Article | Received 7 May 2026; Revised 11 July 2026; Accepted 20 July 2026; Published 6 August 2026
https://doi.org/10.55092/let20260012

Anonymization dilemma of generative artificial intelligence and its countermeasures

Xinhua Fu[1] and Junhan Yin[2,*]

[1] The Data Law Research Center, Beijing Jiaotong University, Beijing, China
[2] School of Law, Beijing Jiaotong University, Beijing, China

* Correspondence author; E-mail: 23261069@bjtu.edu.cn.

Highlights:

- Generative AI heightens the risk that anonymized data may be re-identified.
- Anonymization should move from a one-off legal determination to ongoing risk-based governance.

Abstract: The development of generative AI relies on large-scale corpora that often contain personal information, making anonymization central to data-protection compliance. In comparative law, anonymization generally determines whether personal-information protection rules apply, with the key inquiry turning on identifiability and the reasonable likelihood of re-identification. In generative AI, however, parameterized learning, dynamic content generation, and long-term interaction have transformed anonymization from a static processing outcome into a probabilistic and context-dependent risk condition. This creates three main challenges: a mismatch between formal anonymization and substantive risk, which may lead to identity disclosure or sensitive-attribute inference; the erosion of anonymization's function as a regulatory boundary, which may enable regulatory circumvention and undermine public trust; and uncertainty over responsibility allocation, which may hinder risk prevention, enforcement, and remedies. Existing rules offer only limited responses to residual risks, jointly produced risks, and dynamic changes in model operation. Accordingly, anonymization should be reconceptualized as an ongoing governance framework based on context-sensitive identifiability standards, tiered legal consequences, responsibility allocation, and lifecycle oversight, thereby reconciling personal-information protection with AI development.

Keywords: generative artificial intelligence; anonymization; risk; re-identification risk; risk-based governance; legal determination

1. Introduction

The development of generative AI has made the collection, training, and use of large-scale corpora a key basis for developing and deploying model capabilities [1]. The presence of substantial volumes of personal information in training data makes anonymization a central component of data-protection compliance. Anonymization is a legally significant status determination that directly concerns whether

the relevant information remains personal information and whether personal-information protection rules continue to apply. Accordingly, jurisdictions such as China and the European Union have established rules concerning anonymization to varying degrees.

Existing scholarship on anonymization can be grouped into three main strands. The first strand concerns institutional critique. This strand begins with the difficulties of applying existing law and the functional paradox of anonymization, examining why anonymization rules have failed to perform their intended role as the dividing line between data circulation and personal-information protection [2]. Some scholars have also described anonymization as a broken promise of privacy [3]. The second strand concerns the reconstruction of identifiability criteria. This line of research focuses primarily on how anonymization should be determined, seeking to interpret its meaning in a relative and contextual manner through theories such as “relative anonymization” [4], and “functional anonymization.” It should be noted that current scholarship generally recognizes the relativity of personal information anonymization [5]. The third strand concerns reform of the regulatory model. This approach focuses on how anonymization should operate, shifting the discussion from a static judgment of anonymized results to the continuous control of re-identification risks after anonymization. It emphasizes the need to establish an end-to-end, context-sensitive governance framework tailored to specific settings [6]. These studies have advanced the theory of anonymization from the perspectives of institutional function, identification criteria, and governance logic, and they also provide important theoretical resources for re-examining the normative function of anonymization in the context of generative AI.

Nevertheless, existing research remains limited in that it has not yet offered a sufficient and systematic response to the risks brought about by the technological development of generative AI. In particular, in large-model settings, information relating to individuals may be dispersed across model parameters, embedding representations, retrieval indexes, and generated outputs. Unlike tabular data, unstructured data such as text, images, audio, and video may contain personal information hidden within semantics, context, image features, or voice characteristics, making identification and deletion more difficult [7]. Meanwhile, identification risks may be continuously generated during model deployment, user interaction, and inference output. Anonymization assessments therefore require a clear specification of both the assessment time and the relevant technical layer. Furthermore, where risks are jointly produced by the conduct of multiple parties, anonymization must also be connected with the existing legal system of obligations and rules of liability.

On this basis, this article further focuses on two core questions in the context of generative AI. First, under the technical conditions of parameterized learning, distributed storage, and dynamic generation, does the normative foundation of anonymization as a regulatory boundary still hold? Second, within a processing chain in which training, deployment, and inference are intertwined, how should the scope, timing, and legal consequences of anonymization assessments be redefined?

This article argues that generative AI has not rendered anonymization meaningless, but it has indeed affected the manner in which anonymization is legally evaluated. Therefore, anonymization should no longer be understood as a processing result that, once completed, can be used to determine whether information constitutes personal information. Instead, it should be reconstructed as an ongoing risk-based governance mechanism embedded throughout the entire life cycle of AI, with the control of re-identification risks at its core.

2. The legal function and technical logic of anonymization

2.1. Anonymization as a boundary of personal-information protection rules

To understand why anonymization rules face new difficulties in the context of generative AI, it is necessary first to examine the concept and function of anonymization. Anonymization primarily serves to delimit the scope of personal-information protection rules. Where information can be directly or indirectly linked to a specific individual, the relevant processing activities are subject to those rules. Where, after processing, information can no longer identify a specific individual and cannot readily be restored or re-associated with that individual, it is, in principle, no longer governed by personal-information protection rules.

From a comparative-law perspective, although different jurisdictions adopt different terminology and institutional structures, they all recognize, to varying degrees, that anonymization affects the scope of personal-information protection rules. Anonymization is therefore not merely a technical processing measure, but a legally significant mechanism for determining regulatory status. The core question is whether the relevant information should still be treated as personal information and whether corresponding compliance obligations should continue to apply.

To further illustrate the normative foundations of anonymization in different jurisdictions, this article examines representative rules from China, the European Union, California in the United States, and Japan.

As Table 1 shows, major jurisdictions generally impose relatively stringent anonymization requirements, yet they differ in terminology, assessment criteria, and associated legal consequences. Existing scholarship has summarized the legal standards for anonymization in Europe, the United States, and China respectively as the "exhaustion of all possibilities standard," the "privacy-rights-based protection standard," and the "non-identifiability and irreversibility standard" [8]. Taking China's Personal Information Protection Law of the People's Republic of China (PIPL) as an example, it employs the concept of "anonymization" and emphasizes that, after processing, personal information must reach a state in which "a specific natural person cannot be identified and the information cannot be restored." Once the relevant information satisfies the requirement of being "unable to identify a specific natural person and unable to be restored," it is, in principle, excluded from the direct regulation of the Personal Information Protection Law.

