WEP Warwick Evans Publishing

International Journal of Social Sciences and Public Administration

ISSN: 3005-9836 (Print), ISSN: 3005-9585 (Online)
Volume 10, Number 8, Year 2026
DOI: https://doi.org/10.62051/ijsspa.v10n8.03

A Control-Based Differentiated Framework for the Content Governance Obligations of Generative Artificial Intelligence Service Providers

Shengtao Tai

Law School, Beijing Normal University, Beijing, China

ABSTRACT

The content governance obligations of generative artificial intelligence (GAI) service providers constitute a central issue in contemporary AI regulation. Existing legal frameworks, however, impose substantially uniform content governance obligations on GAI service providers without adequately accounting for the significant differences among various service models. Drawing upon risk-control theory, this Article argues that the content governance obligations of GAI service providers should be differentiated according to the degree of substantive control they exercise over the content-generation process. Based on differences in such control, GAI service models may be categorized into three types: pure prompt-based generation services, retrieval-augmented generation (RAG) services, and foundation model fine-tuning services. Correspondingly, each category should be subject to a distinct set of content governance obligations calibrated to its respective level and form of control. A differentiated regulatory framework that allocates content governance obligations on the basis of substantive control can better reconcile the dual objectives of safeguarding the security of AI-generated content and promoting technological innovation, thereby achieving a more proportionate and refined balance between regulation and innovation.

KEYWORDS

Generative Artificial Intelligence; Content Governance Obligations; Duty of Care; Risk-control Theory.

1. INTRODUCTION

Generative artificial intelligence (GAI), with its powerful capacity to generate content, has become increasingly embedded in processes of knowledge production and information dissemination. At the same time, it has given rise to significant legal and regulatory concerns regarding the safety and accuracy of AI-generated content. Against this backdrop, the scope of the duty of care borne by GAI service providers, together with the circumstances under which they may incur liability for unlawful or infringing content, has emerged as a central and complex issue in the age of artificial intelligence [1]. Existing legal frameworks fail to adequately account for differences in the degree of substantive control exercised by GAI service providers over the content-generation process. As a result, current regulatory approaches often impose equivalent standards of obligation on service providers with different service models and varying capacities of control. Such a uniform regulatory approach may lead to a mismatch between legal obligations and providers'actual capabilities, thereby reducing the effectiveness of regulation.

To achieve an appropriate balance between safeguarding the security of generated content and promoting innovation in artificial intelligence, regulatory frameworks should take into full

consideration the degree of control exercised by GAI service providers and establish a tiered, categorized, and differentiated framework of content governance obligations.

2. THE CURRENT REGULATORY FRAMEWORK AND CHALLENGES OF CONTENT GOVERNANCE OBLIGATIONS FOR GAI SERVICE PROVIDERS

2.1 The Current Regulatory Framework Governing Content Governance Obligations for GAI Service Providers

The content governance obligations imposed on GAI service providers are not established through a single legal instrument; rather, they are constituted by a multi-layered regulatory framework comprising various legal norms. With respect to the substance of these obligations, existing legal rules primarily encompass the following requirements:

Content Safety Review Obligations. Content safety review obligations constitute the most fundamental and strictly regulated obligations imposed on GAI service providers. These obligations may be further divided into two levels.

The first level concerns baseline prohibitions on content generation, under which service providers are prohibited from generating content expressly prohibited by law, including content that incites the subversion of state power, endangers national security, promotes terrorism or extremism, advocates violence, contains obscene or pornographic materials, or disseminates harmful information.

The second level involves procedural obligations relating to governance and management. Once illegal or undesirable information is identified, GAI service providers are required not only to take proactive measures to address such content but also to fulfill reporting obligations to relevant regulatory authorities.

It should be noted that the ex ante review obligations imposed on service providers do not require the complete identification of all unlawful content. Rather, such obligations are limited to content that is manifestly unlawful. Specifically, the ex ante review obligations of service providers primarily focus on content that poses clear and substantial threats to fundamental human rights and public security [2].Content Identification Obligations. Content identification obligations represent the technical manifestation of the principle of transparency in GAI services. Their primary purpose is to ensure the traceability of AI-generated content and the identifiability of its source, thereby providing an essential foundation for the subsequent allocation of dissemination-related responsibilities and enabling users to make informed judgments regarding the provenance and authenticity of information.

