

The Legal Framework for Source-Based Governance of AIGC Disinformation: A Comparative Perspective Between China and the EU

Il quadro giuridico del governo fondato sull'origine della AIGC Disinformation: una prospettiva comparata tra Cina e Unione europea

JINGKUN LIU 

Full Professor of Law China University of Political Science and Law
Visiting Expert UNIDROIT

YUANJUN LI

Ph.D. University of Rome "Tor Vergata"

Abstract

The rapid iteration of generative AI has engendered complex forms of disinformation of AI-generated content (AIGC Disinformation), which constitutes a systemic risk spanning technical, legal, and social dimensions. Comparative research of Chinese and EU legal frameworks indicates that effective source-based governance of AIGC Disinformation requires further development and harmonization. In terms of training data governance, China emphasizes the legality of data sources, quality assurance, and security management to ensure the authenticity of outputs. The EU focuses on transparency and traceability mechanisms, alongside data quality management systems, to mitigate AIGC Disinformation risks. Regarding AI model risk management, China has established lifecycle-based security assessment mechanisms aligned with national recommended standards, applied on a voluntary basis. The EU employs a risk-based regulatory framework, imposing stringent obligations particularly on general-purpose AI (GPAI) models with systemic risks and models used in high-risk AI systems. Despite differences, both jurisdictions treat training data governance as the first line of defense and AI model risk management as the second line of defense. Both also emphasize clarifying the primary responsibilities and liabilities of AI providers. Looking forward, China and the EU should promote global consensus on AIGC Disinformation governance through collaborative frameworks and mutual technical recognition, fostering the high-quality development of generative AI.

Abstract

La rapida iterazione dell'intelligenza artificiale generativa ha generato forme complesse di disinformazione basata su contenuti generati dall'IA (AIGC Disinformazione), che costituisce un rischio sistemico esteso alle dimensioni tecnica, giuridica e sociale. La ricerca comparata sui quadri normativi della Cina e dell'Unione europea indica che il governo efficace, fondato sull'origine, della AIGC Disinformazione richiede ulteriore sviluppo e armonizzazione. Per quanto riguarda il governo dei dati di addestramento, la Cina enfatizza la liceità delle fonti dei dati, la garanzia della qualità e la gestione della sicurezza al fine di assicurare l'autenticità degli output. L'Unione europea si concentra su meccanismi di trasparenza e tracciabilità, unitamente a sistemi di gestione della qualità dei dati, per mitigare i rischi connessi alla AIGC Disinformazione. Con riferimento alla gestione dei rischi dei modelli di IA, la Cina ha istituito meccanismi di valutazione della sicurezza basati sull'intero ciclo di vita del modello, allineati a standard nazionali raccomandati e applicati su base volontaria. L'Unione europea adotta un quadro normativo fondato sul rischio, imponendo obblighi stringenti in particolare ai modelli di IA di uso generale (GPAI) che presentano rischi sistemici e ai modelli impiegati in sistemi di IA classificati come ad alto rischio. Nonostante le differenze, entrambe le giurisdizioni considerano il governo dei dati di addestramento come la prima linea di difesa e la gestione dei rischi dei modelli di IA come la seconda linea di difesa. Entrambe sottolineano altresì la necessità di chiarire le responsabilità e gli obblighi primari dei fornitori di IA. In prospettiva, la Cina e l'Unione europea dovrebbero promuovere un consenso globale sul governo della AIGC Disinformazione attraverso quadri collaborativi e meccanismi di mutuo riconoscimento tecnico, favorendo uno sviluppo di alta qualità dell'intelligenza artificiale generativa.

Keywords: AIGC; disinformation; source-based governance; training data governance; AI model's risk management; AI provider's liability

Summary: [1. The Necessity of Source-Based Governance for AIGC Disinformation.](#) – [1.1 Root Causes of AIGC Disinformation.](#) – [1.2 The Necessity of Source-Based Governance for AIGC Disinformation.](#) – [2. Dimension I of Source-Based Governance: Governance of Training Data.](#) – [2.1 China: The Control of Data Compliance.](#) – [2.2 The EU: The Requirements of Transparency and Explainability.](#) – [2.3 Comparison and Optimization.](#) – [3. Dimension II of Source-Based Governance: Risk Management of AI Models.](#) – [3.1 China: The Combination of Principled Legislative Norms and Lifecycle-oriented Model Risk Assessment in National Recommended Standards.](#) – [3.2 The EU: A Risk-based Classified Framework in Unified Legislation.](#) – [3.3 Comparison and Optimization.](#) – [4. Dimension III of Source-Based Governance: Liability of AI Providers.](#) – [4.1 China: A Common Liability for Generative AI Service Providers.](#) – [4.2 The EU: Allocation of Liability Across the AI Value Chain for Providers.](#) – [4.3 Comparison and Optimization.](#) – [5. Conclusions.](#)

1. The Necessity of Source-Based Governance for AIGC Disinformation.

AIGC Disinformation¹ is not merely a digital extension of traditional disinformation but arises from the interaction of data, models, and algorithms. Legal responses should therefore move beyond the verification of falsehood and harm, and instead trace risks to their technological sources, thereby constructing a full-chain governance system encompassing the generation, dissemination, and reception of AIGC Disinformation. Within this broader governance framework, this paper focuses on source-based governance as the starting point of full-chain regulation.

1.1 Root Causes of AIGC Disinformation.

Effective source-based governance requires an examination of the technical origins of AIGC Disinformation, which primarily involve three dimensions:

First, data bias functions as a latent risk factor in the generation of disinformation. Data constitutes the core foundation of large models. On the one hand, factual inaccuracies and value-based prejudices embedded in training datasets may be internalized by the model, thereby influencing the quality and reliability of AIGC.² This dynamic is often summarized as “bias in, bias out”.³ Because models generate outputs probabilistically rather than storing data, embedded biases cannot simply be removed by deleting files from a storage device. On the other hand, temporal lags and insufficient diversity in training data may lead to errors when models address real-time or highly complex issues. Although connecting models to the internet may supplement or update available information, such access does not amount to model iteration, nor does it fundamentally resolve structural limitations in data access.

¹ The AIGC Disinformation, as defined herein, refers to the information that is inconsistent with objective facts and is generated through generative AI technologies. Specifically, AIGC Disinformation consists of false or misleading content that is automatically fabricated or falsified by generative AI in accordance with a user's instructions, or that is automatically produced as a result of AI hallucinations, and presented in single or composite digital forms such as text, images, audio and video. Depending on whether the actor possesses subjective deliberate intention, AIGC Disinformation may be divided into two categories: one category comprises disinformation deliberately generated through the use of AI by the actor, while the other comprises disinformation generated through the use of AI in the absence of the actor's subjective deliberate intention. The former may be further subdivided into disinformation maliciously generated through the use of AI by the actor, and the content intentionally generated for deepfake through the use of AI that may produce harmful effects of disinformation. The latter may be further subdivided into disinformation generated by AI in accordance with the actor's erroneous instructions without subjective malicious intention (for example, due to limited knowledge or comprehension, or failure to verify out of negligence), and disinformation generated by AI as a result of hallucinations where the actor inputs normal instructions.

² With respect to training data containing value prejudice, the value prejudice themselves may not necessarily constitute disinformation. However, due to the input of training data that including value prejudice, the model may output disinformation that reflects value prejudice.

³ See E Kasneci and others, 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education' (2023) 103 Learning and Individual Differences <<https://doi.org/10.1016/j.lindif.2023.102274>>; S G. Mayson, 'Bias in, Bias out' (2019) 128(8) Yale Law Journal 2218 ff.

Second, model hallucinations reflect an inherent limitation of probabilistic text generation. Models produce outputs based on token probability distributions learned during training, which may result in factual fabrications inconsistent with reality.⁴ Particularly in contexts characterized by information scarcity or high uncertainty, models may generate content that appears coherent and plausible but lacks factual grounding. This feature renders AIGC Disinformation especially deceptive and may undermine trust in AI systems.

Third, algorithmic opacity creates governance blind spots. Owing to technical complexity and confidential constraints, an algorithmic black box is formed,⁵ so that model decision-making processes are often difficult to explain or interpret. Such opacity complicates the prediction, tracing, and legal attribution of liability for AIGC Disinformation arising from data bias or technical defects.