EU law does not provide a unified statutory definition of "anonymization" in its binding provisions. Instead, Recital 26 of the General Data Protection Regulation (GDPR) indicates by implication that anonymous information falls outside the scope of data-protection rules. At the same time, EU law clearly distinguishes anonymization from pseudonymization: the latter reduces the risk of direct identification but does not, by itself, remove the relevant data from the personal-data protection regime.

The California California Consumer Privacy Act (CCPA), by contrast, more frequently employs concepts such as "deidentified information," "aggregate consumer information," and "pseudonymized information." Qualified deidentified information and aggregate consumer information may be excluded from the scope of personal information. Japan's Act on the Protection of Personal Information (APPI) establishes two categories: "anonymously processed information" and "pseudonymously processed information." The former is closer to a legal regime governing the use of information after anonymization, whereas the latter retains a stronger connection with personal information protection norms.

Table 1. Legal bases and rules on anonymization in major jurisdictions.

Jurisdictions	Legal Basis	Relevant Legal Provisions
China	PIPL [9]	Article 4 "Personal information" refers to various information related to an identified or identifiable natural person recorded electronically or by other means, but does not include anonymized information; Article 73 For purposes of this Law, the following terms shall have the following meanings: (4) "anonymization" refers to the process of processing personal information to make it impossible to identify specific natural persons and impossible to restore.
the European Union	GDPR [10]	Article 4(1) defines the scope of "personal data"; Article 4(5) provides for "pseudonymisation"; and Recital 26 provides that the principles of data protection apply to information relating to an identified or identifiable natural person, that anonymous information falls outside the scope of the GDPR, and that, in determining identifiability, account should be taken of all the means reasonably likely to be used.
California, USA	CCPA [11]	Section 1798.140(b) defines "aggregate consumer information"; Section 1798.140(m) defines "deidentified information"; Section 1798.140(v)(3) provides that personal information does not include deidentified information or aggregate consumer information; and Section 1798.140(aa) defines "pseudonymization."
Japan	APPI [12]	Article 2, paragraph 5 provides for "pseudonymously processed information"; Article 2, paragraph 6 provides for "anonymously processed information"; Article 43 sets out the rules for the creation of anonymously processed information; Article 44 provides for the obligations of publication and express indication when anonymously processed information is provided to a third party; and Article 45 prohibits identifying the principal by means such as comparison with other information.

The relevant concepts across different jurisdictions therefore cannot simply be equated. Anonymization is not a single legal concept; rather, it is a legal status shaped by identifiability, reversibility, reasonably foreseeable re-identification risks, and controls over subsequent use.

In traditional data-processing contexts, this regulatory arrangement allows anonymization to operate as a relatively stable regulatory boundary. Some scholars have noted that "data protection law protects individuals against risks associated with the processing of 'personal data' [13]. In traditional Internet judicial practice, courts have generally adopted a tolerant attitude toward the commercial use of anonymized information. For example, in China's Zhu v. Baidu case, the court held that Baidu's use of cookie technology to collect users' anonymized search keywords for personalized recommendations did not constitute an infringement of privacy, because Baidu had fulfilled its notification obligation and provided an opt-out mechanism [14]. Similarly, in China's Taobao v. Meijing case, the court not only recognized the legality of Taobao's in-depth development of anonymized behavioral information, but

also conferred upon such information independent competitive property interests [15]. The above examples show that, in traditional Internet scenarios, anonymization once served to delineate the institutional boundary between personal information protection and the commercial use of data: information that had undergone effective desensitization was generally regarded as having fallen outside the scope of personal information protection.

2.2. The probabilistic nature and relative characteristics of technical anonymity

From a technical perspective, anonymization does not eliminate identification risk altogether. Rather, by controlling data distribution, attribute combinations, and modes of information exposure, it reduces the likelihood of identifying an individual to an acceptable level.

As the scale of data has expanded, traditional desensitization methods have gradually become insufficient to meet the needs of privacy protection. Consequently, a series of anonymization mechanisms based on statistical theory have emerged.

Among these mechanisms, L. Sweeney and others first proposed and developed the k-anonymity method [16]. The k-anonymity model reduces the possibility of individual identification by ensuring that each record is identical to at least k-1 other records with respect to key attributes. It is usually implemented "through generalization and suppression" [17]. Since the k-anonymity model may still be exposed to problems such as the disclosure of sensitive attributes, subsequent studies further proposed models such as l-diversity [18] and t-closeness [19]. The above models function more as evaluation criteria or risk-control approaches for anonymization. In concrete implementation, however, anonymization must also rely on various technical means. Existing research has pointed out that "the anonymization methods currently proposed include generalization, suppression, clustering, microaggregation, decomposition, and permutation" [20]. Taking generalization as an example, data generalization reduces identification risk by lowering data precision. For instance, replacing the specific age of 15 with the age range of 10–20 makes it difficult to accurately restore the original data.

Unlike the anonymization models discussed above, which were developed mainly for data publication, differential privacy focuses on controlling the influence of a single individual's data on aggregate outputs in statistical queries or published results. By adding random noise during data queries or publication, it ensures that the presence or absence of a single individual's data does not significantly affect the overall statistical results, thereby mathematically limiting the risk of information leakage.

However, even where the above technical methods are adopted, anonymized data may still face the risk of re-identification. With the continuous accumulation of data resources, attackers may re-establish the link between data and specific individuals through methods such as correlation with external datasets, supplementation with background knowledge, or model-based inference. Re-identification risk should therefore not be understood in absolute terms. An empirical study by Nyffenegger *et al.* on court judgment texts shows that, in the experimental scenarios they designed, the overall success rate of using large language models to re-identify anonymized judicial documents was relatively low, and the risk was not significant in most cases [21]; Alexin calculated, from a statistical perspective, that the re-identification risk of a specific database after anonymization was below 0.1%, thereby demonstrating that "fair anonymization" does indeed exist [22]. These studies indicate that generative AI does not necessarily render anonymization ineffective, nor can it be simply presumed that anonymized data will inevitably face a high risk of re-identification. The central concern is the highly context-dependent nature of

re-identification risk: data type, textual structure, external auxiliary information, model capability, prompting methods, the attacker's objectives, and the specific use setting may all affect the effectiveness of anonymization in practice. As demonstrated by Altman and other scholars in their discussion of an integrated "legal–technical" governance framework [23], anonymization should be understood more as a risk-management mechanism jointly shaped by law and technology. Its objective is not to completely eliminate every possibility of identification, but to achieve a balance between data use and privacy protection within an acceptable range of risk.

