learning model into two parts at an intermediate layer (the cut layer). The client holds and executes the first part of the model, then sends the generated intermediate representations (not the raw data) to the server. The server completes the forward and backward propagation of the remaining part. This method further reduces the amount of information the client needs to transmit and hides the raw data and its corresponding model structure.

Federated learning and split learning protect privacy by altering the computational architecture, representing a significant shift in privacy-preserving technologies from simply "processing data" to "restructuring the process."

For scenarios with extremely high-security requirements, relying solely on noise addition or gradient aggregation might still be insufficient. Here, powerful tools from cryptography—Homomorphic Encryption (HE) (Acar et al. 2018) and Secure Multi-Party Computation (MPC) (Goldreich 1998)—provide an alternative approach: performing computations directly on ciphertext.

Homomorphic encryption allows specific algebraic operations (like addition and multiplication) to be performed directly on encrypted data. The decrypted result matches the result obtained if the same operations were performed on the plaintext. This means data owners can send encrypted data to an untrusted cloud server. The server performs tasks like machine learning model inference on the ciphertext and returns the encrypted result. The data owner then decrypts it to get the final output. Throughout this process, the cloud server never accesses any plaintext data, achieving a very strong level of privacy protection.

A common feature of these technologies is the establishment of a "technical boundary" within the data processing workflow. The operations of data owners and algorithm providers are confined to either side of this boundary, with access permissions and information exposure scope technically defined.

Legal Foundations of Privacy-Preserving Technologies

The emergence of privacy-preserving technologies represents not only a breakthrough in computing and algorithms but also a direct response to increasingly strict data protection laws. The design principles of these technologies align closely with core data protection principles embodied in regulations like the EU's GDPR and China's Personal Information Protection Law (PIPL). They can be seen as examples of legal principles being implemented through engineering. This section analyzes the legal foundations behind these technologies.

Technical Embedding of the Data Minimization Principle

The data minimization principle requires that personal information processing be "limited to the minimum scope necessary to achieve the processing purposes." It prohibits the collection and use of personal information unrelated to these purposes (Calzada 2022).

Federated Learning strictly confines data processing activities to their source under a "move models, not data" paradigm. Participants only send locally trained model updates, such as gradients, to a central server for aggregation. They do not send the original raw data. This design dramatically reduces the requirement for raw data to leave the device. This implements data minimization at the level of information content. Algorithm developers only access refined parameters directly related to model optimization, not the broad set of original personal data.

Split Learning takes this a step further. This technique splits a deep learning model at an intermediate layer. The data holder transmits only the intermediate features (activations) to the collaborator. These features are high-dimensional and abstract. They are not directly interpretable by the collaborator and are extremely difficult to reverse-engineer into the original data. This achieves an even more thorough form of data minimization than Federated Learning. It ensures, at the source, that any shared information is strictly limited and necessary.

Engineering Implementation of the Privacy by Design Principle

The "Privacy by Design" principle requires integrating privacy protections throughout the entire lifecycle of a product or service. It must be proactively embedded starting from the design stage (Calzada 2022). Privacy-preserving technologies do not simply add patches to existing systems. Instead, they make privacy protection the starting point and core of the system architecture.

Federated Learning is a typical engineering practice of this principle. From its inception, this architecture aimed to proactively prevent the privacy risks associated with centralized data collection. It focuses on prevention rather than remediation after a data breach occurs.

Similarly, Differential Privacy provides provable privacy guarantees by introducing carefully calibrated mathematical noise. This approach quantifies privacy protection into a precise parameter and systematically embeds it into the computational process. It represents the ultimate embodiment of the "Privacy by Design" principle at the algorithmic level.

Fulfillment of Security Obligations and Reshaping of Responsibility Boundaries

Data controllers have a legal obligation to adopt necessary technical measures to ensure data security (Calzada 2022).

Adopting privacy-preserving technologies serves as strong evidence that data controllers are actively fulfilling this security obligation. Simultaneously, these technologies reshape the responsibility boundaries in data processing activities.