Where AI-generated content is likely to mislead the public or cause confusion as to its origin or authenticity, such content undermines individuals' informational autonomy. In such circumstances, content dissemination platforms should be required to fulfill verification obligations with respect to the identification of AI-generated content. Moreover, in fields that place a premium on originality, such as scientific research, education, and the arts, content dissemination platforms should likewise undertake verification obligations in order to safeguard the moral interests of human creators and promote the overall welfare of the relevant creative and academic communities [3]. Generative artificial intelligence systems are frequently susceptible to model hallucinations, producing information that appears plausible but is factually inaccurate, meaningless, or detached from reality. Accordingly, regulatory rules impose content identification obligations on service providers in order to reduce the risk that users may suffer harm as a result of false, inaccurate, or incomplete information [4].Obligations to Ensure the Accuracy of Generated Content. With respect to the accuracy of AI-generated content, existing legal frameworks do not impose an absolute outcome-based obligation on GAI service providers. Although accuracy and reliability represent important normative objectives for the development of generative artificial intelligence, current regulations do not require service

providers to guarantee the factual correctness of generated content, nor do they impose ultimate responsibility for reviewing the accuracy of every output.

The absence of an outcome-based obligation, however, does not imply the absence of legal responsibility. Instead, existing regulations require GAI service providers to take effective measures to improve the accuracy and reliability of generated content. In addition, it is generally accepted that service providers are subject to a duty to provide appropriate warnings or disclosures, informing users of the inherent limitations of AI-generated content and the possibility that such content may be inaccurate or unreliable. These requirements reflect a risk-based regulatory approach, under which service providers assume obligations to enhance the accuracy of generated content and to provide appropriate user warnings, rather than to guarantee the accuracy of every individual output.

2.2 Regulatory Challenges Concerning the Content Governance Obligations of GAI Service Providers

In the context of generative artificial intelligence (GAI), the legal standards for determining the content governance obligations of service providers remain insufficiently defined under the current regulatory framework. Although existing legal rules establish a basic framework of obligations applicable to GAI service providers, they provide little guidance regarding the specific criteria for determining those obligations or the standards by which they should be assessed. While courts have adjudicated a number of disputes involving GAI-generated content, judicial practice has yet to develop a consistent approach to defining the scope of the duty of care. As a consequence, both the analytical reasoning and the legal standards adopted by courts vary considerably across individual cases.

A review of existing judicial decisions reveals three principal approaches to determining whether GAI service providers have breached their content governance obligations.

The first approach adopts a dual criterion based on foreseeability and preventability. Under this approach, whether a service provider has breached its duty of care depends on two considerations: whether, under the prevailing technological conditions, it could reasonably have foreseen the occurrence of the infringing conduct, and whether it possessed the ability to prevent or halt such infringement.

The second approach evaluates compliance with the duty of care by reference to institutional deficiencies. Rather than assessing the duty in the abstract, courts operationalize it through a series of objectively verifiable institutional safeguards, including whether the service provider has established effective complaint and reporting mechanisms, provided adequate warnings regarding potential risks, and clearly identified AI-generated content.

The third approach assesses the duty of care in light of the contemporaneous state of technological development. It focuses on whether the service provider failed to adopt necessary measures that were technologically feasible at the time the infringement occurred, despite having the capacity to do so. More specifically, courts employ the state of the art as the relevant benchmark, taking into account temporal, industrial, and geographical factors in determining whether the service provider has exercised the degree of care expected of a reasonably prudent service provider under comparable circumstances [5].

The three approaches outlined above demonstrate the judiciary's efforts to develop operational standards for assessing the content governance obligations of GAI service providers in response to the novel risks posed by generative artificial intelligence. Nevertheless, the existing regulatory framework imposes content governance obligations of substantially uniform intensity on all GAI service providers, despite significant differences in service architecture, application scenarios, user scale, and-most importantly-the degree of substantive control that providers exercise over generated content.