Technical anonymity has therefore been inherently relative from the outset. The effectiveness of anonymization depends on multiple factors, including sample size, attribute dimensions, external knowledge, the attacker's capabilities, and the specific use setting. It therefore cannot be treated as a once-and-for-all determination made at the completion of processing; rather, it protects privacy primarily by reducing the likelihood of re-identification. As relevant research has noted, "anonymization is not an algorithm applied to a dataset to achieve measurable security, but an ongoing process aimed at reducing the risk of data re-identification to an acceptable level" [24].

2.3. The transformation of information forms in generative AI

Generative AI does not fundamentally alter the relative nature of anonymization. Rather, it changes the forms in which personal information exists, the pathways through which identification risks arise, and the allocation of responsibility across multiple actors. In generative AI systems, particularly large language models, information relating to individuals may, during the training process, no longer exist in the form of complete records, but may instead be transformed into part of model parameters [25], text embeddings [26], retrieval-augmented generation (RAG) [27], or generative capabilities. In other words, the assessment of anonymization can no longer remain confined to the stage before training data enters the model. It must further ask whether identifiable personal information may still be contained in, or reconstructed from, model parameters, embedding representations, retrieval components, context windows, and generated outputs. This change has fundamentally shifted the object to which anonymization is directed. As the UK Information Commissioner's Office (ICO) has pointed out, during the training process, generative AI models do not merely parse personal data in the same way as traditional software; rather, by "learning" from such data, they may retain its imprint [28]. Accordingly, personal-related information no longer exists merely in the form of records in traditional datasets, but may instead be transformed into a distributed existence within the model system.

Dynamic output generation in generative AI further amplifies this problem. Traditional anonymized data usually takes the form of relatively fixed content at the time of release, and identification risk mainly arises when an attacker cross-compares that dataset with external data. Generative AI operates differently. Its outputs are not predetermined, but are generated in real time through the combined operation of user prompts, contextual information, model parameters, and external tools. Accordingly, identification risk does not necessarily manifest as the direct restoration of the original data. Rather, it may take the form of generating names, contact information, biographical trajectories, identity profiles, or sensitive attributes relating to a specific subject through prompt inducement, statistical association, and contextual inference.

Although some model service providers reduce the risks associated with the processing of personal information through measures such as data filtering, training constraints, and output filtering, Open AI,

for example, states in its explanation of model training materials that "we take active steps to limit the processing of personal information during training. For example, we exclude sources that aggregate large amounts of personal data and train models not to respond to requests for private or sensitive information about individuals" [29]. However, relevant research and practice still show that, under certain conditions, large models may memorize and output personal information or sensitive content contained in training corpora [30]. Furthermore, generative AI is evolving from single-turn question-answering systems into agentic systems with capabilities such as persistent memory, task planning, tool use, and environmental perception. In agentic scenarios, identification risk no longer arises merely from a particular model output, but may gradually accumulate over sustained interactions. Fragmented information provided by users in different rounds may, in isolation, be insufficient to identify a specific individual. Yet under the combined effects of the agent's memory mechanism, task continuity, and ability to call external tools, such information may be continuously integrated, associated, and inferred, ultimately forming a relatively complete personal profile.

In sum, the key difference between generative AI and traditional anonymized data lies not merely in the larger scale of data or the greater complexity of processing techniques, but in the changing forms in which information relating to individuals exists, the ways in which outputs are generated, and the pathways through which risks arise. Accordingly, the scope, timing, and basis for allocating responsibility in anonymization assessments have also changed.

3. The multiple dilemmas of anonymization in the context of generative AI

As discussed above, generative AI does not merely expand the scale of data processing; it reshapes how information relating to individuals exists, is generated, and gives rise to risk. This challenges the assumptions underlying anonymization and creates difficulties in legal determination, regulatory consequences, and responsibility allocation. Yet technological change alone does not justify stronger regulation. Reconsidering anonymization becomes necessary only when its failure may cause individual harms, such as identity disclosure, sensitive attribute inference, or ineffective remedies, or public harms, such as regulatory arbitrage, legal uncertainty, and declining public trust, while existing legal duties and remedies cannot adequately address these risks.

Accordingly, the following discussion focuses on three linked issues: the mismatch between formal anonymization and substantive risk, which distorts risk assessment; the erosion of anonymization's function as a regulatory boundary, which destabilizes legal application and regulatory consequences; and uncertainty over responsibility allocation, which further impairs accountability and remedies. Together, these issues justify shifting anonymization from a final determination to ongoing risk-based governance.

3.1. The mismatch between formal anonymity and substantive risk

The first dilemma arises at the level of determining anonymization. A mismatch may arise between the legal question whether the data qualifies as anonymized and the technical reality that the model system may still generate or reconstruct identifiable information.

This mismatch is first reflected in the expansion of the assessment object from datasets to model systems. Traditional anonymization mainly assesses whether original or released data remains identifiable. In the context of generative AI, however, the issue is no longer limited to whether the

original training data has been anonymized. It also includes whether the model system and its outputs may once again refer to, or identify, a specific individual. Existing research has pointed out that the development of generative AI has transformed the mechanism through which risks relating to personal data arise. In other words, such risks no longer stem solely from the data itself, but increasingly arise from the model capabilities developed through training, inference, and generation [31]. Therefore, if the assessment of anonymization remains confined to the stage of data input, it may overlook the identification risks generated during the operation of the model.

