

## Article

# Research on the Risk Prevention and Control of Data Poisoning in the Age of Artificial Intelligence

Bo Long<sup>1</sup>, Zilong Han<sup>2</sup>, Yukun Yue<sup>3</sup> and Yulin Tong<sup>4,\*</sup> 

<sup>1</sup> School of Law, Central South University of Forestry and Technology, Changsha 410004, China; t20252945@csuft.edu.cn

<sup>2</sup> Anhui Vocational College of Police Officers, Hefei 230031, China; fxyledge@gmail.com

<sup>3</sup> Hubei Provincial Research Center for Grassroots Governance, Yangtze University, Jingzhou 434023, China; z5616032@ad.unsw.edu.au

<sup>4</sup> The Faculty of Law and Justice, The University of New South Wales, Sydney, NSW 2052, Australia

\* Correspondence: yulin.tong@student.unsw.edu.au; Tel.: +86-159-8192-5998

## Abstract

The development of large artificial intelligence models relies on the integrity and accuracy of training data. Malicious contamination of training data introduces false information. This information becomes fixed as parameter bias during model training. Consequently, it affects judgment logic and model output. Therefore, data contamination has become a primary method for attacking the security of large models. Despite this threat, legal regulations concerning data poisoning lack systematic research. This paper employs a normative analysis method. It systematically reviews regulatory rules regarding data poisoning. These rules are found within Chinese criminal and data security laws. Current legislation primarily enforces general data security duties. It also relies heavily on accountability after an incident occurs. This approach makes early intervention difficult. Risks often evolve into model defects and cause damage before effective measures are taken. Therefore, the main contribution of this paper is moving beyond the current retrospective regulation model. We advocate establishing a risk control mechanism for the entire training process. This mechanism spans from prior review to continuous monitoring and subsequent accountability. Furthermore, it clearly divides the responsibilities among four main entities. These entities include data providers, model developers, service providers, and regulatory agencies. This mechanism helps resolve practical difficulties associated with data poisoning. These challenges include multiple participants, complex causal relationships, and unclear responsibilities. The proposed framework responds to the genuine needs of data security. It aligns with the governance goal of improving hierarchical and classified security supervision. Ultimately, it provides a feasible analytical framework for protecting training data.



Academic Editor: Patricia Easteal

Received: 29 June 2026

Revised: 23 July 2026

Accepted: 24 July 2026

Published: 28 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** data poisoning; data security; artificial intelligence; risk prevention and control

## 1. Introduction

As large AI models are widely deployed in critical sectors, including finance, health-care and government administration, security risks triggered by data poisoning have become increasingly prominent. Once a compromised model is put into practical use, erroneous information will spread continuously through model outputs. These vulnerabilities are no longer hypothetical. In March 2026, the annual consumer rights broadcast by China Central Television exposed a clandestine industry known as generative engine

optimization. Vendors seeded the open web with fabricated promotional articles. Within hours, a fictitious product became the standard answer recommended by mainstream artificial intelligence models (Xinhua News Agency 2026; OECD.AI 2026). Beyond China, the Kremlin-aligned Pravda network has published millions of falsified articles. These publications specifically target the training and retrieval data of artificial intelligence chatbots. One audit revealed that leading chatbots repeated these narratives in approximately one third of the tested answers (Sadeghi and Blachez 2025). Independent academic studies further demonstrate the fragility of these systems. Replacing a mere 0.001 percent of tokens in a medical training corpus measurably increases harmful outputs. Meanwhile, the compromised model continues to perform normally on standard benchmarks (Alber et al. 2025). Such consequences range from misleading public perceptions at a minor level to endangering national security in severe cases. Research on risk prevention and control against data poisoning therefore carries urgent practical significance.

Existing scholarship on data poisoning can be divided into two principal strands: technical research and legal research. Technical studies have developed a relatively mature account of the mechanisms through which poisoning attacks occur, as well as the methods by which they may be detected, thereby providing essential technical support for risk prevention and control (Steinhardt et al. 2017; Khraisat et al. 2025). In contrast, legal research remains comparatively underdeveloped. Relevant discussions are often subsumed within broader debates on artificial intelligence governance and cybercrime. In the field of AI governance, scholars have advocated the establishment of legally binding quality standards for training data (Chen et al. 2017). Others have proposed amending the Convention on Cybercrime to incorporate specific provisions on data poisoning or adopting a dedicated international treaty addressing AI-related crime. At the domestic level, legal scholarship has focused primarily on the criminal-law characterization of data-poisoning conduct, especially the interpretive application and regulatory viability of existing computer-crime offences (Li and Cao 2025; Khraisat et al. 2025). More recent studies have begun to examine civil liability and administrative supervision in this context. These studies address evidentiary difficulties, causation, risk allocation, and regulatory duties, while exploring possible pathways for strengthening the legal governance of data poisoning (Liu et al. 2025).

The existing literature nevertheless presents several significant limitations. First, although prior studies address the quality of training data, they tend to emphasize general data-quality control rather than malicious contamination through poisoning. These risks arise from distinct sources. Quality defects typically arise from negligence. Conversely, poisoning necessitates a motivated adversary. Section 2.1 elucidates the actors responsible for documented poisoning incidents. Second, existing research often subsumes data poisoning within conventional categories of cyberattacks, without sufficiently elucidating its distinctive attributes or the mechanisms through which its risks are transmitted (Shafahi et al. 2018). Its potential harms therefore remain inadequately examined (Chu and Liu 2025). Third, prevailing criminal-law studies generally follow a paradigm that extends existing offences to novel technological problems, rather than developing an independent evaluative framework tailored to data poisoning. This approach is insufficient. It fails to analyse the causal chains specific to AI model training with adequate precision and exposes clear deficiencies in the attribution of liability. Taken as a whole, the current body of research remains predominantly technology-oriented and pays insufficient attention to legal governance. It relies heavily on ex post regulatory logic, while neglecting front-end governance concerns such as risk identification, prior prevention, and continuous monitoring. This imbalance is especially problematic. Few studies have systematically examined data poisoning from the perspective of risk prevention and control (Yan et al. 2025). Nor has a full-chain regulatory framework been established to cover critical stages

such as contamination detection, traceability, remediation, and consequence management. Consequently, existing research cannot fully address the governance challenges posed by data poisoning.

Beginning with the concept and inherent characteristics of data poisoning, this paper analyses the liabilities and obligations of four categories of stakeholders from the perspective of risk prevention and control. They include data providers, model developers, service providers, and regulatory authorities. It argues that the prevention and control of data-poisoning risks should extend across the entire development process of large AI models. The proposed framework comprises three core components: prior review, continuous monitoring, and ex post accountability. By clarifying the responsibilities and obligations of all relevant actors, the framework safeguards each stage of AI model training and remedies deficiencies in existing regulatory rules. It also reorients governance from post-harm response to pre-risk intervention. In doing so, it enables effective intervention before model defects materialize and reduces the security risks generated by data poisoning (People's Daily 2025). It may also provide a robust analytical foundation for improving the security governance of AI training data.

## 2. The Concept and Features of Data Poisoning

### 2.1. The Concept of Data Poisoning

Data poisoning is an adversarial attack aimed at the training data of AI models, in which the attacker pollutes the training data in order to control or damage the algorithmic model (Sudhakar 2024). The cost of this attack is low. The attacker only needs to mix malicious data disguised as normal samples into the process by which the model takes in data, or to directly alter existing data, so that the model learns wrong patterns during training. This can have a deep effect on model performance, on decision accuracy, and even on system safety (Carlini 2021).