This uniform regulatory approach fails to reflect the heterogeneous nature of GAI services and therefore cannot ensure a proportionate allocation of legal responsibilities. Accordingly, the regulatory framework should move beyond a "one-size-fits-all" model and adopt a differentiated, tiered approach.Specifically, the content governance obligations imposed on GAI service providers should be calibrated according to the type of service provided, the provider's degree of control over generated content, and the level of risk associated with different application scenarios, thereby establishing differentiated standards and corresponding obligations.

3. THEORETICAL FOUNDATIONS FOR THE DIFFERENTIATED STRUCTURING OF CONTENT GOVERNANCE OBLIGATIONS FOR GAI SERVICE PROVIDERS

3.1 The Doctrinal Foundation of Differentiated Obligations: Risk-Control Theory

Risk-control theory constitutes one of the principal doctrinal approaches to defining the duty of care in tort law. Grounded in considerations of risk allocation and distributive justice tort law allocates responsibility for harm by identifying the party who exercises control over the source of the risk that gives rise to the injury [6]. Under Chinese tort law, the theoretical foundation for imposing liability for omissions primarily derives from the German doctrine of Verkehrspflichten (duties of safety in social interaction), which establishes a general legal obligation requiring individuals to protect the legitimate interests of third parties. Based on the specific relationship between the actor and either the source of risk subject to control or the legally protected interests exposed to danger, Verkehrspflichten may be further classified into two categories: duties to control sources of danger and duties to protect the legally protected interests of third parties. The former constitutes the primary theoretical basis relied upon by the majority of scholars in arguing that GAI service providers should assume safety assurance obligations [7]. The core proposition of risk-control theory can be summarized as follows: the responsibility borne by the creator of a risk derives from its capacity to control the source of that risk. The allocation of the duty of care should be proportionate to the actor’s actual ability to control the relevant risks; the greater the capacity for risk control, the more extensive the obligations that the actor is capable of and should be expected to assume.

More specifically, the imposition of safety assurance obligations is justified on the basis that the relevant actors may possess the ability to control risks and distribute losses, derive economic benefits from their activities, and reduce the overall social costs associated with the prevention of tortious harm. Where a potential responsible party has created a certain level of risk while also possessing the capacity to control such risk, it is appropriate to require that party to undertake corresponding safety assurance obligations to a reasonable extent [8].

3.2 The Justification for Applying Risk-Control Theory in the Context of GAI

The application of risk-control theory in the context of GAI is theoretically justified. Data providers, developers, service providers, users, and other relevant actors collectively constitute the machine learning value chain. Although the conduct of each actor at different stages of this chain does not, in itself, necessarily constitute an infringement, such conduct may increase potential risks associated with AI-generated content. At the same time, the relevant actors generally possess certain technical capacities to mitigate or eliminate these risks [9]. As the central actors responsible for the development and deployment of GAI technologies, GAI service providers initiate a new form of information service and communication activity by offering interactive services based on large language models to a broad user base. As initiators of such specific and abstract risks, service providers should, in the course of providing GAI services, adopt necessary measures based on the circumstances of each case to protect users’ legitimate interests from infringement.

Compared with ordinary users, service providers are better positioned to foresee potential risks and harms arising from the operation of GAI systems and possess the technical capacity to adopt appropriate measures to prevent or mitigate such harm. Therefore, by deploying artificial intelligence systems that inherently involve certain risks, GAI service providers continuously create and maintain a risk environment while simultaneously possessing actual capacities for risk control. Accordingly, they should assume corresponding duties of care.