This mismatch is also reflected in changes in the manner of re-identification. In traditional scenarios, re-identification is usually understood as reconnecting anonymized data to a specific individual, for example, by restoring identity information such as names, addresses, or account numbers through cross-comparison. In generative AI, however, the risk does not necessarily take the form of the complete reconstruction of a particular original record. Rather, it may manifest as identity targeting, profile reconstruction, the stitching together of personal experiences, or the inference of sensitive attributes. Even if a model does not reproduce the original training data word for word, substantive identification risk may already arise so long as it is capable of generating information sufficient to point to a specific individual with the assistance of particular prompts, contextual information, or external data.

At the same time, generative AI may further amplify both the mosaic effect and the aggregation effect. The "mosaic effect" refers to a phenomenon in which, "like tiny shards of rock and glass, bits and bytes of data are pieced together in ways that create a fuller picture of you" [32]. This phenomenon is closely related to the aggregation effect [33], namely, that even where a single item of data has been anonymized, individuals may still be re-identified through correlational analysis across different data sources. Existing research has already noted that "the wide adoption of machine learning to solve a large set of real-life problems came with the need to collect and process large volumes of data, some of which are considered personal and sensitive, raising serious concerns about data protection" [34]. This is particularly true in agentic scenarios, where long-term memory, multi-round interaction, tool calling, and task coordination enable fragmented information to be continuously accumulated, associated, and reactivated.

It can thus be seen that generative AI has changed the risk structure confronting anonymization. Formally, data may already have undergone anonymization. Substantively, however, the model system may still generate, reconstruct, or infer information relating to a specific individual during subsequent operation. It is in this sense that generative AI gives rise to a mismatch in anonymization, namely, a situation in which data is formally anonymized while substantive risks remain.

Accordingly, the first dilemma is not merely a matter of technical misjudgment; it directly concerns whether individual rights and interests fall within the scope of legal protection. The model's reconstruction of identity, personal experiences, or sensitive attributes may lead to harassment, fraud, discriminatory decision-making, and harm to personality interests. More importantly, if upstream data have already been classified as anonymized, the relevant processing activities may prematurely fall outside obligations relating to notice, purpose limitation, security safeguards, and responses to individual rights requests.

Existing law is not entirely powerless where model outputs have reconstituted personal information or caused infringement. However, such ex post remedies often depend on the victim's discovery of the risk and ability to prove subject-specific linkage and causation. They are therefore difficult to apply where the risk remains at the level of model capability or continues to accumulate over time.

3.2. The erosion of anonymization's function as a regulatory boundary

If the foregoing mismatch mainly concerns technical risk, the erosion of anonymization's function as a regulatory boundary creates uncertainty over legal consequences. Anonymization was originally intended to distinguish personal information from non-personal information, thereby providing a basis for determining when compliance obligations are triggered and when they cease to apply. In the context of generative AI, however, risks may arise and evolve across training, deployment, and inference, making this boundary unstable.

Although the Netflix case occurred in the traditional context of data release, it illustrates the inherent fragility of the anonymization boundary. As early as 2006, Netflix publicly released 100 million anonymized movie ratings and offered a prize of USD 1 million for an algorithmic competition. Researchers later found, however, that even though user IDs had been replaced, individuals could still be re-identified by cross-comparing the data with publicly available information such as IMDb data. Some students even succeeded in identifying several users, including themselves [35]; Generative AI further amplifies this fragility. Foundation models are "based on standard deep learning and transfer learning, but their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization" [36], and, through training on massive corpora, form parameterized representations across contexts, sources, and tasks. Fragments that appear non-identifying within a single dataset may be reconnected with other information during model training and generation, thereby increasing the possibility of identifying individuals. In other words, anonymization risk no longer depends solely on whether a particular dataset itself has been "de-identified," but increasingly depends on the model system's capacity to reorganize, retrieve, and infer from the data.

Further research shows that large language models can even re-identify Hacker News users and human interview participants with high accuracy based only on pseudonymous online profiles and conversations, achieving results close to those produced by professional investigators after hours of work [37].

Accordingly, re-identification risk is no longer a simple binary matter that arises only before or after anonymization is completed, nor does its source depend solely on whether the original data is retained. Rather, it depends increasingly on whether the system, in its subsequent operation, retains the capacity to generate, aggregate, or infer information relating to a data subject. Even if data has been formally anonymized, this does not necessarily mean that it has, in substantive terms, been removed from identification risk. More troublingly, anonymization may also become a regulatory grey area for large-scale data processing. Some scholars have pointed out that "anonymization in fact provides technology developers with a grey zone, enabling them to conduct large-scale data processing and analysis without violating the law, while disregarding the impact of subsequent data generation activities on the rights and interests of individuals or groups" [38]. Because different deployment methods, application settings, and interaction patterns may continuously alter the level of risk, and because, in agentic settings, long-term memory, task continuity, and environmental linkage may cause subject-related associations to accumulate over time, a legal regime that treats a one-off determination as a sufficient basis for regulatory exclusion may allow anonymization to preserve the boundary formally while failing to constrain subsequent risks in practice.

The second dilemma concerns both individual and public interests. For individuals, unstable boundaries may result in different levels of protection for materially similar risks. For public governance,

if “anonymized” status is treated as a sufficient basis for permanent exclusion from regulatory rules, operators may use formalistic processing to evade obligations relating to lawful basis, impact assessment, purpose limitation, and supervisory audits, thereby creating systemic regulatory arbitrage.

Existing personal-information protection law uses qualification as personal information as its threshold condition. This approach has the advantage of clarity and operability, but it lacks nuanced arrangements for intermediate situations in which risks have been substantially reduced but not completely eliminated.

3.3. Uncertainty in allocating responsibility for anonymization

Generative AI makes responsibility for anonymization harder to allocate. The allocation of responsibility under traditional personal-information protection law usually presupposes relatively clear processing purposes, processing means, and processing entities. Whether in the distinction between controllers and processors, or in the division between personal information processors and entrusted processors, the basic logic is that the subject determining the purpose and means of processing bears the primary compliance obligations, while the subject that specifically carries out the processing activity assumes corresponding responsibility. However, in large model and agentic scenarios, data provision, pretraining, fine tuning, deployment, application integration, plugin calling, and user use are often carried out by different actors, and re-identification risks may also be jointly produced across multiple stages. As a result, the traditional single-actor model of attribution cannot fully account for the way such risks emerge.

