DOI: 10.53469/jssh.2026.8(01).05

China's Labeling Regime for AI-Generated Misinformation: Implementation Challenges and Reform Priorities

Zihan Lu

Ocean University of China, Qingdao, Shandong, China

Abstract: ***While generative artificial intelligence (generative AI) is reshaping digital content production, it also amplifies the risks and governance challenges associated with misinformation. China's current approach centers on a labeling regime that establishes a full-chain responsibility framework across content generation, dissemination, distribution, and use, and that adopts a dual-track system of explicit and implicit labels. However, implementation challenges persist. At the level of governance subjects, key stakeholders often lack sufficient incentives to comply, and the allocation of responsibilities among actors remains unclear. Technically, labels are vulnerable to removal or tampering, undermining traceability and enforcement, while fragmented standards limit cross-border interoperability. In terms of regulatory efficacy, labeling indicates the mode of content production but does not resolve substantive questions of authenticity; moreover, generalized labeling may induce labeling fatigue and weaken the warning function over time. To address these problems, China should clarify labeling duties across the content lifecycle, strengthen anti-tampering and detection technologies, promote interoperability with international standards, and adopt a tiered and risk-based labeling framework to mitigate labeling fatigue. These reforms can support a healthier, more orderly, and sustainable digital information environment.***

Keywords: Generative artificial intelligence, Labeling regime, Misinformation governance.

1. Framing the Problem

Today, artificial intelligence (AI) technologies are evolving rapidly. Large models such as ChatGPT, DeepSeek, and Sora have accelerated the diffusion of generative AI. These developments can boost productivity, drive socio-economic growth, and enhance everyday convenience [1]. As exemplified by ChatGPT, generative AI models are trained on massive text datasets, enabling them to generate fluent text and perform language tasks at scale, and they have been widely applied in fields such as news writing, artistic creation, and scientific research. While generative AI makes creation easier, the high fidelity and scalability of AI-generated content (AIGC) can erode trust online, increasing low-quality content and degrading information ecosystems [2]. These risks can affect public trust, social and economic activity, threaten individual rights, and challenge current legal and regulatory frameworks. Therefore, AIGC labeling has emerged as a foundational governance instrument. It protects users' right to know and improves transparency in online information ecosystems. At the same time, it provides a governance tool to encourage socially beneficial applications and responsible development of the AI industry.

In response, China has issued a series of framework regulations, including the Interim Measures for the Administration of Generative Artificial Intelligence Services and the Provisions on the Administration of Deep Synthesis of Internet-Based Information Services. In March 2025, the Measures for the Labeling of Artificial Intelligence Generated and Synthesized Content (the Labeling Measures) were promulgated, preliminarily establishing a regulatory framework for AIGC labeling. These Measures have made meaningful progress in establishing mandatory labeling duties, improving the governance of content dissemination and distribution, and seeking a balance between regulatory oversight and technological innovation. Yet practice has exposed several gaps. First, the allocation of labeling duties among regulated actors remains overlapping and unclear, which weakens incentives for compliance. Second, current technical capabilities are insufficient to operationalize the dual-track labeling model required by the Labeling Measures, and relevant standards remain fragmented. Third, although AIGC labeling can help the public and regulators identify AI-generated content and trace its provenance, it does not curb the production of misinformation at the source [3]. Accordingly, China's current framework still lacks a differentiated labeling scheme; a one-size-fits-all approach may dilute the informational value of labels and contribute to label fatigue, ultimately weakening their warning function.

Overall, the governance of AI-generated misinformation has reached a stage where it must move beyond the preliminary question of whether labeling is necessary toward the more practical challenge of how to label in a scientific and effective manner. Accordingly, this article examines the practical dilemmas arising from the labeling requirements under the existing Labeling Measures and seeks to incorporate a tiered and classified methodological approach into the design of AIGC labeling rules. The objective is to reconcile competing governance priorities and to develop a labeling regime for AI-generated misinformation that is tailored to China's regulatory environment and developmental stage.

2. Governance of AI-Generated Misinformation under the Labeling Regime

The Labeling Measures pursue traceability for AI-generated content through mandatory disclosure, without unduly interfering with ordinary content creation and dissemination. They allocate labeling duties across multiple actors and establish a dual-track system of explicit and implicit labels, enabling end-to-end governance from generation to use.