4. A TYPOLOGICAL ANALYSIS OF THE DIFFERENTIATED CONSTRUCTION OF CONTENT GOVERNANCE OBLIGATIONS FOR GAI SERVICE PROVIDERS

The internal logic of risk-control theory requires that the allocation of duties of care be proportionate to service providers' actual capacity to control relevant risks. In other words, differences in risk-control capacity determine corresponding differences in the intensity of legal obligations. With respect to the specific control capacities of GAI service providers, different service models may be broadly classified into the following three categories for further analysis:

4.1 Pure Prompt-Based Generation Services

Pure prompt-based generation services represent one of the most common forms of GAI service provision. These services are centered on foundation models and rely on deep learning or other machine learning techniques. Users interact with the system by providing natural-language prompts, based on which the model directly generates outputs such as text, images, or video content [10]. Under this service model, providers typically possess proprietary foundation models and exercise substantial control over key elements of the model, including its architecture, training data, and parameter configurations. Users interact with the model primarily through prompts, while the entire process from data input and model inference to content generation is carried out by algorithmic systems designed and controlled by the service providers. From the perspective of risk control, GAI service providers exercise substantive influence over generated content through various means, including the design of model architectures, the selection of training data, the adjustment of algorithmic parameters, and the establishment of content filtering mechanisms. Accordingly, given their capacity to shape and mitigate risks arising from AI-generated content, providers should assume more significant legal responsibilities in the prevention of infringement [11]. From the perspective of risk control, service providers can intervene in specific types of generated content before or during the generation process through technical mechanisms such as content safety filtering models and keyword-based interception systems. Accordingly, they exercise a relatively high degree of control over generated outputs. However, such control primarily operates through preventive technical mechanisms implemented before or during content generation, rather than through comprehensive ex ante review of each individual output, given the massive volume and dynamic nature of generated content.

Therefore, in the context of pure prompt-based generation services, service providers possess relatively strong control capacities and should accordingly be subject to comparatively higher standards of content governance obligations. Nevertheless, such control is primarily reflected in the design, deployment, and improvement of ex ante and in-process technical safeguards, rather than in an obligation to guarantee the legality or accuracy of every individual output.

4.2 Retrieval-Augmented Generation (RAG) Services

Retrieval-augmented generation (RAG) technology integrates language models with information retrieval techniques and constitutes a hybrid architecture that enhances model capabilities by incorporating external knowledge sources to compensate for the inherent limitations of foundation

models' internal knowledge [12]. Such services provide foundation models with necessary informational support by integrating external knowledge bases and employ information retrieval techniques to obtain contextual information relevant to users' instructions or queries. By augmenting the original queries and prompts with retrieved information, RAG services enhance the generation process of foundation models [13]. RAG-based services possess the dual characteristics of both search engines and GAI systems and are therefore often referred to as AI-powered search services. In RAG-based service models, however, the degree of control exercised by service providers over the content-generation process exhibits distinct characteristics compared with pure prompt-based generation services.

On the one hand, during the indexing stage, RAG service providers construct external knowledge bases covering a wide range of subjects and domains and organize their contents into manageable text chunks to facilitate efficient indexing and retrieval. Consequently, by determining the scope of the knowledge base, configuring retrieval strategies, and adjusting the ranking and filtering mechanisms applied to retrieved information, service providers are able to exercise substantive control over the sources of information before the content-generation process begins [14]. Moreover, because the outputs of RAG services are typically accompanied by citations or references to the underlying information sources, the factual basis of the generated content is inherently traceable. As a result, service providers possess the practical ability to assess and verify the reliability of information sources within a reasonable scope.

On the other hand, RAG-based systems and their providers do not possess sufficient capacity to fully verify the authenticity and accuracy of retrieved information, and AI hallucinations may still arise due to inherent technical limitations or deficiencies in the retrieved content itself. When the external information retrieved by the system contains errors or biases, the content generated by the model based on such information may likewise be inaccurate.

Therefore, in RAG-based services, the control exercised by service providers is primarily reflected in their ex ante control over the underlying AI models and retrieval strategies, as well as their obligations to label generated content and verify content involving serious unlawful information. By contrast, their ability to control specific generated outputs remains relatively limited.