The AI Safety Governance Framework 1.0 issued in China lists data poisoning as an important kind of AI data security risk. It points out that data poisoning mainly takes the form of an attacker altering data, injecting wrong or misleading data, and polluting the probability distribution of the training data, which then weakens the accuracy and trustworthiness of the model (National Technical Committee 260 2024; Biggio and Roli 2018). By the way it is carried out, data poisoning can be roughly divided into active poisoning and passive poisoning. The former means that the attacker actively injects or alters samples in the training set to pollute the model. The latter means that the attacker spreads false information to pollute samples that may be drawn into training and, in this way, pollutes the model (Yang et al. 2024; Hartle et al. 2025). Data poisoning happens in the training stage, and by injecting malicious data, it changes the learning process of the model and affects the very base on which the model forms its judgement (Chen et al. 2021). Research shows that mixing only zero point zero one per cent of false text into the training set raises the harmful content in the model output by eleven point two per cent (Ministry of State Security of the People's Republic of China 2025). The main target of a data poisoning attack is the data used for model training. Its effect is not limited to the model itself (Bursztein 2018). Given the wide use of large AI models today, once a model outputs a polluted result, it may affect public understanding and may further affect national security.

Recent events elucidate the perpetrators and motivations behind data poisoning. Foreign actors affiliated with state governments execute poisoning campaigns to exert geopolitical influence. The Pravda network supports the Kremlin by flooding the open web with fabricated articles. These articles specifically target chatbot training and retrieval data. Audits reveal that leading chatbots repeat these narratives in up to one third of tested responses (NewsGuard 2025). Scholars continue to debate whether this phenomenon

demonstrates successful manipulation or highlights deficiencies in credible data (Alber et al. 2025). Business competitors and marketing vendors also manipulate data to secure commercial advantages. For example, generative engine optimization vendors in China planted fabricated product claims. Mainstream large models subsequently presented these claims as standard answers (Xinhua News Agency 2026; OECD.AI 2026). Furthermore, a court recently held the operators of an artificial intelligence tool liable for unfair competition (Hangzhou Intermediate People's Court 2025). This tool generated massive volumes of false consumer experience posts. Current or former employees constitute a third category of perpetrators. Their lawful access to data pipelines facilitates manipulation without requiring external intrusion. Publicly attributed cases remain scarce. This scarcity inherently reflects the covert nature of the risk (Metcalfe and Churilla 2026).

## 2.2. The Essential Features of Data Poisoning

Unlike traditional network attacks and ordinary data security problems, data poisoning has three essential features, which are the indirect nature of the attacking act, the covert and spreading nature of the actual harm, and the composite nature of the legal interests it harms (Ma 2025).

Aggressive data poisoning is an indirect adversarial technique. Attackers embed camouflaged samples into large language model conversational agents through legitimate data pipelines, disrupting the model's inherent reasoning mechanisms (Carlini 2021). This causes the system to deliver inaccurate responses during user consultations and professional diagnostic evaluations. At present, such attack vectors are currently predominantly deployed within commercial settings, with healthcare, professional service consultation and other sectors as secondary targets. In the commercial field, attackers such as commercial promoters or competitors inject falsified retrieval data into AI training corpora to mislead models into recommending fictitious or low-quality products (Xinhua News Agency 2026; OECD.AI 2026). This tactic can also be leveraged to generate fabricated user feedback data (Hangzhou Intermediate People's Court 2025). And in the healthcare sector, malicious actors target large medical language models by deliberately implanting misinformation into their training datasets via legitimate open internet channels, enabling low-cost indirect adversarial attacks (Alber et al. 2025). It is evident that unlike traditional cyberattacks that rely on external network intrusion, system resource occupation or direct data destruction, data poisoning sabotages the fundamental construction of AI models. It contaminates core data at the data collection and training stages, thereby fundamentally compromising the overall operational performance of models (An and Lu 2018).

The attacking act of data poisoning is indirect. The attacker injects disguised samples through lawful channels of data flow and changes the judgement logic of the model from inside the model. The attacker does not directly delete or modify system functions, and unlawful intrusion into the system is not a necessary condition. Traditional network attacks usually concern the availability of the external network, the use of system resources, or direct damage to data, while data poisoning changes the inner judgement logic of the model in a basic way (An and Lu 2018).

The harm of data poisoning is covert and shows up late, and it clearly tends to spread. The pollution effect hides within huge amounts of training data, and the harm and the act are separated in time and space. The link between the act and the harmful result is weak, which sets up an obstacle for the traditional idea of result-based liability. Poisoned samples are very close to normal samples in their statistical features, and ordinary detection methods can hardly screen them out one by one, so the cost of finding them rises sharply (Raghavan et al. 2022). Even if defenders put in large resources, they can only check suspicious data within a limited range. The remedy for the result is hard to reverse. Malicious data

becomes part of the model parameters and is hard to find by ordinary cleaning methods, and repair often means discarding and rebuilding the whole model. At the same time, a polluted model passes its defect to many downstream systems through transfer learning, and poisoning in one place may move the whole model chain (Pankajakshan et al. 2024). This spread may also set off a chain reaction. In the medical field, a diagnostic model affected by poisoning may be used as the base of other medical models, which causes the range of misdiagnosis to keep growing and may even endanger public health (Koh and Liang 2017).

The legal interests harmed by data poisoning are composite. The legal interest harmed by active data poisoning is the computer system that the large model stands for, or the data used to complete or to make up that computer system; that is, the computer system in a broad sense. If the view shifts to the poison data itself, or to passive data poisoning, the legal interest harmed by the false information that the poison data carries extends to the public right to know. When the matter touches a major disaster or another field where information carries high importance, the legal interest harmed by the poison data builds up layer by layer as the model passes it on. It reaches from the data security and the truth of information of a single system to public order and national security.

These features add up together. They make the existing rules built around traditional harm fall short when they face data poisoning, and they bring a challenge to the regulation of data poisoning under current laws.

### 3. An Analysis of the Regulation of Data Poisoning Under Current Laws and Regulations

At present, the rules that target data poisoning are scattered across texts of different legal levels, and a special and unified regulatory system has not yet taken shape. The relevant rules include the Criminal Law and the basic laws in the data field, and they also cover administrative regulations, departmental rules, various policy documents, and national standards, forming a regulatory network of crossing levels. Yet these rules were made at different historical stages and were meant to answer the risk concerns of particular periods. Their objects of regulation, their logic of regulation, and their means of regulation all differ, and there are weak links between them. More important, the current rules all target the review and regulation of data sources and output content, while the regulation of whether the data itself has problems is relatively lacking. It is therefore necessary to stand on the essential features of data poisoning, sort out the relevant rules in a systematic way, and examine their actual regulatory effect.

#### 3.1. *The Regulation of Data Poisoning Under Current Laws and Regulations*

The Criminal Law, the Data Security Law, and policy documents step in from different angles, which are after-the-fact punishment, data governance, and risk identification. Yet the focus of each kind of rule is not the same, and their coverage of the risk chain of data poisoning is not complete.

##### 3.1.1. The Regulation of Data Poisoning Under the Criminal Law

The Criminal Law has not set up a special clause that gives a unified evaluation of data poisoning. Instead, it regulates the conduct by interpreting and applying existing offences according to the specific attack setting.