For example, deploying a federated learning system means the data controller has implemented technical measures like "local data storage" and "encrypted transmission of model updates." This itself demonstrates the fulfillment of the security obligation. More profoundly, technologies like Homomorphic Encryption enhance security to a cryptographic level by allowing computations to be performed directly on ciphertext. This achieves "data usability without visibility."

The legal significance of this architecture is that it clearly separates the "custodial responsibility" for data from the "right to use data for computation." The data controller (e.g., a hospital) may retain the security responsibility for safeguarding the data. Meanwhile, the data processor (e.g., a research institute) performs computations only on the encrypted data. This technically minimizes the risk of original data leakage resulting from improper operations by the processor. Consequently, it alters the responsibility structure and risk allocation between the parties.

Architectural Assurance of the Purpose Limitation Principle

The purpose limitation principle requires that personal information collected shall not be used for purposes incompatible with the original purposes (Calzada 2022). Privacy-preserving technologies provide architectural assurance for complying with this principle by limiting the exposure of the data itself.

When algorithm developers can only access gradient updates aggregated through federated learning or intermediate features from split learning, it becomes technically difficult for them to use this information for any purpose other than model training. These features are highly specialized and lack value as general-purpose data. In contrast, if they possessed the original dataset, the potential and risk of using it for other analyses would exist. Therefore, privacy-preserving technologies establish a technical barrier at the operational level. This transforms the legal requirement that data can "only be used for specific purposes" into a built-in constraint of the system's operation.

Risk Prevention Concept and Continuous Evaluation Requirements

Modern data protection laws emphasize a risk-based regulatory approach. However, technological limitations reveal a challenge the law must confront: absolute privacy security is difficult to achieve. Risk can only be managed, not eliminated.

For instance, re-identification attacks against k-anonymity models have been confirmed by multiple studies. This has made regulators recognize that the risk of de-identified data is not zero. Similarly, although differential privacy provides strong theoretical guarantees, setting its privacy budget is itself a social decision involving risk trade-offs. It is not merely a technical issue. Therefore, the law cannot exempt controllers from all responsibility simply because they adopt a specific de-identification technology. Instead, it needs to establish a continuous risk assessment and regulatory framework. This framework should urge controllers to continuously adjust and upgrade their privacy protection measures based on technological developments and evolving attack methods.

From Ex-Post Punishment to Ex-Ante Design: A Paradigm Shift in Legal Approach and New Responsibility Allocation

The most profound legal significance of privacy-preserving technologies lies in their facilitation of a paradigm shift in legal thinking—from ex-post punishment to ex-ante design. Changes in technical architecture directly lead to clearer and more precise legal responsibilities.

Paradigm Shift: From External Contractual Constraints to Internal Architectural Constraints

In traditional data collaboration models, the law primarily relies on contract terms to define the rights and obligations of data controllers and processors. Compliance is usually achieved through ex-post audits and liability for breaches. The limitations of this model are clear. Breaches can only be addressed after harm occurs. Furthermore, facing complex and non-transparent data processing flows makes evidence collection difficult and blurs the chain of liability.

Privacy-preserving technologies fundamentally change this landscape. Taking Split Learning and Federated Learning as examples, their architectural design physically and logically prevents the direct transfer of raw data. Algorithm developers (data processors) can technically only access intermediate features or model parameters that are not directly interpretable, not the raw personal data. This "architectural isolation" makes unauthorized access to raw data impossible at the system level. Its protective effect far exceeds any contractual clause relying on ex-post liability. As scholars point out, the function of anonymization (and related de-identification techniques) should shift from merely protecting individual privacy to systematically reducing re-identification risk during data utilization.

This means the method of achieving legal compliance has fundamentally changed. It has transformed from an external, backward-looking review standard into an internal, pre-set operational characteristic built into the system. This reflects the deepening of the "Privacy by Design" legal principle, turning abstract compliance requirements into concrete, executable, and sustainable technical norms.