4.3 Foundation Model Fine-Tuning Services

Foundation model fine-tuning services represent an emerging form of GAI service provision. During the pre-training stage, foundation models acquire powerful general-purpose representation capabilities and extensive world knowledge, thereby providing a strong starting point for a wide range of downstream tasks. However, when applied directly to specific downstream applications, their performance is often suboptimal. To address this limitation, fine-tuning, a transfer learning technique, adapts pre-trained foundation models to the distinctive characteristics and requirements of particular tasks by adjusting their model parameters, thereby significantly improving task-specific performance [15]. Against this backdrop, parameter-efficient fine-tuning (PEFT) has emerged as an important approach to model adaptation. By training only a limited subset of model parameters, PEFT substantially reduces computational and storage requirements while preserving model performance to the greatest extent possible [16]. More specifically, in AI applications such as text-to-image and image-to-image generation, PEFT techniques enable users to fine-tune existing foundation models using only a limited number of images or textual examples, often through low-rank adaptation, thereby allowing the models to consistently generate specific characters, individuals, visual styles, or other categories of content without retraining the entire foundation model [17]. Under this service model, the platform provides both the underlying foundation model and the fine-tuning tools, while users upload a relatively small amount of task-specific data to customize the foundation model for their particular needs. Through this process, the model is capable of consistently generating outputs that conform to users' intended characteristics and content preferences.

By contrast, under this service model, the degree of control exercised by service providers over the content-generation process is comparatively attenuated and partially transferred to users. Unlike the two service models discussed above, fine-tuned models are customized to meet the specific needs of individual users, and the style, subject matter, and output characteristics of the generated content are largely determined by users' choices and inputs. By uploading training data and configuring fine-tuning parameters, users actively participate in shaping the model's behavior, thereby substantially limiting the service provider's ability to control the outputs of the customized model. Accordingly, the AI model fine-tuning technology provided under this service model is, in itself, relatively neutral, and the service provider's degree of substantive control over specific customized outputs is significantly weaker.

In summary, risk-control theory provides the theoretical foundation for the differentiated allocation of content governance obligations among GAI service providers. The degree of substantive control exercised by service providers over the content-generation process varies significantly across different service models and should therefore serve as the principal criterion for allocating legal obligations. In pure prompt-based generation services, providers exercise comprehensive control over model training, deployment, and safety alignment, although their ability to foresee individual outputs remains limited. In RAG services, providers additionally control the selection and management of external information sources, giving rise to corresponding content governance obligations with respect to retrieval mechanisms and source verification. By contrast, in foundation model fine-tuning services, users play a substantial role in customizing model behavior, resulting in a significant dilution of the service provider's control over generated content. Consequently, the content governance obligations imposed on service providers should be correspondingly recalibrated.

5. PATHWAYS FOR THE DIFFERENTIATED ALLOCATION OF CONTENT GOVERNANCE OBLIGATIONS FOR GAI SERVICE PROVIDERS

5.1 Content Governance Obligations for Pure Prompt-Based Generation Services

Under this service model, GAI service providers exercise a relatively high degree of substantive control over the development and operation of the underlying model. The training, fine-tuning, evaluation, deployment, and optimization of the model are all undertaken under the direction of the service provider. Consequently, service providers possess a comparatively high capacity to foresee and control the infringement risks that may arise from the operation of the model. Compared with dispersed end users or upstream data providers, they are therefore better positioned-and correspondingly more responsible-for mitigating such risks through improvements in algorithm design, the refinement of content-filtering mechanisms, and the enhancement of training data quality [18].

Nevertheless, the autonomous nature of GAI systems imposes inherent limits on the service provider's ability to foresee or control every individual output. Given the enormous volume and dynamic nature of AI-generated content, it is neither technically feasible nor reasonably practicable to conduct an ex ante review of every generated output before its dissemination.

Given the foregoing allocation of control, providers of pure prompt-based generation services exercise the most comprehensive control over training data and model architecture at the ex ante stage. They should therefore be subject to comparatively stringent content governance obligations before the deployment of GAI systems.

First, service providers should assume rigorous obligations with respect to the governance of training data. They should take reasonable measures to ensure that training data are obtained from lawful sources, maintain appropriate standards of data quality, and take reasonable measures to prevent training datasets from containing unlawful or harmful content. In addition, service providers should

bear obligations relating to model safety assessment and alignment. Before a model is deployed, they should conduct comprehensive safety evaluations and implement appropriate alignment measures-including reinforcement learning from human feedback (RLHF), system prompt design, and other safety-oriented techniques-to mitigate the risk of systematically generating unlawful content at the earliest stage of the content-generation process.