1.2 The Necessity of Source-Based Governance for AIGC Disinformation.

The emergence of AIGC Disinformation is not merely the product of accidental errors but a systemic risk embedded within the technological architecture. This complex causation, shaped by the interconnections of data, models, and algorithms, underscores the necessity of source-based governance. As the initial phase in addressing AIGC Disinformation, source-based governance directs regulatory attention to the design, development, and testing stages of AI systems. Its necessity may be understood along three principal dimensions.

First, the generation of AIGC Disinformation exhibits a high degree of technological coupling. This structural characteristic distinguishes it from traditional disinformation and confers upon source-based governance a level of technical relevance absent in conventional regulatory approaches.

Second, as the cornerstone of a full-chain governance framework, source-based governance establishes the foundation for subsequent regulatory phases. On the one hand, by imposing *ex ante* technical requirements, it reduces the probability of disinformation generation and produces a risk-mitigation conduction effect that enhances the efficiency of downstream governance measures. On the other hand, it serves as the starting point for allocating responsibilities among relevant actors, thereby facilitating subsequent liability attribution and remedies.

⁴ With regard to the term of “hallucination”, no unified definition has been established yet, but it may be divided into two categories: one is factuality hallucination, in which AIGC is inconsistent with objective facts, including inconsistency with facts and complete fabrication; the other is faithfulness hallucination, in which AIGC is inconsistent with user input, including inconsistency with user instructions, inconsistency with the contextual information provided by the user, and internal logical inconsistency within AIGC. The AIGC disinformation discussed in this paper relates only to the factuality hallucination. For a discussion of this classification and its rationality, see L Huang and others, ‘A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions’ (2025) 43(2) ACM Transactions on Information Systems 1-55 <<https://doi.org/10.1145/3703155>>.

⁵ See V Hassija and others, ‘Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence’ (2024) 16 Cognitive Computation 45-74.

Third, source-based governance exerts a constructive influence on the development of AI technologies. Rather than functioning solely as a constraint on technological risk, it serves as a regulatory mechanism that promotes innovation through structured oversight. By strengthening source-based requirements such as accuracy, robustness, and explainability, it enhances the credibility of AI systems and advances the objective of Trustworthy AI.⁶ Moreover, the early consideration of liability encourages developers to treat authenticity as a core evaluative criterion alongside performance optimization. In this manner, compliance pressures are transformed into incentives for responsible innovation, guiding AI development toward a balanced emphasis on efficiency, safety and quality, and fostering technology aligned with the public good.

To operationalize source-based governance effectively, the framework should be structured around three core dimensions corresponding to the principal nodes of the technical generation chain: governance of training data, risk management of AI models, and allocation of liability to AI providers, thereby reducing the likelihood of generating AIGC Disinformation.

2. Dimension I of Source-Based Governance: Governance of Training Data.

Training data, as defined herein, refers to the collection of sample data used for the pre-training and post-training of AI models or systems, excluding test data. By providing large quantities of samples containing features (information describing sample attributes - inputs) and optional labels (the corresponding results - outputs), training data enable AI models or systems to learn patterns and mapping relationships, ultimately equipping them to process new and unknown data or to perform specific tasks. As the foundation of AIGC, the quantity, quality, and diversity of training data determine the overall performance of generative models and their systems. Accordingly, risks of source contamination and selective input within training data constitute significant causes of AIGC Disinformation. Governance of training data is therefore a prerequisite for safeguarding the authenticity of AIGC.

2.1 China: The Control of Data Compliance.

China has established a regulatory framework for training data governance primarily through administrative regulations, such as the Regulations on Network Data Security Management;⁷ departmental regulations, such as the Interim Measures for the Administration of Generative Artificial Intelligence Services (hereafter “Interim Measures”);⁸ and national standards, such as Cybersecurity Technology - Security Specification for Generative Artificial

⁶ See United Nations, A/RES/78/265, 21 March 2024.

⁷ Regulations on Network Data Security Management (网络安全管理条例) <https://www.gov.cn/zhengce/zhengceku/202409/content_6977767.htm>.

⁸ Interim Measures for the Administration of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) <https://www.gov.cn/zhengce/202311/content_6917778.htm>.

Intelligence Pre-training and Fine-tuning Data (GB/T 45652-2025).⁹ In general, these norms impose three core requirements on training data.

First, the framework requires the legality of data sources. Article 7(1) of the Interim Measures stipulates that generative AI service providers shall use lawfully sourced data for training activities.¹⁰ The core objective of this requirement is to prevent the injection of false information into training datasets by establishing mechanisms of authorization and traceability.

On the one hand, lawful data sources direct training data toward a trusted information basis. Collection methods such as user authorization, copyright verification, and compliant procurement can reduce the risk of incorporating unreliable or tampered data. Selecting data from authoritative institutions, such as government agencies and their public service platforms, industry-certified entities, or formal academic databases, can facilitate the establishment of trustworthy datasets and lower the probability of AI generating false or misleading content. On the other hand, lawful data sources support traceability. By recording the origins of collected data, it becomes possible to trace the sources of AIGC Disinformation and identify liable parties where false information exists within the underlying data, thereby increasing the cost of disseminating disinformation.

Second, the framework emphasizes data quality assurance, encompassing both data content quality and data annotation quality.

With respect to data content quality, Article 7(4) of the Interim Measures requires improvement of training data quality and enhancement of authenticity, accuracy, and diversity.¹¹ National standards such as GB/T 45652-2025 and Cybersecurity Technology - Basic Security Requirements for Generative Artificial Intelligence Service (GB/T 45654-2025)¹² provide more detailed operational guidance. These standards classify “false and harmful information” as a subset of “illegal and unhealthy information” and further group the latter, together with “inaccurate and unreliable content” in specific

⁹ Cybersecurity Technology - Security Specification for Generative Artificial Intelligence Pre-training and Fine-tuning Data (GB/T 45652-2025) (网络安全技术 生成式人工智能预训练和优化训练 数据 安全 规范) <<https://openstd.samr.gov.cn/bzqk/qb/newGbInfo?hcno=82710B59110419C285BDC48AB4D7D1F3>>.

¹⁰ Article 19 of the *Interim Measures* further provides that, when the authorities conduct supervision and inspection of generative AI services in accordance with their duties, generative AI services providers shall lawfully explain the sources of training data as required. Article 5.1(d) of GB/T 45652-2025 also requires generative AI services providers to record the sources of pre-training data collected from external data sources, but does not require such information to be disclosed to the public.

¹¹ Paragraph 2 of Article 6 of the Interim Measures further provides that high-quality public training data resources should be expanded, and that the use of secure and trustworthy data resources should be encouraged.

¹² Cybersecurity Technology - Basic Security Requirements for Generative Artificial Intelligence Service (GB/T 45654-2025) (网络安全技术 生成式人工智能服务安全基本要求) <<https://openstd.samr.gov.cn/bzqk/qb/newGbInfo?hcno=F67D3F376E0A0A0FF5317FB36B32A30A>>.

generative AI applications,¹³ under the broader concept of “security risk information”. The proportions of the latter two types of information serve as the key criteria for, respectively, data source selection¹⁴ and content filtering¹⁵.

In particular, before training, filtering mechanisms, such as keyword libraries and classification models, should be established for data of different modalities and languages, to remove illegal and unhealthy information, including false and harmful information. During the post-training stage, datasets should be examined and corrected for erroneous knowledge, incorrect expressions, misleading implications, fabrication, or tampering. Mechanisms, including fine-tuning and value alignment, are required to ensure the professional integrity of data and consistency with human values and ethical norms. Where AIGC is used as training data, a hallucination risk assessment mechanism should be established to identify and correct erroneous content. Notably, the standards introduce dedicated authenticity assessment requirements for both pre-training and post-training data, thereby directly targeting data reliability. With respect to diversity, the standards prohibit reliance on single-source or poorly diversified datasets.¹⁶ By enhancing

¹³ According to Appendix A.5 of GB/T 45654-2025, such information refers to content generated through the use of generative AI in specific service types with relatively high safety requirements, such as critical information infrastructure, automatic control, medical information services, psychological counseling, financial information services and so on, that exhibits: (a) inaccurate content that seriously deviates from scientific common sense or mainstream understanding; or (b) unreliable content that does not contain serious errors but nevertheless fails to provide meaningful assistance to users.