For active poisoning, scholars mainly discuss the application of the crime of damaging a computer information system (Zhai and Li 2025). This crime punishes the act of interfering, in violation of state regulations, with the functions of a computer information system and thereby causing the system to fail to run normally (National People's Congress of the

People's Republic of China 2020, art. 286). Data poisoning interferes with the running of an AI system by injecting and altering data, so on its plain meaning there is room to bring it within this crime. The difficulty in application lies in whether the drop in model performance and the bias in decisions can be treated as the system failing to run normally (Ma 2025). If the attacker first intrudes into the target system or accesses the training data beyond authority, the act may also touch the crime of illegally intruding into a computer information system and the crime of illegally obtaining computer information system data (National People's Congress of the People's Republic of China 2020, art. 285). However, the focus of the aforementioned type of crimes is the taking of data, which runs in the opposite direction from poisoning, whose core act is injection and alteration. The intrusion-type crimes are usually evaluated on their own only when the poisoner uses real technical means to break through protection and intrude into the system (Cai 2026). These crimes mainly regulate the act of intruding into a system or obtaining data without authority, and their focus is not the alteration and pollution of data. For this reason, unless the actor carries out an intrusion, such as breaking through the protective measures of the system in order to poison, these crimes can hardly evaluate data poisoning itself. When the poisoning targets a particular firm or a particular production or business activity, the crime of damaging production and business operations and the crime of harming business reputation and product reputation have room to apply. The former regulates the act of damaging production or business by other methods, and the latter punishes the act of fabricating and spreading false facts that harm business reputation (National People's Congress of the People's Republic of China 2020, arts. 276 and 221). For passive poisoning, the false content produced by a polluted generative system may be spread through the network and disturb public order. It can then be regulated through the crime of fabricating or intentionally spreading false information (National People's Congress of the People's Republic of China 2020, art. 291-1). Providers and deployers may bear a safety management duty but fail to act or act poorly (Pi 2026). The crime of refusing to perform the information network security management duty can then hold them liable for their neglect (National People's Congress of the People's Republic of China 2020, art. 286-1).

Overall, although existing crimes can address the harms caused by data poisoning from different perspectives, their regulation remains fragmented and subject to significant limitations. No single crime can fully cover all the harm of data poisoning, and inconsistencies in legal evaluation and regulatory gaps are hard to avoid in the application of different crimes.

Improving the criminal regulation of data poisoning requires the establishment of criteria for protected legal interests, causation assessment, and liability attribution that are consistent with the characteristics of AI systems. With respect to protected legal interests, existing crimes address different types of harms, including computer system disruption, unauthorized access, dissemination of false information, and damage to business operations. However, the legal interests protected by these crimes and their focus of evaluation do not fully correspond to the distinctive risks associated with data poisoning. Regarding causation, the traditional approach in cybercrime, which relies primarily on direct and observable harmful consequences, should be appropriately adjusted. The assessment should focus on whether the poisoning conduct has materially altered the state of training data, whether such alteration has further affected model performance or outputs, and whether these effects have ultimately resulted in concrete harm. Regarding liability attribution, existing crimes are primarily based on the direct conduct of individual actors and clearly identifiable harmful consequences, making them insufficient to address the multi-actor and distributed-risk characteristics of the AI ecosystem. Data poisoning may involve multiple actors, including attackers, data providers, model developers, and AI service providers.

Attackers who directly conduct data contamination should bear corresponding liability where the elements of a crime are satisfied. By contrast, the liability of data providers, model developers, and AI service providers should be assessed based on their degree of control over data quality, model training, and system operation, as well as whether they have violated relevant security management obligations.

### 3.1.2. The Regulation of Data Poisoning Under the Data Security Law

The Data Security Law regulates data poisoning mainly through the security protection duty of data processors. The law requires data processors to set up and improve a whole-process data security management system, to take technical measures and other necessary measures to ensure data security, and to turn the upkeep of data security into a legal duty ([Standing Committee of the National People's Congress 2021](#), art. 27).

First, with regard to the role of firms in the flow of data, the law places the duty to keep data complete and usable on every stage of processing activity. It requires processors to take management responsibility for the whole life cycle of data collection, storage, use, processing, and transmission. Poisoning pollutes the training set and directly damages the completeness of the data, so it falls within the scope of the data security protection duty. Second, the Data Security Law requires processors to strengthen risk monitoring and to take remedial measures at once when data security defects and vulnerabilities are found, which extends the duty of risk control into the middle stage of processing activity ([Standing Committee of the National People's Congress 2021](#), art. 29).

The focus of the Data Security Law is on visible security risks such as data leakage, alteration, and loss, and it lacks a special mechanism to deal with hidden pollution that is injected through lawful channels and lies within huge numbers of samples. This gap is most clear when facing the hidden poisoning of the training stage. The precision of the regulation needs to be raised, and the basis for holding the poisoner liable is rather weak, which makes it hard to identify the responsible party and allocate responsibility with precision.

### 3.1.3. The Regulation of Data Poisoning Under Normative Documents

Apart from the laws and regulations above, normative documents mainly take providers and developers of AI services as their object and form a supplementary regulation of data poisoning.

The Interim Measures for the Management of Generative Artificial Intelligence Services require that the source of training data be lawful, and they make clear the safety management responsibility of large model providers and deployers ([Cyberspace Administration of China 2023](#), art. 7). This shows the position of China in encouraging technical innovation while carefully guarding against data security risks. The Provisions on the Management of Algorithmic Recommendation of Internet Information Services require algorithm providers to bear a filing duty and the necessary duties of disclosure and explanation. They also require providers to keep technical documents and use records that can be supervised. Yet they do not set a unified and workable standard for the degree to which explanation should reach ([Cyberspace Administration of China 2021](#); art. 7; [Cyberspace Administration of China 2023](#), art. 17). The Regulations on the Management of Network Data Security stress that data processors should carry out annual risk assessment and set up an internal control system that matches the scale and importance of their data processing. This gives a normative basis for the self-risk assessment of large model suppliers. At the level of policy documents, the White Paper on Artificial Intelligence Data Security treats data poisoning as a sub-concept of data pollution. It defines data poisoning as the act of adding disguised data and malicious samples into training data, damaging the com-

pleteness of the data and thereby causing bias in the decisions of the model, which gives textual support for the recognition of the concept (China Academy of Information and Communications Technology 2019). The AI Safety Governance Framework issued by the National Technical Committee on Cybersecurity Standardisation points out by way of listing that data poisoning mainly appears as training data being altered and as wrong or misleading data being injected. Its later version further puts forward a governance path of improving risk classification and exploring risk grading and dynamic adjustment, but the framework still stays at the level of exploratory principle statements.

The features of this kind of rule are that it is flexible and timely and can fit the real need of fast technical change. Yet its binding force is relatively limited, and it mainly depends on the self-management of firms and the supervision and enforcement of administrative authorities. When a rigid constraint is not enough, the compliance motive of market actors is easily weakened, and some actors may tend to take a strategy of minimum compliance, which then reduces the control function of these documents.

### 3.2. *The Legal Dilemma in Governing the Risk of Data Poisoning*

Current rules lack a prior review of whether the training data itself is true and reliable, lack ongoing monitoring of how poison data changes the model during training, and lack after-the-fact regulation of the harm and responsibility once the harm appears. The dilemma in preventing and controlling data poisoning thus shows up mainly in the three stages of before, during, and after.

#### 3.2.1. *The Missing Prior Review of the Data Itself*

Current rules lack a prior review of the authenticity and reliability of the data itself before it enters the training set. The focus of regulation falls on the lawfulness of the data source and the compliance of the generated content. There is no independent step to check whether the training data has been polluted by human action. The Interim Measures for the Management of Generative Artificial Intelligence Services require providers to improve the quality of training data, but this duty stays at the level of principle and lacks a workable review standard and sample verification process (Cyberspace Administration of China 2023, art. 7). The Regulations on the Management of Network Data Security require processors that provide generative AI services to strengthen the security management of training data. Yet the management points to network data security in a broad sense and does not reach the real identification of poison data (Network Data Security Regulations 2024, art. 19). Distorted training data is the source of false model output, and prior review should be carried out with training data at its centre (Xie and Wu 2026). The current arrangement leaves the check on data authenticity to firms to do by themselves. Without a mandatory standard and outside verification, the compliance input of firms is easily kept low, and finely disguised poison samples can pass self-review and enter the training set. Prior review is therefore absent. The authorities need to add the duty of prior review of training data at the level of rules, and make clear the subject, the content, and the method of review.