2.1 Labeling Subjects: Generative and Synthetic Service

Providers, Internet Application Distribution Platforms, Network Information Content Dissemination Service Providers, and Users

The Labeling Measures establish a responsibility framework for labeling that spans the lifecycle of AI-generated content — from generation to end use—by assigning differentiated duties to distinct actors.

First, Articles 4 and 5 impose source-level labeling obligations on generative and synthetic service providers (generative service providers), shifting responsibility for the legality of AI-generated content upstream to the point of production. This allocation reflects a risk-attribution rationale: those who create risks should bear the costs of controlling them. Generative service providers are also required to ensure label persistence. Where functions such as downloading, copying, or exporting are available, they must ensure that labels remain attached as files circulate, thereby extending supervision across the content's lifecycle.

Second, internet application distribution platforms (app distribution platforms) act as digital gatekeepers. As the primary channel through which generative-AI applications enter the app ecosystem, they bear ex ante review responsibilities. At the app-listing stage, app distribution platforms are required to verify whether an application has the capability to generate and maintain compliant labels, thereby preventing non-compliant AI applications from entering the market.

Third, network information content dissemination service providers (dissemination platforms) serve as key nodes in the dissemination chain and are responsible for ongoing verification and differentiated labeling. Article 6 requires dissemination platforms to verify and detect AI-generated content, apply appropriate labels, and provide user notifications. Depending on the level of certainty, platforms must use differentiated labels such as "generated," "possibly generated," or "suspected AI-generated." In this sense, dissemination platforms operate as an additional line of defense by identifying and intercepting unlabeled AI-generated content during dissemination.

Finally, users, as both the ultimate publishers and recipients of content, bear terminal responsibilities. Article 10 requires users who obtain content through generative service providers to comply with labeling provisions in service agreements and to refrain from intentionally removing or tampering with labels. In addition, when users publish content that contains generative or synthetic elements on online platforms, they must disclose that fact and apply labels using the tools provided by the platform. Accordingly, even if upstream service providers fail to attach labels, users remain subject to a duty of honest disclosure. This rule clarifies users' responsibilities in maintaining information integrity and prevents users from becoming a loophole through which labeling requirements can be bypassed.

2.2 Labeling Methods: The Dual-Track System of Explicit and Implicit Labels

Article 3 establishes a dual-track labeling model combining explicit and implicit labels. Explicit labels are prominent, user-visible notices—such as text, symbols, or watermarks — that indicate the content is AI-generated. The main purpose of explicit labeling is to protect the audience's right to know. With generative AI producing increasingly human-like content, the risk of confusion and deception has risen. Explicit labeling requires generative service providers to display an "AI-generated" notice on or alongside AI-generated content — for example, in the opening frames of a video, in a conspicuous area of an image, or next to a text post. By making the content's origin transparent at the point of exposure, this disclosure duty shifts part of the identification burden from individual users to service providers and reduces the overall social costs of verification. Explicit labels also perform a warning function by prompting users to scrutinize labeled content more carefully. In this way, they help preserve the informational conditions for meaningful public deliberation in the online environment.

Unlike explicit labeling, which targets users, implicit labeling is primarily machine-readable. It embeds provenance information in metadata or imperceptible watermarks to support traceability. Article 5 of the Labeling Measures requires generative service providers to include key fields in file metadata, including attribute information, provider identifiers, and content numbers. Implicit labels matter because visible labels can be easily removed during dissemination—for example, by cropping, masking, or overlaying. By contrast, implicit labels are embedded at the file or signal level and are intended to be more resistant to routine transformations. Even after repeated forwarding or compression, they can help trace content back to its origin and provide technical evidence for ex post accountability and enforcement against online rumors. Scale makes manual checks unrealistic. Given the volume of content circulating on dissemination platforms, Article 6 therefore requires platforms to verify implicit labels. Automated verification enables large-scale screening and interception of unlabeled or suspicious AI-generated content, forming a key foundation for automated misinformation governance.