¹⁴ For data source selection, during the stages of data collection, pre-training and post-training, the proportion of “illegal and unhealthy information” contained in data from different sources serves as a criterion for determining whether a data source may be selected. Typically, manual sampling inspections are conducted, whereby no less than 10% of records are randomly sampled from data source security assessment records and verification records. Data sources containing more than 5% illegal and unhealthy information must not be collected, and already collected data containing more than 5% illegal and unhealthy information must not be used for training. For data that has undergone the above pre-processing, additional sampling inspections must be conducted to examine whether samples contain illegal and unhealthy information, with a minimum sample size of 1,000. Where self-collected training data are used, at least 1,000 training data items must be randomly sampled from the entire self-collected dataset, all of which must include collection records such as timestamps and source information; from these sampled training data items, a further random sample of 100 items must be drawn, none of which may contain data that any third party has explicitly indicated must not be collected.

¹⁵ For data content filtering, before training, the proportion of security risk information contained in different types of data (such as text, images, audio, and video) serves as the filtering standard. Through manual sampling inspection, no fewer than 4,000 data items must be randomly sampled from the full set of training data for each modality, with a qualified rate of no less than 96%. Through technical sampling inspection using keyword matching, classification models and similar methods, no fewer than 10% of the total data volume for each modality must be randomly sampled, with a qualified rate of no less than 98%. Data that fails to meet the qualification threshold must not be used. During the post-training stage, no fewer than 10% or 1,000 data items must be randomly sampled from the filtered dataset to examine whether they contain security risk information.

¹⁶ Different data sources must include multiple data providers, including but not limited to internet websites, other organizations or individuals, and service users; within the same type of data, the proportion contributed by each data source must not be less than 1%, where “type” includes but is not limited to code, images, audio, video and text in the same language; data from overseas must be reasonably combined with domestic data.

representativeness and breadth, these measures strengthen model reasoning capacity and increase the likelihood of factually accurate outputs when addressing novel issues.

Regarding data annotation quality, Article 8 of the Interim Measures requires data annotation quality assessments during research and development, with sampling inspections to verify annotation accuracy. GB/T 45654-2025 and Cybersecurity Technology - Generative Artificial Intelligence Data Annotation Security Specification (GB/T 45674-2025)¹⁷ further establish rules concerning annotation verification standards and annotator management.¹⁸ By improving annotation accuracy, these measures enhance models' capability to recognize and refuse false and harmful content.

Third, the framework requires data security management throughout data processing activities. Article 19 of the Regulations on Network Data Security Management mandates strengthened security management of training data and related processing activities to prevent data security risks. Similarly, Article 14(1) of Provisions on the Administration of Deep Synthesis in Connection with Internet Information Services (hereafter "Deep Synthesis Provisions")¹⁹ requires deep synthesis service providers and technical supporters to adopt necessary measures to ensure training data security.²⁰

GB/T 45652-2025 specifies technical measures relating to storage and transmission security, operational log auditing, and security testing.²¹ Through

¹⁷ Cybersecurity Technology - Generative Artificial Intelligence Data Annotation Security Specification (GB/T 45674-2025) (网络安全技术 生成式人工智能数据标注安全规范) <<https://openstd.samr.gov.cn/bzqk/std/newGblInfo?hcno=407584DD0FA2BA19E62E85D3469290B0>>.

¹⁸ Specifically: (1) With respect to annotation verification standards, data annotation is divided into two categories, namely functional data annotation (used to train generative AI models to acquire the capability to perform specific tasks) and safety data annotation (used to train generative AI models to enhance the safety performance of AI's outputting). Different verification requirements are imposed on the quality of each type of annotation. For functional annotation, only manual sampling inspection of each batch of annotated data is required; where inaccurate content is identified, the entire batch must be re-annotated, and where illegal and unhealthy information such as false and harmful content is identified, the entire batch of annotated data must be invalidated. For safety annotation, because it covers all categories of security risk information, including false and harmful information, each annotated data item must be examined, and must be reviewed and approved by at least one verification personnel. (2) With respect to annotator management, a functional separation between execution and verification roles is implemented: annotation personnel and verification personnel assigned to the same task must not be the same individual, thereby establishing a dual-layer safeguard against improper or erroneous annotations.

¹⁹ Provisions on the Administration of Deep Synthesis in Connection with Internet Information Services (互联网信息服务深度合成管理规定) <https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm>.

²⁰ Although the *Interim Measures* do not explicitly stipulate "data security", the requirements that the development of generative AI adhere to the principle of security in Article 3, that generative AI service providers offer secure services in Article 13, and that generative AI service providers with public opinion influence or social mobilization capabilities should conduct security assessments and algorithm registration in Article 17, may be understood as indirectly emphasizing data security.

²¹ Specifically, the national standard provides that data processing activities shall, at a minimum, meet the requirements of Level 3 protection under the data classification system, and that technical measures such as identity authentication, access control, encrypted data storage and

lifecycle security controls over data collection, pre-processing, and use, risks such as deletion of data traceability records, malicious data injection or tampering, and misuse of sensitive data for introducing and/or generating false information can be mitigated. By combining security safeguards with quality controls, the framework seeks to enhance the reliability and authenticity of AIGC.

Despite these normative requirements, several practical challenges remain.

First, inconsistencies persist in the definition of “training data”. The Interim Measures do not define the term, while Article 23 of the Deep Synthesis Provisions defines it as “annotated or benchmark datasets used to train machine learning models”, thereby limiting its scope to machine learning. Although GB/T 45654-2025 defines training data broadly as all data directly used as model training inputs, including pre-training and post-training data, GB/T 45652-2025 narrows the concept to data “used solely for generative AI pre-training and post-training”. These inconsistencies create uncertainty and exclude training data used for non-generative AI models from the national standards’ sphere of application.

Second, from a governance-value perspective, although the national standards provide operational details, they are recommended rather than mandatory. Binding rules still lack substantive criteria for assessing authenticity, accuracy, and diversity. Principled provisions are difficult to apply directly for identifying and correcting false information in training data, thereby limiting regulatory effectiveness.

Third, from a technological development perspective, compliance is often secondary to performance optimization for AI developers. Compliant datasets may lack sufficient content breadth, and models trained exclusively on such datasets may perform less effectively. Moreover, ensuring full compliance prior to model development, such as through database procurement, may impose high costs and reduce innovation incentives. Consequently, the practical observance of training data requirements remains uncertain.

Finally, existing norms insufficiently address purely factual false information generated by AI. Current regulations focus primarily on “illegal and unhealthy information”, “false and harmful information”, and “fake news”, but do not comprehensively cover the factually false information of outputs in areas such as health or academic knowledge.

In light of these limitations, the European Union’s regulatory framework offers instructive comparative insights.

transmission (using secure channels such as TLS/SSL), security isolation, backups, and operational log auditing shall be adopted to protect data security. Where industry data are involved, corresponding protective measures shall also be implemented in accordance with relevant industry standards. With respect to data security testing, the standard requires the identification and remediation or filtering of poisoned data within training datasets (such as data containing large-scale annotation errors or data exhibiting specific characteristics), and further requires the establishment of emergency response mechanisms for data security incidents.

2.2 The EU: The Requirement of Transparency and Explainability.

The EU has developed a relatively structured legal framework for training data governance.²²

First, EU law provides explicit conceptual definitions. Article 3 of the Artificial Intelligence Act (hereafter “AI Act”) distinguishes between “training data”, “validation data”, and “testing data”. Although, for the purpose of this paper, training data corresponds broadly to the combined notion of training and validation data under EU law, the AI Act clarifies these concepts at the level of applicable objects and stages. Compared with Chinese law, the AI Act offers a more systematic categorization across the AI lifecycle.

Second, the AI Act establishes transparency requirements for general-purpose AI (GPAI) models and their training data.²³ Generative AI models

²² This paper focuses on the Artificial Intelligence Act with respect to training data governance. See Regulation (EU) 2024/1689 <<http://data.europa.eu/eli/reg/2024/1689/oj>>. In fact, however, the EU’s data governance framework is also closely connected with legislation such as General Data Protection Regulation (GDPR), see Regulation (EU) 2016/679 <<http://data.europa.eu/eli/reg/2016/679/oj>>; the Data Act, see Regulation (EU) 2023/2854 <<http://data.europa.eu/eli/reg/2023/2854/oj>>; and the Data Governance Act, see Regulation (EU) 2022/868 <<http://data.europa.eu/eli/reg/2022/868/oj>>; and so on.