#### 3.2.2. *The Blank in Ongoing Monitoring of Data Training*

The training process lacks ongoing monitoring aimed at poisonous data, and the causal link between the poisoning act and the abnormal behaviour of the model is hard to establish in law. It is not transparent which data a model learns which behaviour from, and the training process that cannot be explained limits any account of how data affects output and blocks the tracing of the working mechanism (Y. Wang 2025). Current rules do not require developers to set up continuous monitoring of data access and data flow in the training stage, and they do not require keeping log records that allow tracing back. Once the records of the middle stage are missing, the link between the injection of poison

and the drop in model performance cannot be restored. The finding of a causal link lacks a clear legal standard, and the harmful result is hard to trace to a specific poisoning act. A legal judgement of cause needs a chain of facts that can be restored, and the lack of ongoing monitoring and log records makes that chain hard to rebuild. To prove that poisoning did happen and to trace the abnormality of the model to poisoning need full evidence, and the burden of proof falls on the harmed side, which raises the cost of seeking a remedy (Ding 2024). The blind spot of regulation falls exactly on the training stage, and the actual effect of poison data on the model lacks a visible record and a traceable mark. Ongoing monitoring is therefore missing.

### 3.2.3. The Insufficient After-the-Fact Regulation of Data Poisoning

After the harm appears, current law lacks a workable after-the-fact mechanism for damage assessment and accountability. First, the mechanism for damage assessment and remedy is missing. After it is confirmed that the model problem comes from data pollution, current rules have no matching standard for damage assessment. There is also no firm requirement for emergency isolation, data tracing, and model repair. These often depend on the conscience of the firm, the range and degree of the harm are hard to measure, and the remedy therefore lacks a basis. The newly enforced national standard sets technical requirements for identifying and filtering poison data in the training stage, but its focus is still on detection and cleaning, and it puts too little effort into assessment and accountability after the harm occurs (Standardization Administration of China 2024). Faced with harm that cannot be measured, the harmed side often cannot get a remedy that matches it. Second, the responsible party is hard to identify. Data poisoning involves several stages, namely data providers, model developers, and service providers, the lines of rights and duties are blurred and crossed, and it is often impossible to find out at which stage the pollution arose (Li and Xue 2025). Current law has not made targeted arrangements for the duty of care of the parties at each stage. After the harm occurs, it is therefore hard to lock onto the party that should bear responsibility and hard to weigh the degree of fault of each side. Once the chain of responsibility breaks, the allocation of fault behind a single poisoning act loses its support.

The present state of results-based regulation and difficult accountability can hardly answer the real risk of data poisoning. The step in of criminal law takes the appearance of a serious harmful result as its premise, and the risk built up by poisoning before and during training often falls outside criminal evaluation. Current rules put more effort on the two ends, data source and generated content; put less effort on the training stage in the middle; and lack a matching response for handling harm after it appears. As a result, the chain from poison data entering the training set to causing harm lacks early control and effective response. Administrative regulation leans towards formal filing and review and does not verify the training act itself enough, so it can hardly achieve continuous identification and feedback (J. Zhang 2023). The timing of regulation is late, its basis leans towards results, and accountability often comes to nothing. The three together make it hard for the law to step in before the risk turns into a model defect, and hard to give a matching remedy and punishment after the harm occurs. The failure of accountability weakens the deterrence of the rules, and the poisoner may use unclear responsibility to escape punishment. These gaps show that relying on after-the-fact punishment can hardly reach the risk that poisoning buries in the training stage. The focus of governance needs to move forward, and the legal mechanism should be rebuilt from the angle of risk prevention and control.

## 4. Building a Legal Mechanism for the Prevention and Control of Data Poisoning Risk

The risks posed by data poisoning have attracted increasing attention among governments and regulators worldwide. Although neither the European Union nor the United States has established a dedicated legal regime specifically addressing data poisoning, both have incorporated it into broader AI governance frameworks (Wang and Qi 2025). The EU Artificial Intelligence Act adopts a risk-based approach by classifying AI systems used in areas such as biometrics, education, and the administration of justice and democratic processes as high-risk and imposing quality requirements on their training datasets. The United States has not enacted a comprehensive federal AI statute but primarily relies on voluntary guidance issued by the National Institute of Standards and Technology (NIST). For example, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* classifies data poisoning as a type of adversarial machine learning attack and incorporates it into AI risk identification, assessment, and management (NIST AI 100-2e2025, NIST 2025, sct. 3.2). Although the EU and the United States have adopted different regulatory approaches, both emphasize risk management, offering useful insights for the legal regulation of data poisoning.

Against this background, the prevention and control of data poisoning risk needs to start from three sides, which are prior review, ongoing monitoring, and after-the-fact accountability, and these match four kinds of parties, namely data providers, model developers, model service providers, and regulators. The requirement of data authenticity assurance can be brought into the current AI governance framework through interpretation and amendment of the law (Zhao 2025). The three stages then link into a control chain that covers the whole training process.

### 4.1. The Prior Review Mechanism for Data Poisoning

The prior review mechanism requires data providers to bear the duty of checking the authenticity and reliability of the data that enters the training set. Although the current Interim Measures for the Management of Generative Artificial Intelligence Services require providers to improve the quality of training data at the level of principle, this duty lacks a workable check standard and sample verification process. For this reason the check on authenticity has long depended on the conscience of the firm (Cyberspace Administration of China 2023, art. 7). The direction of legal interpretation is to break the broad requirement of improving data quality down into review steps that can be carried out, so that the check duty of the provider has a clear boundary. Data providers should screen the source of samples at every stage of collection, cleaning, and labelling. They should use hash fingerprint comparison, spot checks of data, and quality assessment to make sure the data source is trustworthy, and start tracing at once when a risk is found (Cao et al. 2025). For malicious data disguised as normal samples and for abnormal data distribution, providers need to identify and remove them precisely before the data flows into training, in line with the features of poisoning, checking layer by layer from collection and feeding to training (Liu et al. 2025). Once polluted samples flow into the training set, they are hard to remove, and the cost of finding them rises sharply as training goes on, so setting the checkpoint before the data enters training can block the risk at a lower cost. Prior screening therefore takes on the function of protection at the source. Keeping a record of the review becomes the starting point of prior control.

The carrying out of the review duty depends on a clear assignment of responsibility. Under the Data Security Technology Rules for Data Classification and Grading, a data quality responsibility system should be truly set up. The duties of data producers, processors, and users in data quality management should be made clear, and a traceable arrangement

of responsibility should be formed (Standardization Administration of China 2025). On this basis, data providers become the main responsible party in prior review, bearing the duty to check the authenticity of the data they supply and giving a verifiable account of its source. The rights and duties involved in data quality assurance need to be set out at the institutional level so that the review standard has a unified basis (Su 2025). The setting of the review standard can refer to practice abroad. The EU Artificial Intelligence Act requires high-risk systems to carry out strict filtering at the data input stage and to guard against the injection of malicious data through a data source tracking mechanism. In this way data is checked for entry before it is put to use (European Commission 2024). The rules of China can take in the practice of prior filtering, set the check of the data source as a mandatory duty of providers, and change the present state in which the check on authenticity is left entirely to firms. Mandatory checking and administrative supervision work together to turn the review standard from an advice into a constraint. The introduction of a mandatory duty narrows the room for disguised samples to pass self-reviews and keeps polluted samples out of the training set (Peng and Zuo 2024). Prior review thus turns from broad safety management towards the real identification of poison data.