Furthermore, model developers may not interact directly with end users, deployers may not possess the underlying training data, and application operators may substantially alter model outputs and identification risks through prompt design, knowledge-base access, workflow orchestration, and long-term memory mechanisms. Accordingly, anonymization risks are often not directly caused by a single actor, but are the result of the combined effects of technical decisions, organizational arrangements, and use behaviors by multiple actors at different stages.

In agentic scenarios, this difficulty in allocating responsibility becomes even more pronounced. Through long-term memory, task orchestration, tool calling, and environmental feedback, agents continuously integrate user information, causing identification risks to span multiple technical layers and points in time. Such risks can no longer be simply attributed to a particular output or to the conduct of a particular actor. Since AI systems themselves do not possess intentions in the legal sense [39], traditional liability rules centered on the subjective state of the actor are difficult to apply directly. Legal responsibility must therefore be allocated among developers, deployers, operators, and users.

At the same time, the emergence of risk may be jointly affected by historical conversations, external knowledge bases, system prompts, results returned by plugins, and the user’s current input. Thus, the responsibility question in anonymization is no longer merely one of identifying a single legally responsible actor. Rather, it requires looking beyond the formal division of labor among modules, stages, and actors to identify those who exercise substantive control over the creation of risk.

The regulatory concern underlying the third dilemma is that uncertainty over responsibility allocation weakens both ex ante prevention and ex post remedies. Individuals often cannot know whether the risk originates from training data, model parameters, retrieval-augmented knowledge bases, system prompts, long-term memory, or user input. Regulators may also find it difficult to reconstruct the risk pathway in the absence of logs and cooperation mechanisms.

Existing rules on personal information processors, entrusted processors, and joint processing can cover certain cooperative relationships. However, responsibility attribution under this framework still largely depends on clear control over the purposes and means of processing. Where model developers, deployers, application operators, and tool providers each control different risk factors, the framework may produce both overlapping obligations and responsibility gaps.

4. From one-off legal determination to ongoing risk-based governance

As discussed above, the development of generative AI has altered several premises on which anonymization traditionally rests, making it necessary to move from a one-off legal determination to ongoing, risk-based governance. This shift should be grounded in existing personal-information protection norms and proceed along four dimensions: identifiability standards, legal consequences, responsibility allocation, and accountability mechanisms.

At its core, this approach seeks to transform anonymization from a one-off assessment of results into a dynamic governance framework spanning the process of data processing, model operation, and application deployment. In this way, re-identification risks can be continuously assessed and addressed in a tiered manner, while providing a basis for identifying and allocating responsibility in multi-actor settings.

4.1. Contextual adjustment of identification standards

4.1.1. From a static state of anonymity to dynamic assessment of re-identification risk

Under the traditional framework of anonymization, once data meets the standard of anonymity, it is usually regarded as falling outside the scope of personal information protection rules. In the context of generative AI, however, identification risk may fluctuate dynamically with changes in model capabilities, access to external data, modes of interaction, and application scenarios.

More specifically, anonymization should exclude information from personal-information protection rules only where the information no longer enables substantive identification of a data subject under ordinary technical conditions, within the scope of reasonably available auxiliary information, and in the specific use context. Conversely, if the model system may still generate, aggregate, or infer information pointing to a specific individual through prompt inducement, retrieval augmentation, multi-turn interaction, long-term memory, or access to external knowledge, it cannot be assumed to fall outside the scope of personal-information protection rules merely because the input data has already been anonymized. In other words, in the context of generative AI, anonymization should be understood as a legal status subject to continuous risk assessment, rather than a fixed result that comes to an end once processing has been completed. Existing research has pointed out the need to "promote the shift of anonymization from static absoluteness to dynamic relativity, and, on the premise that risks are controllable, to explore lawful pathways for the use of personal information, thereby establishing a coordinable middle ground between data use and the protection of rights and interests" [40]. This also indirectly confirms the necessity of making anonymization standards more dynamic and context-specific. At the same time, relevant views in comparative law may also support the foregoing assessment. The UK Information Commissioner's Office, in its guidance on anonymization, emphasizes that "anonymisation is intended to reduce the risk of an individual being identified, or being identifiable, to a

sufficiently remote level. This will depend on a range of factors, specific to the context" [41]. It can thus be seen that the assessment of anonymization is itself context-dependent. Introducing this understanding into the context of generative AI requires legal evaluation to pay further attention to the impact of model capabilities, the data environment, external tools, and interactive processes on re-identification risk. This context-dependent approach is further supported by the European Data Protection Board's (EDPB's) Guidelines 02/2026 on Anonymisation, adopted for public consultation in July 2026 [42]. By emphasizing the means reasonably likely to be used by relevant entities and the need for reassessment as technologies and data environments evolve, the Guidelines reinforce the view that anonymization should be treated as a status requiring continuing review rather than a one-off determination.

4.1.2. A three-pronged assessment of re-identification risk

In the context of generative AI, the assessment of anonymization should not remain confined to whether direct identifiers have been removed from the data. Rather, it should shift to whether the relevant information still presents a reasonably foreseeable risk of re-identification during the operation of the model system. The so-called reasonably foreseeable risk of re-identification does not refer to any abstract or theoretical possibility of identification, but to the real possibility that, under specific technical conditions, data environments, and use scenarios, the relevant information may once again point to a specific individual.

Drawing on the provisions of the Opinion on Anonymisation Techniques [43], this risk may be assessed along three dimensions: incentive, capability, and context. First, with respect to incentive, the assessment should examine whether the relevant actors have a real incentive, purpose of use, or interest-based connection for re-identification. If the possibility of identification remains only a highly abstract theoretical assumption and lacks any real interest-driven motivation, anonymization should not be rejected on that basis alone. If, however, the relevant actors attempt to carry out re-identification through intelligent algorithms and benefit from it, they should bear the corresponding risks arising therefrom [44]. Second, with respect to capability, the assessment should examine whether the relevant actors possess the technical capability, data resources, and organizational conditions necessary to carry out re-identification. In the context of generative AI, particular attention should be paid to whether the model may, through prompt inducement, contextual accumulation, retrieval augmentation, long-term memory, or tool calling, generate, complete, stitch together, or infer information relating to a specific individual.