First, Article 2(7) of the AI Act explicitly provides that Union laws concerning the protection of personal data remains applicable. Where personal data are used as training data for AI, compliance with GDPR is required (e.g. Article 6(1)(a), in GDPR). At the same time, GDPR grants individuals a series of rights for the protection of personal data. Accordingly, risks of AIGC Disinformation arising from data misuse and the associated infringements of rights can be effectively restrained through GDPR. However, the two legislative instruments contain overlaps, inconsistencies, and enforcement difficulties with respect to certain obligations governing data-related activities.

Second, the Data Act, which aims to ensure users’ access rights to data generated by connected AI products and services and to promote data sharing, interacts with the AI Act. The Data Act primarily applies to non-personal data and mixed datasets. Where an AI provider is both a data holder under the Data Act and simultaneously subject to provider obligations under the AI Act, overlaps or conflicts may arise.

Third, the Data Governance Act, by establishing frameworks for trusted data intermediaries and data altruism, promotes voluntary data sharing. This can support the data governance requirements of the AI Act and create potential synergies in governance. Data intermediary services under the Data Governance Act may serve as a mechanism to provide assurances regarding training data quality and compliance for AI providers, particularly providers of high-risk AI systems, thereby alleviating compliance burdens. The optimization of the relationship between the two regimes depends on enhanced cooperation among EU-level governance bodies (such as the AI Office and the European Data Innovation Board), including the joint issuance of guidance to ensure that data sharing in the AI domain complies both with the safety and quality standards of the AI Act and with the principles of trust and neutrality enshrined in the Data Governance Act.

Fourth, it should be noted that the Digital Omnibus proposal released by the European Commission in November 2025 aims to simplify and consolidate various EU digital legislation. One of the most consequential proposed amendments concerns the definition of personal data in Article 4(1) of GDPR. See <<https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal>>. As the concept of personal data not only determines the scope of application of GDPR but also has practical implications for the aforementioned legal instruments within the EU digital regulatory framework, any modification to this definition would exert significant downstream effects.

²³ For the transparency requirements relating to training data for GPAI models, see Article 53 of the AI Act and Template for the Public Summary of Training Content for General-Purpose AI

generally fall within the scope of GPAI regulation. In particular, providers of GPAI models with systemic risk should disclose information concerning data types and sources, acquisition and selection practices, measures for detecting inappropriate sources, and methods for identifying biases and other additional information.²⁴ Substantial fines may be imposed for non-compliance. The independent regulations for GPAI models distinguish the EU approach from those of China.

Third, the AI Act adopts a risk-based approach, imposing differentiated requirements according to risk level. High-risk AI systems must comply with specific data governance obligations.²⁵ Article 10 requires that training, validation, and testing datasets for AI models in the high-risk systems meet quality standards, thereby indirectly regulating the generation of AIGC Disinformation.²⁶ These provisions resemble Chinese rules but are more detailed and incorporate human rights impact considerations. Article 15 further promotes accuracy, robustness, and cybersecurity,²⁷ reinforcing safeguards against both the injection of false information into training data and the re-entry of AIGC Disinformation into iterative training cycles. Unlike China's reliance on recommended technical standards, the EU law requires quantifiable performance metrics and systematic evaluation of accuracy and robustness. Moreover, Article 17(1)(f) mandates quality management systems and comprehensive data management records, ensuring traceability of training data.²⁸ Such traceability facilitates rapid identification of problematic data sources and provides a clear basis for accountability where AIGC Disinformation causes harm.

2.3 Comparison and Optimization.

China and the EU adopt distinct regulatory paradigms for training data governance, reflecting differing legislative philosophies.

China follows a model of "legislative abstraction combined with standard specification". Binding legislation articulates principles, while recommended

models. See <<https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models>>.

²⁴ However, the above information is not required to be disclosed to the general public. Moreover, to assist AI model providers in fulfilling the above obligations, the EU General-Purpose AI Code of Practice provides a Model Documentation Form that may serve as a disclosure template for training data information. The use of this form is not mandatory. See <<https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>>.

²⁵ When the application of a GPAI model falls within the category of high-risk AI systems, the compliance requirements for high-risk AI systems should be applied at the same time. In the extremely limited circumstances, for example, where generative AI performs only a single task and lacks general-purpose characteristics, it does not constitute a GPAI model and is regulated as a general AI system.

²⁶ See Article 10(2)(c) and 10(3) of the AI Act. It is worth mentioning that, with respect to testing data, EU law applies governance requirements identical to those applicable to training data; by contrast, Chinese law adopts a different approach, namely by establishing separate specialized test item repositories for model output risk assessment and optimization, rather than treating testing data as part of data governance.

²⁷ See Article 15 of the AI Act.

²⁸ See Article 17(1)(f) of the AI Act.

national standards provide detailed technical parameters. This approach preserves flexibility during rapid technological development while guiding enterprises toward structured compliance.

The EU, by contrast, adopts a “framework-first and unified legislation” model. The AI Act establishes systematic obligations and procedural requirements, while technical standards are delegated to standardization bodies. This model emphasizes risk prevention, human rights protection, and regulatory coherence across Member States, thereby reinforcing the EU’s “Brussels Effect” in global digital governance.²⁹

From an institutional effectiveness perspective, both models present strengths and limitations. China’s principle-based system may face uneven enforcement, whereas the EU’s detailed requirements increase compliance burdens and costs. Nevertheless, both approaches should converge on the common objective of ensuring data reliability, quality controllability, and security traceability.

Accordingly, training data governance may be optimized along three dimensions.

First, alternative sourcing models should be explored. Instead of collecting large volumes of data and filtering ex post, regulators may stimulate ex ante collecting high-quality or professional closed-source databases. Incentives could include tax reductions for compliant data providers, financial subsidies or insurance incentives for AI providers purchasing certified databases, or the establishment of funds for the public-commercial mixed and shared data. Governments may also organize high-quality public datasets and make them available under defined conditions.

Second, training data assessment mechanisms should evolve toward a dual-track model combining objective static assessment and dynamic synchronous assessment. Static assessment addresses legality, authenticity, diversity, and representativeness before training. Dynamic assessment monitors real-time outputs and their associated links to training data during and after training. Given the nonlinear and probabilistic nature of generative AI, only a closed-loop system integrating both approaches can effectively prevent the generation and dissemination of AIGC Disinformation.

Third, technical tracing and filing mechanisms should be strengthened. Although both Chinese and EU law require logging, distributed technologies such as blockchain should be innovated to establish full-lifecycle records of training data and even the self-evolution of the metadata. Immutable audit trails would enhance traceability and accountability at the source, reinforcing the integrity of training data governance.

3. Dimension II of Source-Based Governance: Risk Management of AI Models.

Risk management of AI models, as defined herein, refers to a comprehensive set of measures for identifying, assessing, monitoring, and addressing the risks

²⁹ See A Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press 2020) 1 ff. <<https://doi.org/10.1093/oso/9780190088583.001.0001>>.

of false content generated by models. These measures are implemented throughout the entire lifecycle of the model, including training, outputting, monitoring, updating, and operation within its environment, in order to prevent, mitigate, and remediate the risks of AIGC Disinformation arising from data deficiencies, technical flaws, improper operations, or external attacks. The ultimate objective is to enhance the authenticity and reliability of AIGC.

Model risk management builds upon training data governance and functions as a second line of defense against AIGC Disinformation. It extends the scope of prevention and control from the “data input layer” to the “model output layer”. It not only verifies the effectiveness of training data governance but also constitutes a critical mechanism for containing the generation and dissemination of AIGC Disinformation.

3.1 China: The Combination of Principled Legislative Norms and Lifecycle-oriented Model Risk Assessment in National Recommended Standards.

Within China’s AI-related legal framework, a distinctive normative model for AI model risk management has begun to take shape, although it remains at an early stage of systematization.