#### *4.2. The Ongoing Monitoring Mechanism for Data Poisoning*

Model developers control key stages of AI development, including training data processing, model construction, and safety testing. They should therefore assume ongoing monitoring obligations commensurate with their degree of control at different stages of AI development and operate within a corresponding accountability framework. Drawing on the regulatory approach of the Artificial Intelligence Act, developers should establish technical monitoring, log-keeping, pre-deployment security review, and incident response mechanisms throughout the training process. They should also designate key responsible persons at critical stages, including data collection, cleaning, labeling, and training, so as to ensure the timely detection, effective control, and traceability of data poisoning risks.

Ongoing monitoring should not be understood merely as a technical recommendation, but as a continuing security obligation imposed on relevant actors throughout the AI development process. Its purpose is not only to detect and mitigate data poisoning risks through technical means, but also to provide the factual basis necessary for causation analysis and liability attribution. Where relevant actors fail to fulfil reasonable monitoring obligations or fail to establish necessary monitoring, log-keeping, or incident response mechanisms, they may incur regulatory liability. Where such failures result in serious harm and satisfy the applicable legal requirements, they may also give rise to civil or criminal liability.

The ongoing monitoring mechanism requires model developers to set up technical monitoring during the training process. The lack of transparency in the training process makes the link between the injection of poison and the abnormality of the model hard to restore, and the blind spot of regulation falls on the training stage. Developers should carry out continuous monitoring of training parameters, training samples, and the running state of the model, stop training at once when abnormal behaviour is found, and look for suspicious signs while tracing and reducing loss (Xu and Zhang 2025). Continuous monitoring covers the swings in training loss, the shift in gradient distribution, and the abnormality in model output, which gives a chance to notice the injection of poison hidden among huge numbers of samples. The density of monitoring needs to match the intensity of training. Foundation model providers bear a system security protection duty, take technical responsibility for the resistance of training data to pollution, and build the safety requirement into the inner stage of model construction (Xie and Wu 2026). Developers stand at the core technical position of the AI system, and their duty to guard against

poisoning bears directly on the safety and steadiness of the model, so the strength of the duty should match the technical control power of the developer (Carlini and Wagner 2017).

The monitoring of data access and data flow needs a complete keeping of logs. Developers need to set up continuous records of data access and data flow in the training stage, save logs that allow tracing back, and keep a factual basis for attribution between the injection of poison and the abnormal output. The Artificial Intelligence Act requires high-risk systems to set up technical documents, event logs, and use records, and it stresses that they be traceable, auditable, and accountable (Zhang and Ji 2026). The governance framework of China can take in the requirement of keeping records, setting data sheets, logs, and lineage records as the legal materials of developers and thus keep a trace of evidence (Sun 2024). The EU takes model providers as the core duty bearers, sets duties of record-keeping and pre-market compliance assessment for high-risk systems, and lets the training process undergo outside review (Wang and Qi 2025). Legislation abroad builds a dynamic monitoring mechanism into the development stage, tracks the flow of data throughout the process, and keeps assessing data security, offering a sample that the middle stage can learn from.

Ongoing monitoring also includes the handling after an abnormality is found and the review before the model goes online. For a model that has finished training, developers should carry out safety testing with model reverse engineering and neurone activation analysis to identify hidden trigger patterns (Tong et al. 2026). Trigger patterns hide within the model parameters, ordinary testing can hardly expose the marks of pollution, and safety testing needs to look deep inside the model and observe abnormal activation paths. The content of a model update needs to be on the record, and the model can go online only after smart algorithms and human review (Susskind 2022). A version management system and an online review mechanism can prevent an unreviewed model from going online at will and polluting other models, and keep an abnormal model isolated from the chain of spread (R. Zhang 2025). Online review acts as a gate before the model enters application, a model that does not pass review is kept out of the chain of spread, and the pollution cannot spread to other systems through model reuse. Developers should set a key responsible person at every stage of data collection, cleaning, labeling, and training, let the key responsible person check stage by stage, and hold the risk of poisoning within the training process (Kumar et al. 2024). Setting a key responsible person brings the broad responsibility of the firm down to a specific post, so once poisoning appears it can be traced to the gatekeeper of the matching stage. Ongoing monitoring thus fixes the factual marks needed to restore the cause.

It should be clarified that the designation of a gatekeeper does not mean that all legal consequences arising from data poisoning should be attributed to the individual employee occupying that position. By identifying key responsible persons, the ongoing monitoring mechanism facilitates the timely detection of risks, the preservation of evidence, and the traceability of harmful events, thereby providing an important basis for causation analysis and the allocation of liability among multiple actors. However, given the multi-actor nature of AI development, data poisoning may result from a combination of organizational management, technical design, data governance, and security management failures. It would therefore be inappropriate to regard individual employees as the sole bearers of legal responsibility. AI enterprises, as the principal organizers of AI development and deployment, should bear corresponding governance responsibilities, while regulatory authorities should perform supervisory functions through compliance oversight and risk assessment. The liability of individual employees should be determined in light of their specific duties, degree of control, and any breach of applicable security obligations. The allocation of liability for data poisoning should take into account each actor's contribution

to the materialization of the risk, degree of control, and breach of legal obligations, so that responsibility remains proportionate to the actor's role.

#### 4.3. *The After-the-Fact Accountability Mechanism for Data Poisoning*

The after-the-fact accountability mechanism targets the emergency handling and the accountability after the harm appears, and it falls on two parties, the model service provider and the regulator. First, when the model outputs an abnormal result, the service provider should take emergency isolation, data tracing, and model repair in time, so as to avoid the harm or stop it from growing. The current Interim Measures require providers, when they find unlawful generated content, to take measures such as stopping generation, stopping transmission, and removal, and to report to the competent authority ([Cyberspace Administration of China 2023](#), art. 14). The service provider faces users and the output end directly and stands at the place that first notices an abnormal result, and the speed of emergency handling directly decides how far the harm spreads. The service provider should also set up internal data security management rules, screen suspicious data, test abnormal samples, report poisoning events in time, and fix the emergency process as a daily system ([C. Wang 2025](#)). Data tracing checks layer by layer along the flow of data to locate the polluted samples and the node of injection, while model repair rolls the model back to a version not yet polluted and trains it again. Fixing the emergency process gives firms rules to follow when an event happens, so that handling no longer depends on a temporary judgement. A quick isolation mechanism stops malicious data from spreading further along the model chain and buys time for later repair.

Second, regulators need to add industry rules and set up a mechanism for damage assessment and accountability. The assessment standard and the basis for accountability after the harm occurs are still missing, so the remedy lacks support. Experience abroad takes damage assessment and reporting as a legal step after an incident, and firms need to assess how badly the incident has damaged the completeness of data and the performance of the system, and report to the competent authority ([European Union 2016](#)). Damage assessment needs to measure the range of samples reached by the pollution, the number of downstream models affected, and the actual impact on public understanding, so that the remedy gains a measurable basis. Regulators can set damage assessment as a mandatory requirement through departmental rules and make clear the subject, the content, and the time limit of the assessment. Operators of critical information infrastructure should carry out safety testing and risk assessment on a regular basis, and this arrangement of regular assessment can extend to the harm checking of high-risk models ([Cyberspace Administration of China 2021](#), art. 17). Data poisoning involves several stages, namely data providers, model developers, and service providers, and a clear definition of the duty of care of the party at each stage helps to lock onto the party that should bear responsibility after the harm occurs ([Zheng 2024](#)). Industry rules and departmental rules need to be improved to fill the gap in the current remedy, and to turn assessment and accountability from the conscience of the firm into a legal procedure.