During the content-generation stage, in addition to obligations relating to content labeling and the provision of prominent notices, service providers should assume appropriate duties of content safety review and filtering. In particular, they should deploy effective technical measures, including intelligent filtering systems with adaptive learning capabilities, and establish a graduated response mechanism based on feature identification, behavioral analysis, and risk classification. Furthermore, through the use of digital watermarking and metadata traceability technologies, service providers should ensure that content governance measures can be implemented with sufficient precision at the level of individual data units where necessary [19]. Nevertheless, given the inherent limits of their substantive control over individual outputs, these obligations should be understood as duties to implement effective and state-of-the-art technical measures, rather than as obligations to guarantee the legality or factual accuracy of every item of AI-generated content. Accordingly, the applicable standard should remain one of reasonable technological diligence, rather than one of strict or outcome-based liability.

At the ex post stage, service providers should establish effective complaint and reporting mechanisms. Upon receiving infringement notices from rights holders, platforms should adopt appropriate technical measures to prevent the repeated generation of substantially similar infringing content. Furthermore, for content that manifestly violates public order and morality or presents a high likelihood of infringement, service providers should implement enhanced preventive measures, including advance risk warnings and proactive interception mechanisms [20].

5.2 Content Governance Obligations for Retrieval-Augmented Generation (RAG) Services

In RAG-based services, providers exercise a certain degree of control over information sources by determining the scope of external knowledge bases, configuring retrieval strategies, and adjusting the ranking and filtering mechanisms for retrieved information. However, when the external information retrieved by the system contains errors or biases, the content generated by the model based on such information may likewise be inaccurate. Given these distinctive characteristics of control, the content governance obligations imposed on RAG service providers should adopt a differentiated approach from those applicable to pure prompt-based generation services. The duty of care imposed on GAI service providers should not be understood simply as an exemption from liability for generated outcomes. Rather, such duties should evolve toward more proactive forms of risk identification, continuous monitoring of generated content, and the provision of explanations regarding potential limitations of outputs. In the context of RAG-based services, these obligations are primarily reflected in the governance of user behavior, control over generated outputs, disclosure and labeling of information sources, and ethical review mechanisms [21]. At the ex ante stage, service providers should establish a technical framework for value alignment by translating fundamental societal values-such as fairness, non-discrimination, and transparency-into operational and verifiable technical parameters. In doing so, they should ensure that the underlying logic of AI systems remains aligned with broadly accepted social values and ethical principles throughout the content-generation process [22]. Service providers should also undertake obligations to verify retrieved information. Such verification should not require providers to examine the authenticity of every individual piece of information contained in external sources. Rather, providers should conduct an overall security assessment and review of the knowledge bases and external information sources incorporated into the retrieval scope, so as to prevent unlawful or harmful sources of information from being introduced into the generation process and causing contamination of retrieved information.

At the generation stage, service providers should disclose and label the sources of information underlying generated outputs, thereby ensuring the traceability of the factual basis upon which such outputs are generated.

At the ex post stage, where generated content is found to contain errors, service providers should bear reasonable obligations of correction and response. Where particular information sources are confirmed to contain erroneous content, providers should, based on their ability to trace the origin of such information, take appropriate measures to correct the generated content or remove, restrict, or block access to the problematic sources within the knowledge base.

5.3 Content Governance Obligations for Foundation Model Fine-Tuning Services

The foundation model layer of GAI systems possesses a dual character. On the one hand, as a foundational infrastructure underlying a wide range of AI applications, it performs a public-oriented function due to its broad social impact and widespread accessibility. On the other hand, by providing access to downstream developers and users who build applications upon foundation models, it exercises a quasi-regulatory governance capacity over the use and adaptation of these models [23]. Based on the foregoing analysis, in foundation model fine-tuning services, the service provider's control over the content-generation process is significantly diluted and partially transferred to users. Accordingly, the content governance obligations imposed on such service providers should be differentiated by taking into account the substantial participation of users in model customization and the resulting redistribution of control over generated content.