First, direct legislative constraints on AI models are relatively limited and fragmented and are primarily reflected in three categories of principle-based norms. The first category consists of “red-line prohibitions”, namely models that must not be developed or deployed, including those that contravene mainstream values, violate ethical norms, or contain illegal and unhealthy information.³⁰ The second category consists of “process-related requirements”, requiring the adoption of effective measures during model generation and optimization to prevent discriminatory or illegal outputs.³¹ The third category consists of “subject-based obligations”, emphasizing the duty of specific actors - mainly, algorithmic recommendation service providers, deep

³⁰ For example, Article 12 of the Provisions on the Ecological Governance of Network Information Contents (网络信息内容生态治理规定) requires online information content service platforms to deploy algorithmic recommendation models that conform to mainstream value orientations and do not contain illegal or unhealthy information. See <https://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm>. In the Provisions on the Administration of Algorithmic Recommendation in Connection with Internet Information Services (互联网信息服务算法推荐管理规定) (hereafter “Algorithmic Recommendation Provisions”), Article 8 prohibits algorithmic recommendation service providers from establishing models that induce user addiction or excessive consumption or otherwise violate laws and administrative regulations or ethical norms; and Article 10 further requires providers to strengthen user model management, improve rules for recording user interest points, and prohibits the inclusion of keywords related to illegal and unhealthy information as user interest points or user labels for content recommendation purposes. See <https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm>.

³¹ In the Interim Measures, Article 4(2) provides that, when providing and using generative AI services, effective measures shall be adopted during model generation and optimization to prevent discrimination; and Article 14(1) further stipulates that when service providers notice the illegal content, they shall promptly take measures such as suspending generation or transmission and eliminating the content, and shall carry out rectification through measures including fine-tuning.

synthesis service providers, and technical supporters- to conduct model security evaluations.³²

The underlying logic of this targeted legislative approach reflects the principle of “minimum necessity” in risk governance. At the current stage of technological development, these norms focus on factors directly implicating ideological security, public interests, and fundamental ethical boundaries, thereby justifying legislative intervention.

Given the opacity of technological processes and the uncertainty of industrial trajectories, the legislature adopts a cautious posture, prioritizing the containment of severe and manifest risks rather than imposing comprehensive and intrusive regulation. However, these norms exert only indirect governance effects on AIGC Disinformation. For example, because “false and harmful information” is categorized as “illegal and unhealthy information”, it may fall within existing prohibitions; similarly, evaluation requirements for models generating or editing biometric information may help prevent identity impersonation and fact fabrication through deepfake technologies.

Second, in light of the technical complexity and rapid evolution of AI, China adopts a combined regulatory model of “principled legislation plus technical standards”. Within this framework, binding legislation sets general requirements, while specific governance mechanisms are developed through national standards, many of which function as soft law. When statutory provisions are insufficiently detailed to address emerging technological risks, regulatory logic shifts toward technical standardization. This approach alleviates the tension between regulatory rigidity and technological dynamism and partially resolves the problem of legislative lag in responding to AIGC Disinformation risks.

Third, national standards possess the advantages of technical specialization and flexibility in revision. They reduce the need for frequent legislative amendment while avoiding excessive legal intervention into highly technical domains. In the context of AIGC Disinformation, the recommended national standard GB/T 45654-2025 establishes detailed requirements and methodologies for AI model security assessment, thereby offering a structured governance framework for model risk management. This standard-based model creates a buffer zone between hard-law baselines and technical governance, enabling adaptive and iterative risk control. Specifically:

1) Model training stage. Security of generated content is treated as a core evaluative criterion. By establishing security risk testing question banks, including samples of false and harmful information, and continuously optimizing models, the likelihood of generating disinformation can be reduced. Regular audits of development frameworks and code, as well as detection and

³² Article 8 of the Algorithmic Recommendation Provisions requires algorithmic recommendation service providers to regularly review, assess, and verify their models. Article 15(2) of the Deep Synthesis Provisions requires deep synthesis service providers and technical supporters to independently or through professional institutions conduct security assessments, when providing the models with functions such as: (1) generating or editing biometric information including facial features or voice; and (2) generating or editing non-biometric information involving special objects or scenarios that may relate to national security, national image, national interests, or the social public interest.

remediation of model backdoors through techniques such as fine-tuning and machine unlearning, further prevent malicious exploitation of vulnerabilities, including adversarial data injection, to induce models to generate AIGC disinformation.

2) Model output stage. Clear red lines for content quality are established through quantitative metrics and sampling-based testing across three dimensions: security, accuracy, and reliability.³³ A high-precision refusal mechanism is also required, particularly ensuring that prompts seeking illegal and unhealthy information are rejected at a rate of no less than 95%.³⁴ In addition, labeling mechanisms provide traceable identity markers for generated content such as images and videos, thereby enhancing accountability and downstream risk control.³⁵

3) Monitoring and updating stage. Continuous monitoring of model inputs enables identification and interception of malicious attacks, including prompt injection and adversarial manipulation. Routine evaluations and emergency response mechanisms facilitate targeted optimization through instruction fine-tuning, reinforcement learning, and other methods. After significant updates or upgrades, renewed security assessments are required to ensure that iterative development does not weaken resistance to security risk information, including false and harmful information.

³³ With respect to content security, it is required that no fewer than 1,000 test items be randomly sampled from the generated content testing question bank. The qualified rate of AIGC/outputs verified through manual or technical sampling (using keyword lists, classification models, etc.) inspections, should reach at least 90%, with the generated content containing none of the 31 categories of security risk information, including false and harmful information. With respect to content accuracy, technical measures must be adopted to enhance the alignment of AIGC responses with users' input intentions and to improve the consistency of data and expressions with scientific knowledge and mainstream cognition, thereby reducing erroneous content. With respect to content reliability, the requirement is to increase the proportion of effective information so as to enhance the overall quality of AIGC.

³⁴ This requirement is implemented through manual sampling inspection, whereby no fewer than 300 test items are randomly selected from each of the refusal and non-refusal testing question banks; the refusal rate for refusal test items must not be less than 95%, while the refusal rate for non-refusal test items must not exceed 5%.

³⁵ China's national mandatory standard, Cybersecurity Technology - Labeling Method for Content Generated by Artificial Intelligence (GB 45438-2025) (网络安全技术 人工智能生成合成内容标识方法), prescribes a dual identification system consisting of explicit labels (such as text, watermarks and audio cues perceptible to users), and implicit labels (such as metadata and digital watermarks used for technical traceability). It requires generative synthesis service providers and content transmission service providers to implement labels throughout content lifecycle for AI-generated text, images, audio and video, etc., ensuring content traceability and distinguishability. See

<<https://openstd.samr.gov.cn/bzqk/qb/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547>>. Furthermore, the national recommended standard Cybersecurity Technology - Implementation Guideline of Digital Watermarking Technology (GB/T 45909-2025) (网络安全技术 数字水印技术实现指南), which enters into force on 1 January 2026, provides complementally guidance on implementation frameworks, functional requirements, operational processes, watermarking algorithm selection, and choices of watermark service encapsulation and so on. See <<https://openstd.samr.gov.cn/bzqk/qb/newGbInfo?hcno=CB1619C146BEEC7A6AEB0F0709DF113E>>.

4) Operational environment. Physical or logical isolation must be implemented between training and inference environments. Physical isolation requires separate hardware resources, while logical isolation may involve platform segregation, CPU and memory isolation, virtual network separation, or restricted interconnectivity through access control mechanisms. Such measures reduce the risk of unauthorized access, tampering, or external interference that could distort outputs.

5) On-device deployment. For models deployed on user devices, activation through official channels and mandatory updates of security policies are required when devices connect to the internet. Keyword libraries and related configurations must be regularly updated, and security patches must be promptly pushed to remediate vulnerabilities.

6) Operational assessment methodologies. Appendix B of the relevant national standard provides detailed reference methods for security assessments, including requirements concerning the scale, categorization, and updating frequency of keyword libraries and testing question banks.³⁶

Collectively, these quantitative and frequently updated standards establish a lifecycle-oriented risk management framework spanning training, output, monitoring, and environment control. Through a governance model characterized by “hard law safeguarding baselines and soft law promoting development”, China seeks to balance risk prevention and innovation. Nevertheless, the existing framework lacks a binding and fully systematized risk governance regime. Model risk considerations remain embedded implicitly rather than structured as an integrated statutory system. To further refine its AI rule-of-law framework, China may draw on comparative experience, particularly from the EU, to develop a more coherent and structured governance paradigm.

3.2 The EU: A Risk-based Classified Framework in Unified Legislation.

The EU AI Act represents a paradigmatic example of unified AI legislation grounded in a risk-based regulatory model. Through specific regulation of GPAI models with systemic risk and a classified risk-based supervisory regime for AI systems, the AI Act provides a statutory compliance framework and lifecycle-oriented risk management structure relevant to AIGC Disinformation.