Third, with respect to context, the assessment should examine the data-processing setting, user scope, access permissions, access to external knowledge, output recipients, and technical and organizational safeguards. The re-identification risk of the same information is clearly different in a closed testing environment and in an agentic system that is open to the public, connected to external knowledge bases, and equipped with long-term memory. As relevant research has noted, "language models like GPT and BERT are trained on large text datasets to predict or generate sequences of text. These models have the potential to memorize specific details from training data, which could include sensitive information such as health records, biometric data, sexual orientation, or other information shared during interactions with language models" [45]. Similarly, the Italian Garante's temporary restriction on ChatGPT in 2023 also shows that, in the context of generative AI, regulators will not be satisfied merely with formal explanations of data processing. Rather, they will further examine the overall lawfulness of personal data processing in model training, content generation, and service operation [46].

4.2. Tiered legal consequences of anonymization

In the context of generative AI, the key issue is how the legal consequences of anonymization should vary according to different degrees of re-identification risk. Although de-identified information has had direct identifiers removed, it may still present varying degrees of re-identification risk under specific technical conditions, with the assistance of auxiliary information and data aggregation capabilities. Anonymization can be established only where it can effectively prevent risks of singling out, linkability, and attribute inference. Merely deleting direct identifiers such as names and identification numbers is not sufficient to remove data from the regulation of personal data [47]. If exclusion from personal-information protection laws and regulations remains the sole legal consequence, the law will be unable to respond accurately to different degrees of re-identification risk. In this sense, the legal consequences of anonymization may be divided into three tiers: high re-identification risk, controlled anonymization, and regulatory exclusion based on anonymization.

First, the high re-identification risk tier. In scenarios where the risk of re-identification is relatively high, or where model outputs may directly generate subject-related information such as names, contact information, identity trajectories, or sensitive attributes, stricter personal information protection rules should continue to apply. These include a clear basis for processing, necessity limitations, protection of sensitive information, access control, testing and auditing, and ex post remedies.

Second, the controlled anonymization tier. In scenarios where measures such as de-identification, access restrictions, output filtering, control of retrieval scope, and memory isolation have been adopted, and where the risk of re-identification has been significantly reduced but not completely eliminated, relatively flexible use may be permitted for specific purposes, in specific scenarios, and under specific control conditions. However, basic requirements such as risk assessment, purpose limitation, security safeguards, and prohibition of re-identification should still be retained.

Third, the regulatory-exclusion tier based on anonymization. Only where, after a comprehensive assessment, the relevant information no longer presents a reasonably foreseeable risk of re-identification, and cannot readily be re-linked to a specific individual under ordinary technical conditions, within the scope of reasonably available auxiliary information, and in the specific use context, should it be excluded from the personal-information protection regime.

Existing research has also pointed out that anonymization requires individualized arrangements based on the specific context of information processing, and that a single uniform standard is unable to respond to differences in risk across different scenarios [48]. At the same time, the boundary between anonymous data and personal data is not absolutely clear in practice, but should instead be assessed in light of the specific technological and application context [13]. Rubinstein and Hartzog also caution that anonymization should not be understood as an absolutely completed state, since doing so may easily obscure the risks of re-identification and disclosure of sensitive attributes [49]. A tiered approach to the legal consequences of anonymization can therefore preserve anonymization's function as a regulatory boundary while leaving institutional space for low-risk data use in generative AI settings.

4.3. Clarifying responsibility for anonymization in multiple actor scenarios

In the context of generative AI, anonymization-related risks usually do not result from the conduct of a single actor, but from the combined operation of multiple stages, including training, deployment,

operation, and use. Existing research has already pointed out that liability in generative AI may involve multiple actors, and may simultaneously involve different actors such as service providers, deployers, and users [50]. Against this background, the allocation of responsibility for anonymization needs to "determine the content of the rights and obligations of specific actors according to the source of social risk, the stage at which it occurs, and the consequences of harm" [51], thereby moving from single-actor attribution to a structured model of responsibility allocation across multiple actors and multiple stages.

Therefore, responsibility allocation should no longer rely mechanically on traditional attribution models, but should instead adopt a three-dimensional framework based on control over risk, benefit derived, and foreseeability. Specifically, actors that control model capabilities, data access, system memory, output filtering, and use settings should bear preventive and governance obligations corresponding to their control capacity. Actors that directly or indirectly benefit from the relevant risk-creating activities should also bear duties of care or risk-sharing obligations corresponding to the degree of benefit they receive. Where a particular actor has a high degree of foreseeability with respect to a specific re-identification risk, or ought to have foreseen such risk but failed to fulfill a reasonable duty of care, the basis for its responsibility should likewise be strengthened accordingly.

On this basis, anonymization should also introduce a mechanism for coordinated responsibility allocation among multiple actors. Coordinated attribution does not simply expand the scope of actors bearing responsibility. Rather, where identification risks are jointly triggered or amplified by multiple parties, responsibility should be reasonably allocated according to the functional role played by different actors in the process of technological implementation and the types of obligations they bear. Data providers should be responsible for data sources, initial processing, and information quality. Model developers should be responsible for training mechanisms, the risk of parameter memorization, basic security capabilities, and security boundaries. Deployers and application operators should be responsible for knowledge access, long-term memory settings, output control, and user interaction mechanisms in specific scenarios. Users who knowingly exploit system defects to actively carry out re-identification should also bear corresponding legal consequences. The objective is to recognize shared responsibilities and obligations in the design, deployment, and use of complex technologies such as AI, thereby potentially avoiding or mitigating responsibility gaps [52].

4.4. Lifecycle governance and accountability mechanisms

Anonymization-related risks in generative AI are not concentrated in a single stage, but arise across input, training, deployment, inference, and feedback. It is necessary to "follow the logic of governance throughout the entire life cycle of AI, guide the establishment of risk management and response mechanisms by relevant actors across the whole process, and, in light of the characteristics of AI applications, strengthen the monitoring and management obligations of relevant actors with respect to AI applications after they have been placed on the market" [53]. Therefore, the reconstruction of anonymization must be embedded within a framework of lifecycle governance. This lifecycle-based approach is also broadly consistent with the EU Artificial Intelligence Act. Although the Act does not establish an independent test for anonymization, its risk-based framework and requirements concerning documentation, logging, and post-market monitoring support the continuous assessment of anonymization risks throughout the operation of AI systems [54]. Stalla-Bourdillon and Knight criticize the anonymization approach of "release and forget," arguing that anonymization is not the end point of

responsibility after data release. Rather, its effectiveness must be continuously assessed in light of the overall data environment, including the data itself, the technical infrastructure, data recipients, subsequent processing purposes, and externally linkable data [55]. Specifically, at the input stage, mechanisms for data screening, classification, and front-end de-identification should be established with regard to data sources, processing purposes, sensitive attributes, combinations of weak identifiers, and context-specific risks. At the operational stage, auditing and testing should be strengthened with respect to model memorization, parameter-leakage risks, retrieval-augmented modules, and long-term memory mechanisms. At the output stage, scenario-specific filtering, blocking, warning, and correction mechanisms should be adopted to prevent the system from generating information that directly or indirectly points to a specific data subject. In this way, anonymization is no longer an isolated data-processing act, but a governance arrangement embedded throughout the operation of the model system.