At the ex ante stage, service providers should ensure that the foundation models they provide do not inherently possess the capacity to systematically generate unlawful or harmful content. Accordingly, they should undertake obligations to implement necessary safety measures at the foundation model level. Specifically, service providers should fulfill data governance obligations by ensuring the professionalism, reliability, and accuracy of training datasets. They should also undertake obligations relating to alignment-oriented fine-tuning by applying domain-specific reinforcement learning and other alignment techniques to ensure that model outputs conform to relevant industry standards and professional knowledge requirements [24]. Furthermore, service providers should establish comprehensive service agreements with users to clarify the respective rights and obligations of both parties. In particular, such agreements should specify prohibited categories of conduct, including the intentional generation of unlawful or misleading content, infringement of third-party intellectual property rights, and dissemination of unlawful content, while establishing corresponding contractual liabilities for violations [25]. At the content-generation stage, one possible governance approach is to establish a dynamic risk protocol mechanism, under which developers and deployers may predetermine risk warning thresholds and conditions for the allocation or reallocation of responsibilities. When model iterations result in changes to the distribution of control over the content-generation process, the parties may reassess and adjust their respective obligations and responsibilities based on the actual circumstances [26]. At the ex post stage, when service providers become aware that a model has been used to generate infringing or unlawful content, they should take appropriate remedial measures, such as removing the relevant content or restricting access to the model or generated outputs.

6. CONCLUSION

Existing legal frameworks generally adopt a uniform approach to allocating content governance obligations among GAI service providers. However, significant differences exist among GAI services in terms of their technical architectures, application contexts, and capacities to control the content-generation process, resulting in regulatory outcomes that may fall short of intended objectives. Risk-control theory provides a compelling theoretical foundation for the differentiated allocation of such

obligations and helps avoid the difficulties arising from the direct application of generalized duty-of-care standards to the emerging and technologically complex field of GAI.

Based on differences in service providers' substantive control over the content-generation process, differentiated obligations should be allocated among different categories of GAI services. In pure prompt-based generation services, providers exercise comprehensive control over model architecture, training data, and safety alignment, and should therefore undertake relatively stringent content governance obligations throughout the generation lifecycle. In retrieval-augmented generation (RAG) services, the focus of providers' obligations should shift toward the overall security assessment of information sources, ensuring the traceability of generated content, and correcting or restricting access to erroneous information. In foundation model fine-tuning services, users participate extensively in the customization of models, significantly reducing the provider's control over specific generated outputs. Accordingly, providers' obligations should be correspondingly adjusted, focusing primarily on ensuring that the foundation model itself does not possess the capacity to systematically generate unlawful content and on defining prohibited uses through comprehensive service agreements.

Only by moving beyond a one-size-fits-all regulatory approach toward a tiered and differentiated framework can legal regulation achieve a more refined balance between safeguarding the security of AI-generated content and fostering artificial intelligence innovation, thereby enabling a mutually reinforcing relationship between technological development and legal governance.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Wu, H. D. (2025). On the reasonable duty of care of generative AI service providers. *Intellectual Property*, (11), 3–23. [In Chinese.]

[2] Wu, Y. R. (2025). The liability allocation mechanism for generative AI service providers. *Academic Research*, (10), 60–68. [In Chinese.]

[3] Liu, X. L. (2026). Legal clarification and path optimization of the obligation to label AI-generated content. *Beijing Social Sciences*, (1), 105–115. [In Chinese.]

[4] Cheng, X. (2026). Duty of care, statutory duty, and fault of generative AI service providers. *Political Science and Law Tribune*, (4), 35–47. [In Chinese.]

[5] Zhang, W. R., & Liu, Y. Z. (2026). Hierarchical construction of imputation principles for generative AI service providers. *Hebei Law Science*, (7), 178–194. [In Chinese.]

[6] Guo, M. R. (2016). The impact of dangerous liability on the development of tort liability law. *Journal of Yantai University (Philosophy and Social Sciences Edition)*, (1), 24–29. [In Chinese.]

[7] Liu, C. W., & Ma, R. C. (2025). Normative sources of the duty of care for generative AI users. *Journal of Shandong University (Philosophy and Social Sciences Edition)*, (5), 152–163. [In Chinese.]