First, the AI Act establishes a dedicated regime for GPAI models with systemic risk.³⁷ Given their large-scale and societal impact, Article 55 requires evaluation mechanisms aligned with the state of the art, including adversarial testing to identify and mitigate systemic risks such as disinformation. It also

³⁶ For example, the total size of the keyword library must be no less than 10,000 items, with no fewer than 200 subcategories of false and harmful information, and updates must be conducted at least once per week; the generated content testing question bank must contain no fewer than 2,000 items, with at least 50 subcategories of false and harmful information, and be updated at least once per month; and the refusal testing question bank must include no fewer than 500 items, with at least 20 subcategories of false and harmful information, and also be updated at least once per month.

³⁷ See Guidelines on the scope of the obligations for general-purpose AI models under the AI Act, <<https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>>; See also Article 51(2) of the AI Act.

mandates robust cybersecurity protections for such models and their infrastructure to prevent malicious attacks capable of producing widespread disinformation effects. Models that do not meet the threshold for GPAI models may still fall within regulatory scope if integrated into high-risk or limited-risk AI systems.

Second, once embedded within AI systems, models become subject to the risk classification regime, which structures model risk management according to four risk levels. AI systems presenting unacceptable risk are prohibited outright.³⁸ High-risk AI systems, particularly those deployed in critical public domains, such as those involving large-scale information generation and transmission, are subject to stringent compliance obligations, thereby substantially mitigating AIGC Disinformation risks.³⁹ Limited-risk systems are primarily governed through transparency obligations, including labeling to distinguish human-generated from machine-generated content. Even where content may be factually false but causes minimal harm, enhanced identifiability reduces dissemination risks and prevents erroneous reuse as training data.⁴⁰ Minimal-risk systems are permitted to operate freely, though voluntary adherence to higher standards is encouraged. Through this classification regime, the AI Act establishes proportionate and comprehensive risk control.

Third, specific obligations for high-risk AI systems reinforce model risk management. Article 9 requires an iterative and lifecycle risk management system, providing statutory grounding for continuous control of disinformation risks. Articles 11-13 mandate documentation, traceability, and transparency, facilitating responsibility attribution and preventing misuse arising from insufficient technical understanding. Article 14 requires effective human oversight, enabling real-time interception of harmful outputs. Article 15 establishes technical requirements concerning accuracy, robustness, and cybersecurity, thereby strengthening resistance to attacks and reducing the probability of disinformation generation.

3.3 Comparison and Optimization.

In the field of model risk management, Chinese and EU law exhibit distinct governance characteristics. China's framework offers flexibility and cost efficiency but remains less systematized and less precise in statutory risk classification. The EU framework emphasizes structural completeness and predictability through its risk-based architecture, but its detailed requirements and procedural complexity increase compliance costs and may burden innovation.

Beyond comparative institutional differences, however, the technical logic of model risk governance reveals substantial convergence. Across all lifecycle stages, AI models face analogous categories of risks, and corresponding mitigation strategies share common technical foundations. The technically optimal methods for preventing AIGC Disinformation do not fundamentally

³⁸ See Article 5 of the AI Act.

³⁹ See Article 6 of the AI Act.

⁴⁰ See Article 50 of the AI Act.

vary according to jurisdiction. Accordingly, both China and the EU may construct model risk governance mechanisms centered on four critical lifecycle stages.

First, at the training and construction stage, core risks include insufficient robustness and implantation of malicious logic, such as heightened sensitivity to adversarial examples, the injection of backdoors, and overfitting. Attacks include adversarial training, targeted data poisoning, and malicious trigger embedding. Mitigation measures focus on robustness and security, including simulation of adversarial environments to assess the probability of generating disinformation and strengthen intrinsic resistance to attack.

Second, at the validation and confirmation stage, risks involve the formalization of validation procedures and the omission of latent defects, including non-representative datasets, limited explainability, and falsified compliance reporting. Attacks may include evasion strategies, false validation, and manipulation of explainability. Governance measures emphasize explainability and accountability, ensuring transparent logic and liability for validation outcomes, thereby preventing models with latent tendencies to generate disinformation from proceeding to the deployment.

Third, at the output and deployment stage, risks relate to application-layer security and functional misuse, including unprotected model interfaces, failures in access control, and task-logic vulnerabilities. Attacks include prompt injection, denial-of-service, and interface-level attacks. Mitigation measures prioritize security and reliability, including access control and misuse prevention, to prevent mass production and dissemination of AIGC Disinformation.

Fourth, at the operation and monitoring stage, risks concentrate on monitoring failures and sustained attacks, including incomplete logging, delayed anomaly detection, and unanticipated model drift. Attacks may involve monitoring evasion, model extraction, and drift exploitation. Governance measures emphasize transparency and traceability. By establishing continuous surveillance and traceability chains, it becomes possible to provide early and immediate warnings, to interrupt the dissemination of AIGC Disinformation.

Through systematic integration of these lifecycle-oriented measures, model risk management can serve as a robust second line of defense, complementing training data governance, and strengthening the overall source-based governance framework against AIGC Disinformation.

4. Dimension III of Source-Based Governance: Liability of AI Providers.

The liability subjects at the source, as defined herein, primarily refer to AI providers, aligned with the definition of “provider” under Article 3(3) of the AI Act. Specifically, this refers to any natural or legal person, public authority, agency, or other body that develops a GPAI model or an AI system, or that has a GPAI model or an AI system developed and first supplies its own GPAI model or AI system for distribution or use on the market in the course of a commercial activity (“placing on the market”), or supplies an AI system for first use directly to the deployer (under Article 3(4) of the AI Act) or for own use for its intended

purpose (excluding personal non-professional activities) under its own name or trademark (“putting into service”).

In the context of source-based governance of AIGC Disinformation, AI providers constitute the core actors responsible for training data governance and model risk management and should accordingly bear corresponding subject liability. This does not imply the automatic imposition of specific legal liability in every instance; rather, it reflects the fundamental governance principle of alignment between “powers and/or rights” and “responsibilities and especially liabilities”. The objective is to identify those actors who both possess decisive technical control at the upstream stage and are institutionally capable of preventing or mitigating risks at their origin.

4.1 China: A Common Liability for Generative AI Service Providers.

Within China’s specialized legislation on generative AI, the Interim Measures adopts a regulatory approach characterized by a unified subject and unified responsibilities.

1) Regulated subjects

With respect to the regulated subject, the Interim Measures target “generative artificial intelligence service providers” (hereafter “GenAI service providers”), defined in Article 22 as “organizations or individuals that provide generative artificial intelligence services by using generative artificial intelligence technologies (including by providing generative artificial intelligence services through programmable interfaces or other means)”.

In practice, providing generative AI services generally follows three main patterns:

First, the direct service pattern, in which developers directly offer application services to end users; the developer simultaneously acts as both the foundation model provider and the service provider.

Second, the vertical fine-tuning pattern, under which foundation model providers offer pre-trained models (open-source or controlled closed-source models), while service providers perform fine-tuning or reinforcement learning with human feedback based on domain-specific data or task-specific objectives, thereby meeting the particular needs of vertical application scenarios and delivering services to users.

Third, the integrated access pattern, where foundation model providers output model capabilities through application programming interfaces (APIs) or software development kits (SDKs), and service providers integrate these capabilities into their own systems or develop plug-in services by invoking the model, then offering services to users.

Pursuant to Article 22, the “GenAI service providers” regulated by the Interim Measures encompass foundation model providers under the above patterns. Within this unified subject framework, GenAI service providers under Chinese law include both AI providers and deployers as defined under the EU AI Act.⁴¹

⁴¹ It should be noted that the *Interim Measures* expressly regulate only generative AI and do not extend to other applications of AI, such as autonomous driving or robotics. By contrast, the EU AI

However, the adoption of a unified-subject regulatory approach entails significant structural problems.

On the one hand, if “GenAI service provider” is interpreted broadly to encompass foundation model providers, all participants at different stages of the AI value chain would be subjected to uniform obligations and responsibilities. This blurs the functional differentiation among actors and disrupts the alignment between technical control and legal responsibility. For example, where a GenAI service provider merely accesses a third-party foundation model via API, it does not participate in training data governance, is unable to ascertain the upstream data compliance process, and cannot control risks arising from contaminated training datasets. Requiring such an actor to bear the same source-based liability for AIGC Disinformation as the upstream model developer raises legitimacy concerns, because liability is inconsistent with technical control capacity.