The carrying out of accountability needs reliable support for tracing and a varied arrangement for sharing the burden. Technologies such as blockchain offer novel technical solutions for accountability. For example, blockchain records data usage through a distributed ledger. This ledger cannot be altered. This immutability enhances the technical reliability of data tracking. It also provides evidence to detect the extent of pollution ([Zheng et al. 2018](#)). However, researchers must note the limitations of these new solutions. Blockchain technology cannot verify external data independently. It must rely on oracles to transmit information. This transmission connects facts from the physical world with smart contracts. Therefore, an attacked oracle will automatically execute distorted inputs

on a massive scale. This causes erroneous liquidation, data manipulation, and inequality among participating entities (Eskandari et al. 2021; Caldarelli and Ellul 2021). Consequently, practical applications should configure verification across multiple nodes. Systems should also include hardware heterogeneity and fault substitution mechanisms (Zhang et al. 2016). Methods like random task allocation can reduce the probability of manipulated oracle results (Alyukov et al. 2025). Cryptographic data commitments, symbolic analysis, and satisfiability modulo theories provide further support. Solvers use these tools to analyze smart contract execution results. This analysis occurs when oracle inputs deviate from expected values. The system then identifies safety parameters and automatically generates protection conditions (Zhang et al. 2020; Deng et al. 2024). Integrating multiple technical solutions avoids significant technical dilemmas. A single centralized data processor struggles to cover the entire process. This integration also clarifies rights and obligations. These responsibilities often become ambiguous during interactions among multiple parties. The burden should therefore be divided among the parties in a reasonable way as the risk passes on (Li and Xue 2025). Cybersecurity insurance can take part in data quality assurance, and after the duty bearer takes out insurance, the third-party services offered by the insurer and partner institutions ensure data quality, so that the risk is spread across society (Zhou 2024). The taking out of insurance spreads the harm that a single firm can hardly bear over a wider range and gives the harmed side a steady source of remedy. Regulators can also lead the building of a data quality sharing alliance, set up a blacklist of poison samples, and share the results of identification within the industry (Chen et al. 2024). The blacklist gathers the scattered results of identification into a shared resource for protection, so that newly appearing poison samples can be blocked within the industry in time. After-the-fact accountability thus extends from plain punishment to a complete arrangement of damage assessment, burden sharing, and industry cooperation, and together with prior review and ongoing monitoring, it forms a control mechanism that covers the whole training process.

## 5. Conclusions

Data poisoning leverages data circulation and large AI model training to covertly solidify contamination into model defects. Its hazards go beyond the scope of technical security and may impair national decision security, public information order and social public interests. Though China's current regulatory system lays down rules for full process security management and risk monitoring of data processing activities, the overall framework still focuses on damage prevention and control after harm occurs. It lacks targeted risk prevention and control mechanisms for the stage before training data is imported into models. Therefore, regulators must move beyond the traditional regulatory logic centred on post-harm mitigation, shift regulatory focus to an earlier stage and establish a full process governance framework covering prior review, ongoing monitoring and after-the-fact accountability. Article 27 of the Data Security Law of the People's Republic of China requires all entities conducting data processing activities to formulate and improve full-process data security management systems and implement necessary measures to safeguard data security (Bi and Zhang 2025). Article 29 obligates enhanced risk monitoring. Entities must take immediate remedial actions once risks including data security flaws and vulnerabilities are identified.

At the prior review stage, the authenticity and reliability of training data shall be included within the scope of statutory data security protection obligations. Data providers shall bear clear verification duties so that access screening of training data does not fully rely on enterprises' self-assessment and internal management. At the ongoing monitoring stage, model developers shall undertake continuous monitoring and other corresponding duties based on their degree of control over training operations and risk management

capacity. Regulators shall specify statutory retention periods for relevant records and official review standards to ensure all large AI model training activities remain under supervisable conditions (Zhou and Zhao 2026). At the after-the-fact accountability stage authorities shall refine institutional arrangements for risk reporting, damage assessment and liability pursuit. Liabilities shall be reasonably allocated according to each stakeholder's control capacity and degree of fault. Prior review, ongoing monitoring and after-the-fact accountability form an interconnected system. This framework expands governance of data poisoning beyond isolated case handling after damages emerge to cover the complete timeline of risk generation and proliferation. Under this framework the authenticity and reliability of training data cease to be mere technical benchmarks for model development. They evolve into legally binding security obligations that run through data provision model training and service delivery.

In conclusion, establishing a dedicated risk prevention and control mechanism for data poisoning fills gaps in regulatory provisions governing training data protection. It also enables legal intervention against artificial intelligence risks to take effect at an earlier stage with updated regulatory approaches. Large AI models have become core technical infrastructure supporting information production, social governance and national decision-making. Authentic and reliable training data serves as the fundamental prerequisite for trustworthy model operation. Legislation shall respond timely to the formation logic of technology-related risks, embed data security obligations into every link of model training and deployment and maintain a reasonable balance between technological innovation and security safeguards.

**Author Contributions:** Conceptualization, B.L. and Z.H.; methodology, B.L., Z.H. and Y.Y.; validation, B.L., Z.H., Y.Y. and Y.T.; formal analysis, B.L. and Z.H.; investigation, B.L., Z.H. and Y.Y.; resources, B.L. and Y.Y.; writing—original draft preparation, B.L., Z.H., Y.Y. and Y.T.; writing—review and editing, B.L., Z.H., Y.Y. and Y.T.; visualization, B.L. and Z.H.; supervision, B.L., Z.H. and Y.T.; project administration, B.L. and Y.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This article is sponsored by The Research Foundation of the Ministry of Education of the People's Republic of China for Humanities and Social Sciences Projects (Project Number: 21YJCZH095).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

- Alber, Daniel Alexander, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A. Valliani, Jeff Zhang, Gabriel R. Rosenbaum, Ashley K. Amend-Thomas, David B. Kurland, and et al. 2025. Medical Large Language Models Are Vulnerable to Data-Poisoning Attacks. *Nature Medicine* 31: 618–26. [\[CrossRef\]](#) [\[PubMed\]](#)
- Alyukov, Maxim, Mykola Makhortykh, Alexandr Voronovici, and Maryna Sydorova. 2025. LLMs Grooming or Data Voids? LLM-Powered Chatbot References to Kremlin Disinformation Reflect Information Gaps, Not Manipulation. *Harvard Kennedy School Misinformation Review*, October 15. [\[CrossRef\]](#)
- An, Keying, and Hong Lu. 2018. The Criminal Law Protection of Cybersecurity in the Age of Artificial Intelligence: From the Perspective of the AI-Enabling of Cybercrime. *Journal of Yunnan Minzu University (Philosophy and Social Sciences Edition)* 35: 145–56.
- Bi, Kun, and Hualin Zhang. 2025. Algorithms, Data, and Computing Power: The Technical Support for the Risk Governance Paradigm of Artificial Intelligence. *Jiangsu Social Sciences* 3: 110.
- Biggio, Battista, and Fabio Roli. 2018. Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning. *Pattern Recognition* 84: 317–31. [\[CrossRef\]](#)