At the same time, accountability mechanisms compatible with process-based governance should also be established. First, process documentation and risk traceability should be strengthened through technical logs, operational records, and risk-assessment documentation, so that, where re-identification risks or information leakage occur, the pathways through which such risks emerge can be traced and the stages at which responsibility arises can be identified. Second, audit, reporting, and external supervision mechanisms proportionate to the level of risk should be introduced. For high-risk models and large-scale personal information processing activities, anonymization risk testing and security assessments should be conducted on a continuous basis, and major risk incidents should be reported to regulatory authorities in a timely manner. Where entities continuously generate high-risk outputs, fail to fulfill risk-control obligations, or intentionally circumvent anonymization requirements, regulatory authorities may impose measures such as warnings, orders to take corrective action within a specified period, service suspension, and administrative penalties in accordance with the law. In this way, anonymization should no longer be understood merely as an isolated data-processing operation, but rather as a continuous governance arrangement embedded throughout the operation of AI systems.

5. Conclusion

This article does not call for stronger regulation merely because generative AI has the capacity for re-identification. Rather, reconsideration is warranted because the above challenges may harm individual rights and interests and impair public governance, while existing rules centered on one-off anonymization outcomes and a single processing actor are insufficient to address these risks. The mismatch between formal anonymization and substantive risk may allow risks such as identity disclosure and sensitive attribute inference to escape ex ante control; the erosion of anonymization's function as a regulatory boundary may lead to unequal protection and regulatory arbitrage; and uncertainty over responsibility allocation may increase the costs of proof, enforcement, and remedies.

Accordingly, anonymization has not lost its significance in the age of generative AI. However, its legal assessment should shift from a static, outcome-based determination to ongoing risk-based governance. Existing rules can still serve as the foundation for governance, but they should be strengthened through context-sensitive identifiability standards, tiered legal consequences, role-based responsibility allocation, and dynamic lifecycle review. Corresponding obligations should also be reinstated in a timely manner when re-identification risks increase. In this way, anonymization can

continue to function as a regulatory boundary between data use and personal-information protection, while avoiding becoming a formalistic device for avoiding regulation and responsibility.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT, developed by OpenAI, solely to improve the clarity, grammar, and readability of the language. ChatGPT was not used to generate substantive arguments, conduct analyses, interpret results, or formulate conclusions. All intellectual content and scientific conclusions are the authors' own. The authors reviewed and revised all AI-assisted text and take full responsibility for the integrity, originality, and accuracy of the manuscript.

Acknowledgments

This work was funded by the 2025 Beijing Social Science Fund Planning Project, "Research on the Reconstruction of Personal Information Protection Norms for Generative Artificial Intelligence," with grant number 25BJ03038.

Authors' contribution

Conceptualization, X.F. and J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, X.F.; funding acquisition, X.F. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Liu Y, Cao J, Liu C, Ding K, Jin L. Datasets for large language models: a comprehensive survey. *Artif. Intell. Rev.* 2025, 58(12):403.

[2] Xu K. Resurrection of the zombie clause: the reformation of the personal information anonymization system (In Chinese). *Law Econ.* 2024, 2024(4):160–177.

[3] Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Rev.* 2010, 57(6):1701–1777.

[4] Fan W. Rethinking of personal data in big-data age (In Chinese). *Inf. Secur. Commun. Priv.* 2016, 2016(10):70–80.

[5] Zhao J. The theoretical foundation and institutional construction of personal information anonymization (In Chinese). *Peking Univ. Law J.* 2024, 36(2):326–345.

[6] Zheng J. The construction of systematic norms of data anonymization (In Chinese). *J. Polit. Sci. Law* 2022, 2022(4):61–71.

[7] Weitzenboeck EM, Lison P, Cyndecka M, Langford M. The GDPR and unstructured data: is anonymization possible? *Int. Data Priv. Law.* 2022, 12(3):184–206.

[8] Li Y, Chen Q. Discussion on legal standards of taxpayers' information anonymization in the context of tax data management (In Chinese). *Tax. Econ.* 2020, 2020(4):71–80.

[9] The National People's Congress of the People's Republic of China. Personal information protection law of the People's Republic of China. 2021. Available: http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm (accessed on 3 May 2026).

[10] European Union. 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) (Text with EEA relevance). 2016. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 3 May 2026).

[11] Rob B. California Consumer Privacy Act (CCPA). 2024. Available: https://www.oag.ca.gov/privacy/ccpa (accessed on 3 May 2026).

[12] Japanese Law Translation. Act on the protection of personal information. 2003. Available: https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en (accessed on 3 May 2026).

[13] Rupp V, von Grafenstein M. Clarifying "personal data" and the role of anonymisation in data protection law: including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection. *Comput. Law Secur. Rev.* 2024, 52:105932.

[14] Nanjing Intermediate People's Court. Civil Judgment (2014) Ning Min Zhong Zi No. 5028. 2014.

[15] Zhejiang High People's Court. Civil Retrial Application (2019) Zhe Min Shen No. 1209. 2019.

[16] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Seattle, USA, June 1–3, 1998, p. 188.

[17] Wang P, Wang J. Survey of research on anonymization privacy-preserving techniques (In Chinese). *J. Chin. Comput. Syst.* 2011, 32(2):248–252.

[18] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discovery Data* 2007, 1(1):1–12.

[19] Li N, Li T, Venkatasubramanian S. Closeness: a new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.* 2009, 22(7):943–956.