[8] Zhang, X. B., & Bian, L. (2025). On the presumption of fault liability of generative AI service providers. *Northern Legal Science*, (5), 5–19. [In Chinese.]

[9] Liu, Y. H. (2026). Research on hierarchical allocation of copyright infringement liability for generative AI. *Law Journal*, (3), 67–87. [In Chinese.]

[10] He, W. (2024). *Generative AI AIGC and multimodal technology application practice guide*. China Science and Technology Press. [In Chinese.]

[11] He, Y. P. (2026). On the liability of model providers for copyright infringement of AI-generated content. *Journal of Soochow University (Law Edition)*, (1), 51–63. [In Chinese.]

[12] Li, Z. H. M. (2026). Secondary hallucination risks and regulation of retrieval-augmented generation technology in judicial applications. *Journal of University of Electronic Science and Technology of China (Social Sciences Edition)*. Advance online publication. https://doi.org/10.14071/j.1008-8105(2026)-3013. [In Chinese.]

[13] Tian, Y. L., Wang, Y. T., Wang, X. X., Yang, J., Shen, T. Y., Wang, J. G., Fan, L. L., Guo, C., Wang, S. W., Zhao, Y. W., Wan, S., & Wang, F. Y. (2025). From RAG to SAGE: Current status and prospects. *Acta Automatica Sinica*, (6), 1145–1169. [In Chinese.]

[14] Wang, X. L., Li, Y., Ma, C. F., & Li, S. (2026). Retrieval-augmented generation (RAG): A survey of methods and applications. *Computer Science*, (7), 101–117. [In Chinese.]

[15] Tang, A. D., & Lin, Z. C. (2026). A survey on parameter-efficient fine-tuning methods for large models: Techniques, trends, and challenges. *Acta Automatica Sinica*. Advance online publication. https://doi.org/10.16383/j.aas.c250451. [In Chinese.]

[16] Qin, D. H., Li, Z. T., Bai, F. B., Dong, L. K., Zhang, H., & Xu, C. (2025). A survey on parameter-efficient fine-tuning techniques for large language models. *Computer Engineering and Applications*, (16), 38–63. [In Chinese.]

[17] Hu, H. F., & Zhou, Y. (2024). Multi-subject personalized image generation method based on LoRA. *Chinese Journal of Stereology and Image Analysis*, (4), 320–332. [In Chinese.]

[18] Ning, D. X., & Wei, T. T. (2026). Hierarchical determination of indirect infringement and liability allocation of generative AI service providers: Also on the imputation dilemma and response of indirect infringement of traditional network service providers. *Social Sciences in Inner Mongolia*, (1), 96–103. [In Chinese.]

[19] Du, M. J. (2025). Systematic deconstruction and institutional reconstruction of tort liability of generative AI service providers. *Journal of Xihua University (Philosophy and Social Sciences Edition)*, (4), 25–39. [In Chinese.]

[20] Wang, L. M., & Lin, T. L. (2026). Judicial regulation of generative AI: Risks, dilemmas, and responses. *Fujian Tribune (Humanities and Social Sciences Edition)*, (4), 143–154. [In Chinese.]

[21] Zhang, H. T. (2025). Imputation and composition of generative AI torts under the dynamic system theory. *Law Journal*, (6), 82–101. [In Chinese.]

[22] Li, D. (2023). On the determination of tort liability for algorithmic discrimination against consumers: An empirical investigation based on judicial adjudication. *Contemporary Law Review*, (6), 75–85. [In Chinese.]

[23] Zhang, L. H. (2023). The legal positioning and hierarchical governance of generative AI. *Modern Law Science*, (4), 126–141. [In Chinese.]

[24] Deng, W. (2026). Research on the regulatory path of generative AI hallucination infringement. *Political Science and Law*, (6), 134–150. [In Chinese.]

[25] Li, D. (2026). Phased duty of care and tort liability of generative AI service providers. *Modern Law Science*, (2), 224–242. [In Chinese.]

[26] Wang, Q. S. (2025). Attribution paradigm and liability mechanism for damages caused by large model operations. *Oriental Law*, (4), 124–137. [In Chinese.]