- Bursztein, Elie. 2018. Attacks Against Machine Learning: An Overview. Available online: <https://elie.net/blog/ai/attacks-against-machine-learning-an-overview> (accessed on 16 July 2026).
- Cai, Shilin. 2026. On the Restrictive Application of the Crime of Illegally Obtaining Computer Information System Data. *Huxiang Forum*, 101–11. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7202987168> (accessed on 16 July 2026).
- Caldarelli, Giulio, and Joshua Ellul. 2021. The Blockchain Oracle Problem in Decentralized Finance—A Multivocal Approach. *Applied Sciences* 11: 7572. [CrossRef]
- Cao, Yifei, Qirui Luo, and Lin Zhang. 2025. Data Mixed Poisoning Attacks Based on Federated Learning. *Computer Engineering*, 1–2, Published online.
- Carlini, Nicholas. 2021. Poisoning the Unlabeled Dataset of Semi-Supervised Learning. Available online: <https://arxiv.org/abs/2105.01622> (accessed on 16 July 2026).
- Carlini, Nicholas, and David Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. New York: ACM, pp. 1–12.
- Chen, Jian, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. 2021. De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks. *IEEE Transactions on Information Forensics and Security* 16: 3412–25. [CrossRef]
- Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems. *Cryptography and Security*, 1–18. Available online: <https://arxiv.org/abs/1712.05526> (accessed on 16 July 2026). [CrossRef]
- Chen, Xuebin, Zhiqiang Ren, and Hongyang Zhang. 2024. A Survey of Security Threats and Defense Measures in Federated Learning. *Journal of Computer Applications* 44: 1663–64.
- China Academy of Information and Communications Technology. 2019. White Paper on Artificial Intelligence Data Security. Available online: <http://www.caict.ac.cn/> (accessed on 16 July 2026).
- Chu, Jiewang, and Youyun Liu. 2025. Research on the Formation Mechanism and Prevention Strategies of Information Pollution in the AIGC Environment. *Information Studies: Theory & Application*. Published online.
- Cyberspace Administration of China. 2021. Provisions on the Management of Algorithmic Recommendation of Internet Information Services. Available online: [https://www.cac.gov.cn/2022-01/04/c\\_1642894606364259.htm](https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm) (accessed on 16 July 2026).
- Cyberspace Administration of China. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. Available online: [https://www.cac.gov.cn/2023-07/13/c\\_1690898327029107.htm](https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm) (accessed on 16 July 2026).
- Deng, Xun, Sidi Mohamed Beillahi, Cyrus Minwalla, Han Du, Andreas Veneris, and Fan Long. 2024. Safeguarding DeFi Smart Contracts against Oracle Deviations. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. New York: ACM, pp. 171:1–171:12. [CrossRef]
- Ding, Daoqin. 2024. The Data Legal Issues in the Training Stage of Generative Artificial Intelligence and Legislative Suggestions. *Administrative Law Review*, 16–28. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7113496793> (accessed on 16 July 2026).
- Eskandari, Shayan, Mehdi Salehi, Wanyun Catherine Gu, and Jeremy Clark. 2021. SoK: Oracles from the Ground Truth to Market Manipulation. In *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies*. New York: ACM, pp. 127–41. [CrossRef]
- European Commission. 2024. Artificial Intelligence Act. Available online: <https://artificialintelligenceact.eu/> (accessed on 16 July 2026).
- European Union. 2016. General Data Protection Regulation (EU) 2016/679, Arts. 33–34. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (accessed on 16 July 2026).
- Hangzhou Intermediate People's Court. 2025. Civil Judgment (2025) Zhe 01 Min Zhong No. 3998. Available online: <https://wenshu.court.gov.cn/> (accessed on 16 July 2026).
- Hartle, Frank, III, Steve Mancini, and Emily Kerry. 2025. Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges. *Issues in Information Systems* 25: 433–34.
- Khraisat, Ansam, Ammar Alazab, Moutaz Alazab, Tony Jan, Sarabjot Singh, and Ashraf Uddin. 2025. Securing Federated Learning: A Defense Strategy against Targeted Data Poisoning Attack. *Discover Internet of Things* 5: 16. [CrossRef]
- Koh, Pang Wei, and Percy Liang. 2017. Understanding Black-Box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*. Sydney: PMLR, pp. 1885–94.
- Kumar, Naveen K., Krishna C. Mohan, Linga Reddy Cenkeramaddi, and Navchetan Awasthi. 2024. Minimal Data Poisoning Attack in Federated Learning for Medical Image Classification: An Attacker Perspective. *Artificial Intelligence in Medicine* 159: 103024. [CrossRef] [PubMed]
- Li, Huaisheng, and Zhi Cao. 2025. The Criminal Law Response to AI-Related Data Poisoning. *Digital Rule of Law*, 114–15. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7202423504> (accessed on 16 July 2026).
- Li, Yang, and Lan Xue. 2025. Subversion, Adaptation, and Synergy: A Study of the Multi-Party Governance Mechanism of Generative AI from the Perspective of Responsibility Ethics. *E-Government*, 40–49. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7201539266> (accessed on 16 July 2026). [CrossRef]