[20] Liu X, Wang L. Advancement of anonymity technique for data publishing (In Chinese). *J. Jiangsu Univ. (Nat. Sci. Ed.)* 2016, 37(5):562–571.

[21] Nyffenegger A, Stürmer M, Niklaus J. Anonymity at risk? Assessing re-identification capabilities of large language models in court decisions. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico city, Mexico, June 16–21, 2024, pp. 2433–2462.

[22] Alexin Z. Does fair anonymization exist? *Int. Rev. Law Comput. Technol.* 2014, 28(1):21–44.

[23] Altman M, Cohen A, Nissim K, Wood A. What a hybrid legal-technical analysis teaches us about privacy regulation: the case of singling out. *B.U. J. Sci. Tech. Law* 2021, 27(1):1.

[24] Elliot M, O'Hara K, Raab C, O'Keefe CM, Mackey E, *et al.* Functional anonymisation: personal data and the data environment. *Comput. Law Secur. Rev.* 2018, 34(2):204–221.

[25] See Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, *et al.* Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, Online, August 11–13, 2021, pp. 2633–2650.

[26] Morris JX, Kuleshov V, Shmatikov V, Rush AM. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 6–10, 2023, pp. 12448–12460.

[27] Zeng S, Zhang J, He P, Liu Y, Xing Y, *et al.* The good and the bad: exploring privacy issues in retrieval-augmented generation (RAG). In *Proceedings of Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, August 11–16, 2024, pp. 4505–4524.

[28] Information Commissioner's Office. Generative AI fourth call for evidence: engineering individual rights into generative AI models. 2024. Available: https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-fourth-call-for-evidence/ (accessed on 27 November 2025).

[29] Open AI. How ChatGPT and our language models are developed: learn more about how we develop our models and apply them in products like ChatGPT. 2025. Available: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed (accessed on 27 April 2026).

[30] Quach K. What happens when your massive text-generating neural net starts spitting out people's phone numbers? If you're OpenAI, you create a filter. The Register. 2021. Available: https://www.theregister.com/2021/03/18/openai_gpt3_data/ (accessed on 27 April 2026).

[31] Ye X, Yan Y, Li J, Jiang B. Privacy and personal data risk governance for generative artificial intelligence: a Chinese perspective. *Telecommun. Policy* 2024, 48(10):102851.

[32] Buffone P. Health data privacy. 2017. Available: https://www.appliedclinicaltrialsonline.com/view/health-data-privacy (accessed on 12 March 2026).

[33] Solove DJ. *The digital person: technology and privacy in the information age.* New York: New York University Press, 2004.

[34] El Mestari SZ, Lenzini G, Demirci H. Preserving data privacy in machine learning systems. *Comput. Secur.* 2024, 137:103605.

[35] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, Oakland, USA, May 18–21, 2008, pp. 111–125.

[36] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, *et al.* On the opportunities and risks of foundation models. *arXiv* 2021, arXiv:2108.07258.

[37] Lermen S, Paleka D, Swanson J, Aerni M, Carlini N, *et al.* Large-scale online deanonymization with LLMs. *arXiv* 2026, arXiv:2602.16800.

[38] Fu X. Structural impact of generative artificial intelligence on the personal information protection law and its response (In Chinese). *Legal Forum* 2025, 40(6):102–112.

[39] Ayres I, Balkin JM. The law of AI is the law of risky agents without intentions. *U. Chi. Law Rev. Online* 2024, 2024:1–11.

[40] Zheng Z, Yuan M. A normative framework for personal information utilization in generative artificial intelligence (In Chinese). *J. Beijing Univ. Aeronaut. Astronaut. (Soc. Sci. Ed.)* 2026, 39(1):38–47.

[41]Information Commissioner's Office. How do we ensure anonymisation is effective? 2025. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/how-do-we-ensure-anonymisation-is-effective/ (accessed on 12 March 2026).

[42] European Data Protection Board. Guidelines 02/2026 on Anonymisation. 2026. Available: https://www.edpb.europa.eu/public-consultations/guidelines-022026-on-anonymisation_en (accessed on 11 July 2026).

[43] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques (WP216). 2014. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed on 27 April 2026).

[44] Zhang L, Yu L. The definition of tort liability in the value chain of generative artificial intelligence (In Chinese). *J. Beijing Admin. Inst.* 2025, (5):72–83.

[45] Liu Y, Huang J, Li Y, Wang D, Xiao B. Generative AI model privacy: a survey. *Artif. Intell. Rev.* 2024, 58(1):33.

[46] Altomani P. Italian Garante bans ChatGPT from processing personal data of Italian data subjects. 2023. Available: https://www.dataprotectionreport.com/2023/04/italian-garante-bans-chat-gpt-from-processing-personal-data-of-italian-data-subjects/ (accessed on 26 April 2026).

[47] Boudewijn A, Ferraris AF. Legal and regulatory perspectives on synthetic data as an anonymization strategy. *J. Pers. Data Prot. Law* 2024, 2024(1):17–31.

[48] Zhang Z. Value trade-offs and normative reconstruction in personal-information anonymization (In Chinese). *J. Nankai Univ. (Philos. Lit. Soc. Sci. Ed.)* 2025, 2025(6):66–81.

[49] Rubinstein IS, Hartzog W. Anonymization and risk. *Wash. Law Rev.* 2016, 91(2):703–760.

[50] Wang L. Re-examining infringement risks and countermeasures in generative artificial intelligence (In Chinese). *Soc. Sci. Guangdong* 2026, 2026(1):249.

[51] Zhao J. Digital jurisprudence: its definition and research orientation (In Chinese). *ECUPL J.* 2024, 27(4):64–75.

[52] Custers B, Lahmann H, Scott BI. From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in AI and other complex technologies. *AI Soc.* 2025, 40:4035–4050.

[53] Zhang X, Wei Y. Study on the basic issues of artificial intelligence legislation in China (In Chinese). *Law Soc. Dev.* 2024, 30(6):5–12.

[54] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Councilof 13 June 2024 laying down harmonised rules on artificial intelligence and amendingRegulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). 2024. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 11 July 2026).

[55] Stalla-Bourdillon S, Knight A. Anonymous data v. personal data: false debate: an EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l L. J.* 2016, 34(2):284–322.