Responsibility Allocation: Using Technical Boundaries to Redefine Legal Liability

The "technical boundary" established by the architecture of privacy-preserving technologies provides an objective and auditable basis for clearly defining the legal responsibilities of all parties in judicial and regulatory practice. This addresses the issue of "responsibility not flowing with data circulation."

Under the new responsibility framework, the core responsibility of the data controller is focused on source guarantee. First, they must ensure the reliability and security of the locally deployed first part of the model (e.g., the client part in split learning or the local training process in federated learning). They must adopt sufficient technical and organizational measures to protect local data and the environment. Second, they must fulfill the notification obligation to data subjects. They must clearly explain, in an understandable way, that their personal data will be used for collaborative computation in a de-identified feature form, not by sharing the raw data directly.

Correspondingly, the responsibility of the algorithm developer is strictly limited to the scope of the de-identified information they receive. Their core obligation is to process the received intermediate features or aggregated models subsequently. They must commit to, and in practice refrain from, attempting to reconstruct the original data through attacks like model inversion or membership inference. If a data processor attempts to breach the technical boundary for "re-identification," such action itself constitutes a clear element of illegality or even crime (e.g., the crime of infringing on personal information).

This responsibility division based on technical architecture greatly enhances the practicality of legal enforcement. If a data security incident occurs, the investigation can clearly focus on whether the data owner's local encoder was compromised or the algorithm developer maliciously misused the features. This clarification of responsibility allocation effectively reduces legal uncertainty in data collaboration and provides market participants with stable compliance expectations.

Evolution of Risk Governance: From Absolute Security to Controlled Risk

The application of privacy-preserving technologies also promotes an advancement in the concept of risk governance within the law. It shifts the focus from pursuing unrealistic "absolute anonymity" or "zero risk" to pragmatic "controlled risk" management. A consensus has formed in academia and practice: in the face of continuously evolving re-identification techniques, absolute, non-reversible anonymization exists only in ideal scenarios.

Therefore, the legal evaluation standard for these technologies should not require the complete elimination of risk. Instead, it should assess whether the technology reduces the re-identification risk to an acceptable level. This necessitates a risk-based governance approach. It involves embedding technical and organizational measures to prevent re-identification throughout the entire data processing lifecycle. For instance, the law can require enterprises to conduct continuous risk assessments of their adopted de-identification technologies. It should consider whether the technical difficulty, time, and economic cost required for reconstruction far exceed what a typical actor can bear. This pragmatic legal standard leaves room for technological innovation while ensuring the core objective of privacy protection is achieved dynamically.

Conclusion

Privacy-preserving AI technologies, such as split learning and differential privacy, represent more than engineering progress. They drive a fundamental shift in legal paradigms. Through architectural design, these technologies transform abstract legal principles into executable technical specifications. This enables a crucial move from passive, ex-post punishment to proactive, ex-ante prevention.

The "technical boundaries" established by these architectures clearly redefine the legal responsibilities of all participants. They provide a viable path for the secure flow and value realization of data within a compliant framework.

The future improvement of AI governance relies on a deep integration of technical insight and legal wisdom. Constructing a legal environment that fosters innovation while protecting fundamental rights and clarifying responsibilities will be a key focus for future work.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Acar, A.; Aksu, H.; Uluagac, A. S.; and Conti, M. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4): 1–35.

Calzada, I. 2022. Citizens' data privacy in China: The state of the art of the Personal Information Protection Law (PIPL). *Smart Cities*, 5(3): 1129–1150.

Dwork, C. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, 1–12. Springer.

Goldreich, O. 1998. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110): 1–108.

Li, N.; Li, T.; and Venkatasubramanian, S. 2006. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, 106–115. IEEE.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; and Venkitasubramaniam, M. 2007. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1): 3–es.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05): 557–570.

Thapa, C.; Arachchige, P. C. M.; Camtepe, S.; and Sun, L. 2022. Splitfed: When federated learning meets split learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8485–8493.