- Liu, Yuxin, Weiqiang Jiang, Zhongyuan Guo, Xiaohu Lu, and Shuanglin Tang. 2025. Research on Solutions to the Security Risk of AI Data Poisoning. *Secrecy Science and Technology*, 12–13. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7200646640> (accessed on 16 July 2026).
- Ma, Runyi. 2025. The Risks and Regulatory Paths of Data Poisoning. *Science and Technology Management Research*, 156–69. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7201097000> (accessed on 16 July 2026).
- Metcalf, Renae, and Matt Churilla. 2026. Data Poisoning in AI Models: The Case for Chain of Custody Controls. *Carnegie Mellon University, Software Engineering Institute Blog*, April 27. Available online: <https://www.sei.cmu.edu/blog/data-poisoning-in-ai-models-the-case-for-chain-of-custody-controls/> (accessed on 16 July 2026).
- Ministry of State Security of the People's Republic of China. 2025. Beware of AI Data Poisoning. Available online: <http://stcn.com/article/detail/2926924.html> (accessed on 16 July 2026).
- National People's Congress of the People's Republic of China. 2020. Criminal Law of the People's Republic of China. Available online: [http://en.npc.gov.cn.cdurl.cn/2020-12/26/c\\_921604.htm](http://en.npc.gov.cn.cdurl.cn/2020-12/26/c_921604.htm) (accessed on 16 July 2026).
- NIST. 2025. NIST Trustworthy and Responsible AI NIST AI 100-2e2025. Available online: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf> (accessed on 16 July 2026).
- OECD.AI. 2026. AI Data Poisoning via GEO Manipulates Recommendations and Misleads Consumers in China. *OECD AI Incidents Monitor*, March 14. Available online: <https://oecd.ai/en/incidents/2026-03-14-e431> (accessed on 16 July 2026).
- Pankajakshan, Rahul, Sumitra Biswal, Yuvaraj Govindarajulu, and Gilad Gressel. 2024. Mapping LLM Security Landscapes: Comprehensive Stakeholder Risk Assessment Proposal. *Cryptography and Security*, 1–12. [CrossRef]
- Peng, Zhongli, and Zedong Zuo. 2024. The Legal Governance of Data Errors. *Journal of Sichuan University (Philosophy and Social Science Edition)*, 157–69. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7113456209> (accessed on 16 July 2026).
- People's Daily. 2025. Xi Jinping Stressed at the 23rd Collective Study Session of the Political Bureau of the CPC Central Committee: Improve the Long-Term Mechanism for Governing the Online Ecosystem and Continuously Build a Clean and Healthy Cyberspace. *People's Daily*, November 30, p. 1.
- Pi, Yong. 2026. On the AI Safety Management Duty and the Criminal Liability of Those Who Bear It. *China Legal Science*, 146–47. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7203404833> (accessed on 16 July 2026).
- Raghavan, Vijay, Thomas Mazzuchi, and Shahram Sarkani. 2022. An Improved Real Time Detection of Data Poisoning Attacks in Deep Learning Vision Systems. *Discover Artificial Intelligence* 2: 18. [CrossRef]
- Sadeghi, McKenzie, and Isis Blachez. 2025. A Well-Funded Moscow-Based Global “News” Network Has Infected Western Artificial Intelligence Tools Worldwide with Russian Propaganda. *NewsGuard*, March 6. Available online: <https://www.newsguardtech.com/special-reports/moscow-based-global-news-network-infected-western-artificial-intelligence-russian-propaganda/> (accessed on 16 July 2026).
- Shafahi, Ali, W. Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. Available online: <https://dl.acm.org/doi/10.5555/3327345.3327509> (accessed on 16 July 2026).
- Standardization Administration of China. 2024. *Data Security Technology—Rules for Data Classification and Grading*. GB/T 43697-2024. Beijing: Standardization Administration of China. Available online: [https://openstd.samr.gov.cn/bzgk/std/showGb?type=download&hcno=F0C385EDC38CBF277AEC021F23126ADE&request\\_locale=zh](https://openstd.samr.gov.cn/bzgk/std/showGb?type=download&hcno=F0C385EDC38CBF277AEC021F23126ADE&request_locale=zh) (accessed on 16 July 2026).
- Standardization Administration of China. 2025. *Cybersecurity Technology—Security Specification for Generative Artificial Intelligence Pre-Training and Optimization Training Data*. GB/T 45652-2025. Beijing: Standardization Administration of China. Available online: [https://openstd.samr.gov.cn/bzgk/std/showGb?type=download&hcno=82710B59110419C285BDC48AB4D7D1F3&request\\_locale=zh](https://openstd.samr.gov.cn/bzgk/std/showGb?type=download&hcno=82710B59110419C285BDC48AB4D7D1F3&request_locale=zh) (accessed on 16 April 2026).
- Standing Committee of the National People's Congress. 2021. Data Security Law of the People's Republic of China. Available online: <http://www.npc.gov.cn/> (accessed on 16 July 2026).
- Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. 2017. Certified Defenses for Data Poisoning Attacks. Available online: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf) (accessed on 16 July 2026).
- Su, Yu. 2025. The Institutional Guarantee of Public Data Quality. *Administrative Law Review*, 85–100. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7201460613> (accessed on 16 July 2026).
- Sudhakar, Prithiv Roshan. 2024. Subverting AI's Intellect: How to Thwart Data Poisoning. *ISACA Journal* 1: 45–48.
- Sun, Qingbai. 2024. On the Normative Construction of Risk Governance for AI Large Model Training Data. *E-Government*, 41–52. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7200028122> (accessed on 16 July 2026).
- Susskind, Jamie. 2022. *The Power of Algorithms: How Humans Live Together*. Translated by Dabai Li. Beijing: Beijing Daily Press, p. 34.
- Tong, Songsong, Kuiwu Yang, Wen Wang, Jianghong Wei, and Haofeng He. 2026. A Feature-Blocking Backdoor Defense Mechanism for Computer Vision Models. *Journal of Image and Graphics* 31: 755–68. [CrossRef]
- Wang, Chen. 2025. Data Poisoning: The Reconstruction of the Subjectivity of Original Authors in the Age of AI Image Generation. *Youth Journalist*, 35–36. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7200571762> (accessed on 16 July 2026).

- Wang, Yijin, and Zhanyong Qi. 2025. Taming the Digital Leviathan: The Legal Regulation of AI Embedded in Education Data Governance—Taking the EU Artificial Intelligence Act as an Example. *University Education Science*, 88–98. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7200504705> (accessed on 16 July 2026). (In Chinese)
- Wang, Yixiang. 2025. The Technical Dilemma of AI Explainability and Its Legal Resolution. *Advances in Social Sciences*, 206–13. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=HS723972025011025> (accessed on 16 July 2026).
- Xie, Zhiyong, and Mengyu Wu. 2026. The System Security Protection Duty of Providers of AI Foundation Models. *Journal of Comparative Law*, 1–17. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7203129011> (accessed on 16 July 2026). [CrossRef]
- Xinhua News Agency. 2026. Cutting Off the Black-and-Gray Industry Chain That Poisons AI. March 23. Available online: <https://www.news.cn/tech/20260323/c91f52ec891a44bab8d7348b98989c2d/c.html> (accessed on 18 July 2026).
- Xu, Wei, and Li Zhang. 2025. Whole-Chain Governance of AI False Information in China: Operating Mechanism, Practical Dilemma, and Optimization Path—A Grounded Analysis Based on 62 Policy Texts. *E-Government*, 1–13, Published online.
- Yan, Yukun, Peng Tang, Rui Chen, Ruochen Du, and Qilong Han. 2025. A Randomness-Enhanced Bilevel Optimization Defense Method against Data Poisoning Backdoor Attacks. *Netinfo Security*, 1074. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7201513482> (accessed on 16 July 2026).
- Yang, Ziyang, Jie Zhang, Wei Wang, and Huan Li. 2024. Invisible Threats in the Data: A Study on Data Poisoning Attacks in Deep Generative Models. *Applied Sciences* 14: 8742. [CrossRef]
- Zhai, Yan, and Xiaobo Li. 2025. The Connotation and Features, Hierarchical Risks, and Governance Paths of Data Poisoning from the Perspective of AI Safety. *Henan Social Sciences*, 113–24. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7202205632> (accessed on 16 July 2026).
- Zhang, Fan, Ethan Cecchetti, Kyle Croman, Ari Juels, and Elaine Shi. 2016. Town Crier: An Authenticated Data Feed for Smart Contracts. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. New York: ACM, pp. 270–82. [CrossRef]
- Zhang, Fan, Sai Krishna Deepak Maram, Harjasleen Malvai, Steven Goldfeder, and Ari Juels. 2020. DECO: Liberating Web Data Using Decentralized Oracles for TLS. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. New York: ACM, pp. 1919–38. [CrossRef]
- Zhang, Jiyu. 2023. On the Algorithm Filing System. *Oriental Law*, 86–98. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7109332634> (accessed on 16 July 2026).
- Zhang, Qihang, and Leilei Ji. 2026. Information Disclosure: A Governance Path for the Use of AI Model Training Data. *Library*, 40–48. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7203167642> (accessed on 16 July 2026).
- Zhang, Rong. 2025. Design of a Data Poisoning Attack Detection and Robust Cleaning Algorithm for AI Model Training. *Information Research* 51: 79–82.
- Zhao, Jingwu. 2025. On the Institutional Construction of the High-Quality Supply of AI Training Data. *China Law Review*, 92–104. Available online: <https://qikan.cqvip.com/Qikan/Article/Detail?id=7200411194> (accessed on 16 July 2026).
- Zheng, Zhifeng. 2024. The Identification of Subjects and the Design of Liability Attribution for AI Application Responsibility. *Law Review* 42: 123–37.
- Zheng, Zibin, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. 2018. Blockchain Challenges and Opportunities: A Survey. *International Journal of Web and Grid Services* 14: 352–75. [CrossRef]
- Zhou, Ruijue. 2024. The Intervention Path of Cybersecurity Insurance in the Governance of Data Breach Risk. *Northern Legal Science* 18: 76–90.
- Zhou, Ruijue, and Jingwu Zhao. 2026. Legislation or Interpretation: Correcting the Governance of AI Data Poisoning. *Exploration and Free Views*. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.