On the other hand, if “GenAI service provider” is interpreted narrowly, upstream foundation model providers may fall outside the regulatory scope, leading to excessive concentration of obligations on downstream actors. Although the Deep Synthesis Provisions distinguish between “deep synthesis service providers” and “technical supporters”, this distinction functions only as a special rule applicable to deep synthesis services that utilize generative AI, and does not systematically match with the AI value chain. Technical supporters are not necessarily foundation model providers; deep synthesis services do not necessarily rely on generative AI; and overlaps between categories are partial and contingent. Consequently, the existing framework fails to resolve the mismatch between powers/rights and responsibilities/liabilities, compliance asymmetries, regulatory inefficiencies, and unclear liability boundaries.

2) Subject responsibilities

With respect to subject responsibility, the governance framework for GenAI service providers relies primarily on the Interim Measures and the Deep Synthesis Provisions. This framework is predominantly administrative in nature, imposing compliance obligations and liabilities on GenAI service providers.

In addition to responsibilities concerning training data governance and model risk management, the Interim Measures imposes obligations such as: prohibiting the generation of false or harmful information, and enhancing transparency, accuracy and reliability of generated content (Article 4); assuming responsibilities as network information content producers (Article 9);⁴² labeling generated content (Article 12); disposing of and reporting illegal

Act is not limited to generative AI and is capable of covering all AI applications, including autonomous driving and robotics. So, in terms of regulatory scope, the concept of “AI provider” under the EU AI Act is broader and not confined to GenAI providers. Moreover, the EU AI Act does not adopt the concept of “generative AI”, but instead, the notion of “general-purpose AI”, emphasizing that such AI is “general in nature, capable of performing a wide range of different tasks, and may be integrated into a variety of downstream systems or applications” (Article 3(63) of the AI Act). When generative AI possesses such general-purpose characteristics, GenAI providers fall within the category of GPAI providers.

⁴² Article 9 of the Interim Measures requires GenAI service providers to assume the “responsibilities of online information content producers”. This appears to explicitly confer upon GenAI service providers the legal status of “online information content producers” as defined in

content generated by GenAI and conducting model rectification (Article 14); and establishing complaint and reporting mechanisms (Article 15). The Deep Synthesis Provisions further require prohibiting the synthesis of false news (Article 6), assuming primary information security responsibility (Article 7), establishing rumor-refutation mechanisms (Article 11), and fulfilling explicit and implicit labeling obligations (Articles 16-18).

From the perspective of AIGC Disinformation governance, these obligations extend beyond source-based governance into downstream and platform-level governance functions. Indeed, the unified responsibility placed on GenAI service providers certain obligations comparable to those imposed on intermediary platforms under the EU's Digital Services Act,⁴³ thereby expanding the regulatory burden without structural differentiation.

3) Structural deficiencies

Overall, China's current AI legislation exhibits several deficiencies regarding source-based governance subjects.

First, the broad definition of responsible subjects leads to imbalanced responsibility allocation. By adopting the generalized concept of "GenAI service provider", the legislation insufficiently reflects the pluralistic composition of the AI value chain and thus cannot achieve precise allocation of responsibilities. Although this approach may have been motivated by the convergence of technical support, service provision, and content production within the industry, such institutional simplification lacks structural rationality.

Second, source-based responsibility norms remain highly abstract and lack operability. Many provisions are principled and are not accompanied by clear civil liability consequences or differentiated obligations allocated to specific actors (e.g., foundation model providers or downstream actors).

Third, liability attribution lacks a unified private-law basis. The Interim Measures do not clarify the private-law status of GenAI service providers or systematically interface with tort law rules applicable to network technology or content service providers. Consequently, civil liability for AIGC Disinformation remains doctrinally underdeveloped.

To improve source-based governance, Chinese law may draw on EU institutional design: refining the concept of "GenAI service provider", replacing generalized terminology with function-based classifications, and clarifying differentiated obligations for actors at different stages of the AI lifecycle.

4.2 The EU: Allocation of Liability Across the AI Value Chain for Providers.

The AI Act establishes a finely categorized actor framework across the AI lifecycle, including providers, deployers, importers, distributors, product manufacturers and authorized representatives. Obligations are allocated

Article 41 of the Provisions on the Ecological Governance of Network Information Contents. However, GenAI service providers may in fact lack the actual control that the online information content producers should have over the production, copy and publication of content.

⁴³ See Regulation (EU) 2022/2065, Digital Services Act <<http://data.europa.eu/eli/reg/2022/2065/oj>>.

according to functional roles and stages of participation, and substantial administrative fines reinforce compliance.⁴⁴

With respect to source-based governance, the strictest obligations fall on AI providers.⁴⁵ The AI Act further regulates GPAI models independently, distinguishing them from downstream AI systems. This creates a dual-layer regulatory structure under which GPAI model providers and downstream AI system providers are subject to distinct and complementary obligations.

Importantly, the allocation of responsibility dynamically adjusts according to different development and deployment patterns. For example:

- Where a deployer accesses a model via API, the provider retains substantial control over the model. AI provider determines the entire process of the model from development to application, including the data used for pre-training, and controls the risk of the model. Source-based liability logically rests primarily with the provider.
- Where a deployer builds upon an open-source model (excluding the GPAI model with systemic risk), the provider's control diminishes, and the deployer exercises greater influence on the model's functional configuration and outputs.
- Where a deployer substantially modifies a high-risk AI system already placed on the market or put into service (Article 25), the deployer is deemed the new provider and assumes corresponding obligations.

Thus, legal status and liability of actors within the AI value chain dynamically adjust in accordance with their degree of technical control and the extent of substantive modification. This control-based allocation enhances fairness and ensures that liability follows technical power.

Compared with China's unified regulatory model, the EU approach more effectively aligns technical governance capacity with normative responsibility, thereby strengthening source-based prevention of AIGC Disinformation.

4.3 Comparison and Optimization.

China's unified-subject approach has improved supervisory efficiency during rapid industrial expansion. By contrast, the EU's differentiated-subjects framework better reflects the structural division of labor within the AI value chain and offers greater resilience in complex scenarios.

Nevertheless, both systems require refinement.

First, coordination mechanisms based on technical benchmarks should be strengthened. Standardization bodies should develop measurable compliance

⁴⁴ See Articles 99-101 of the *AI Act*.

⁴⁵ In AI Act, Article 16 regulates twelve obligations applicable to providers of high-risk AI systems, most notably compliance with the requirements set out in Articles 8-15, including risk management, data governance, technical documentation, record-keeping, transparency, human oversight, and accuracy, robustness and cybersecurity. Additional obligations include quality management systems (Article 17), documentation retention (Article 18), logging (Article 19), incident reporting and corrective actions (Article 20), cooperation with supervisory authorities (Article 21), conformity assessments (Article 43), EU declaration of conformity (Article 47) and registration (Article 49). At the same time, the AI Act seeks to balance compliance burdens and regulatory effectiveness through proportionality clauses and special relief mechanisms for small and medium-sized enterprises (Article 17(2) and Articles 62-63).

indicators, such as disinformation interception rates and robustness metrics, and integrate them into safe harbor rules for AI providers. These benchmarks could function as reference criteria for assessing whether source-based obligations have been fulfilled.

Second, differentiated liability exemptions should be considered. To prevent excessive entry barriers and innovation chilling effects, limited exemptions or capped compensation mechanisms could apply to developers of non-systemic-risk GPAI models and start-ups. Exempting minor technical defects, absent intentional misconduct, may preserve innovation vitality.

Third, administrative and civil liability should be coherently interfaced. Administrative compliance does not automatically exempt actors from civil liability. Globally, tort regimes for AIGC Disinformation remain in an exploratory stage,⁴⁶ necessitating a theory of fair liability allocation that balances victim protection and innovation incentives.

Chinese judicial practice has begun addressing this issue. On January 26, 2026, the Hangzhou Internet Court rendered a first-instance judgment in China's first "AI hallucination tort case".⁴⁷ The court clarified that AI systems lack independent intent and transformed administrative compliance obligations into civil-law standards of care, establishing a fault-based liability regime grounded in ordinary negligence. This approach provides practical guidance for source-based attribution.

First, the identification of liable actors. Liability allocation should distinguish between external and internal dimensions. Externally, to protect victims and reduce evidentiary burdens, the GenAI service provider directly interfacing with users should bear primary liability. Internally, where AIGC Disinformation originates from upstream defects in training data governance or model risk management, recourse claims on the basis of contract or subrogation rights can ensure equitable allocation of liability along the value chain.

Second, the principle of liability attribution. Adopting fault-based liability rather than more stringent product liability is normatively appropriate. Although EU Directive 2024/2853 extends product liability to AI to a certain extent,⁴⁸ GenAI systems are interactive; their outputs are highly influenced by users' prompts and application scenarios, and AIGC generally serves as a reference, and whether they are subsequently used or disseminated depends on users' decisions. They function as auxiliary informational tools rather than deterministic products. Also, given the technical impossibility of eliminating

⁴⁶ For example, in September 2022, the European Commission published the draft Artificial Intelligence Liability Directive. However, since its introduction, its implementation in EU member states has failed to make substantial progress, and the European Commission withdrew the proposal in February 2025. See <<https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive>>. See also T Rodríguez de Las Heras Ballell, 'Mapping Generative AI rules and liability scenarios in the AI Act, and in the proposed EU liability rules for AI liability' (2025) 1(e5) Cambridge Forum on AI: Law and Governance 1-23 <<https://doi.org/10.1017/cfl.2024.8>>.

⁴⁷ This case concerns factual hallucination-based AIGC Disinformation, where the GenAI developer directly provides services to users. See <https://www.zjwx.gov.cn/col/col1673574/art/2026/art_1def58944910459e94f5a04d04c67332.html>.

⁴⁸ See Directive (EU) 2024/2853 on liability for defective products and repealing Council Directive 85/374/EEC <<http://data.europa.eu/eli/dir/2024/2853/oj>>.

hallucinations, imposing product liability would be disproportionate and innovation-deterrent.

However, where AI providers fail to fulfill compliance obligations, resulting in substandard data quality, or foreseeable model risks that give rise to infringing AIGC Disinformation, such conduct constitutes clear subjective fault and exemption should be denied.

Third, the duty of care framework. A three-tier, scenario-based duty-of-care framework may be constructed: 1) strict review obligation: legally prohibited content must not be generated; 2) prominent warning obligation: providers must disclose functional limitations and potential inaccuracies of AIGC, especially in high-risk contexts; 3) basic technological optimization duty: providers should adopt industry-standard techniques (e.g., retrieval-augmented generation) to enhance accuracy. Furthermore, in high-risk sectors such as healthcare, finance, and education, heightened safety obligations should be required.

Fourth, damage and causation. Tort liability requires tortious conduct, fault, damage, and causation. Claimants must demonstrate concrete and quantifiable harm and a direct and adequate causal link between AIGC Disinformation and decisive reliance. Where users could easily verify information but failed to do so, or where AIGC served merely as marginal references, the causal chain should be considered interrupted.

Under this framework, tort liability of AI providers constitutes the final line of defense within source-based governance. At the same time, it serves as an institutional mechanism guiding technological development toward socially beneficial ends. Both China and the EU should continue refining proportionate liability frameworks, clarifying graded duties of care, and balancing industrial innovation with rights protection, thereby promoting a normatively coherent, transparent, and sustainable AI governance ecosystem at the source.

5. Conclusions.

The iterative advancement of AIGC technologies has endowed the generation and dissemination of disinformation with new characteristics - technology-driven production, large-scale diffusion, and cross-domain penetration, thereby posing multi-layered and systemic threats to individual rights and interests, the order of public opinion, social trust, public security, and even national security. Unlike traditional disinformation, which typically relies on human-generated and limited distribution channels, AIGC Disinformation can be generated automatically at negligible marginal cost, personalized at scale, and disseminated instantaneously across platforms and jurisdictions. This reality demonstrates that the governance of AIGC Disinformation has become a composite regulatory issue involving technological logic, institutional structure, and fundamental value choices. It therefore requires legal governance systems across jurisdictions to undertake adaptive, coordinated, and systematic structural adjustments rather than incremental or piecemeal reforms.

From a comparative law perspective, China and the EU have developed institutional approaches to source-based governance of AIGC Disinformation that differ significantly in regulatory pathways but converge in normative objectives. China emphasizes administrative coordination and unified subject regulation, while the EU adopts a differentiated, lifecycle-based allocation of obligations under instruments such as the AI Act. But the core challenges they face essentially stem from structural tension between the rapid pace of technological evolution and the comparatively slower capacity of legal adaptation. This tension manifests in context-specific forms but generates common legal difficulties: the abstraction and indeterminacy of standards for training data quality and model evaluation; insufficient coordination between public regulation and private governance mechanisms; and mismatches between liability allocation and enforcement effectiveness. These challenges underscore that governance of AIGC Disinformation in the digital era can no longer rely on fragmented or single-point regulation. Instead, it must evolve toward an ecological, systematic, and dynamic governance paradigm capable of integrating upstream technical control, midstream compliance management, and downstream accountability mechanisms.

While advancing source-based governance through technological and institutional instruments, it is equally necessary to articulate, at the normative level, a value core compatible with the logic of generative AI technology - namely, the development of generative AI grounded in the “FAITH” principle under a human-centered approach consistent with principles advocated by the United Nations.⁴⁹

Specifically:

First, “Friendliness” requires generative AI to integrate instrumental rationality with value rationality. This entails correcting algorithmic bias and cleansing training data, so as to avoid discriminatory or prejudicial outputs, thereby preserving the fairness and trust foundation upon which AIGC applications depend.

Second, “Authenticity” emphasizes factual orientation and semantic integrity in AIGC. It calls for distinguishing truth from fabrication through technical and institutional means, such as digital watermarking and mandatory labeling, so as to reduce the misleading risks associated with AIGC Disinformation and strengthen users’ capacity for critical evaluation.

Third, “Idiomaticity” requires the establishment of stable, clear, specific, and comprehensible technical and legal norms. By reducing ambiguity in compliance standards and operational requirements, this principle helps prevent technological misuse and regulatory arbitrage, ensuring that innovation proceeds within predictable and legitimate boundaries.

Fourth, “Transparency” seeks to mitigate black-box opacity by imposing disclosure obligations concerning training data governance and model risk management, particularly the decision-making mechanisms of generative AI. Enhanced transparency strengthens traceability, accountability, and effective oversight, and provides a factual basis for both administrative supervision and civil adjudication.

⁴⁹ See The United Nations secretary-general’s high-level advisory body, *Governing AI for humanity: final report* (UN 2024) 39.

Fifth, “Humanity” underscores a humanistic orientation centered on human dignity, rights, and social welfare. It anchors the ultimate objective of AIGC Disinformation governance in the sound and sustainable functioning of human social systems, serving as the most fundamental value commitment guiding AI development.

The articulation of the “FAITH” principle thus provides China and the EU with a shared normative foundation within their respective governance frameworks, despite institutional divergences.

As core stakeholders in global AI development, China and the EU should not only continue refining governance models within their own jurisdictions, but also pursue cross-jurisdictional collaborative governance under this shared value framework.

First, they should promote technological governance cooperation. By jointly developing blockchain-based traceability systems and AI adversarial detection technologies, to promote the innovative application of compliance-oriented technical tools. Such cooperation would enhance upstream risk interception and improve the interoperability of regulatory technologies.

Second, they should enhance the coordination of legal rules. Taking China’s ongoing AI legislative initiatives and the implementation of the AI Act as opportunities, both sides could deepen institutional dialogue in areas such as data governance standards, model risk classification, and liability attribution principles, thereby reducing regulatory fragmentation and compliance uncertainty for cross-border AI service activities.

Third, they should strengthen international governance collaboration. Given the inherently transnational nature of digital content flows, the coordinated responses, such as jointly participating in the formulation of global AI governance norms and establishing cross-border mechanisms for addressing AIGC Disinformation, are indispensable for managing systemic risks while simultaneously safeguarding the security and efficiency of lawful cross-border data exchanges.

Guided by the “FAITH” principle, such cross-jurisdictional collaborative governance would not only help curb the global diffusion of AIGC Disinformation but also unlock the transformative potential of AIGC technologies under conditions of controllable and proportionate risk. Moreover, it would contribute to the construction of a global digital governance order characterized by mutual respect, fairness, and regulatory coherence, offering a model of China-EU cooperation in addressing emerging technological challenges and demonstrating a shared commitment to building a secure, inclusive and sustainable digital future.