Explicit and implicit labels are designed to work together. Explicit labels provide a clear, user-facing disclosure and warning cue, while implicit labels enable traceability and support enforcement by preserving machine-readable provenance information. This combination reduces the risk that labeling information is lost as content moves across platforms. Even if a visible label is removed or obscured, implicit labels can still be detected and used to trigger automated checks or early-warning processes on dissemination platforms. The dual-track model also seeks to balance regulatory objectives with user experience. It reconciles governance objectives with practical usability. Article 9 of the Labeling Measures allows, in specific scenarios such as professional creative contexts, the omission of explicit labels through contractual arrangements, while not exempting service providers from the obligation to add implicit labels. This design balances users' aesthetic and creative preferences with the minimum requirements of social information security.

In sum, the Labeling Measures establish an end-to-end labeling framework. By allocating duties across actors at

different stages and combining explicit and implicit labels, the regime strengthens both the effectiveness and the precision of governance for AI-generated content.

3. Implementation Challenges in Governing AI-Generated Misinformation under the Labeling Regime

Despite an end-to-end framework and a dual-track design, the Labeling Measures face persistent implementation gaps: low compliance incentives, fragile labels that can be stripped or forged, and limited efficacy—labels indicate AI involvement rather than truthfulness and may contribute to labeling fatigue.

3.1 Actor-Level Dilemmas: Willingness to Label and the Delineation of Responsibilities

3.1.1 Negative Externalities of Information Disclosure and Insufficient Incentives to Label

One major practical challenge of the labeling regime is the limited willingness of key actors to comply proactively. While labeling generates public benefits for misinformation governance, such as reducing audiences' identification costs, enhancing traceability, and mitigating the spillover of misinformation, the associated costs are borne privately. These costs include technical investment, human review, dispute resolution arising from misclassification, losses in user experience, and traffic loss or revenue decline. The resulting imbalance creates a classic negative-externality problem and a structural tendency toward underinvestment in labeling. Similar incentive patterns have been observed in disclosure regimes in other domains. In online advertising, for example, key opinion leaders may use ambiguous disclosures to preserve perceived authenticity and maintain organic engagement [4]. In the AIGC context, users especially content creators have similar concerns. Once content is labeled as AI-generated, audiences may automatically downgrade its quality, artistic value, or credibility. This phenomenon, often described as "machine heuristic bias" [5], leads creators to refuse labeling to avoid economic or reputational losses. Moreover, as human–AI co-creation has become widespread, users often perceive their cognitive inputs such as prompt engineering as constituting substantive creative labor, which strengthens their sense of "psychological ownership". Subjectively, they are inclined to regard the resulting artwork or text as written by me rather than written by AI [6]. As a result, they may resist attaching an AI-generated label, viewing it as diminishing authorship and intellectual contribution. Dissemination platforms face parallel disincentives: conspicuous labels and frequent pop-ups can disrupt user experience and reduce engagement, while imperfect detection increases the risk of misclassification and the associated complaint- and dispute-handling costs. When these compliance burdens outweigh the expected benefits, platforms tend to adopt defensive and minimalist labeling strategies [7].

3.1.2 Overlapping Duties and Ambiguous Boundaries

First, overlapping duties across the content lifecycle create practical difficulties. Under the Labeling Measures, generative service providers must attach source-level labels, while dissemination platforms are required to verify labels, add warning labels where appropriate, and append dissemination-related metadata. This structure presumes a downstream supervisory role for dissemination platforms. In practice, however, the two roles can converge: some providers simultaneously generate content and operate dissemination services. In such cases, duplicative obligations may lead to inefficient compliance and, more importantly, create a risk of self-verification, where a platform is effectively asked to verify its own outputs, weakening the intended oversight function. For instance, in a copyright infringement and unfair competition dispute adjudicated by the Hangzhou Internet Court, the court observed that generative AI service providers differ from traditional content providers, hosting service providers, and search-link service providers, and may simultaneously assume the dual roles of content producer and platform manager. Where such role convergence occurs, imposing duplicative labeling obligations on a single service provider may result in resource waste. Moreover, the dual identity may give rise to a form of "self-verification" whereby a service provider acting as a dissemination platform verifies its own generated content, thereby undermining the intended supervisory function of dissemination platforms [8].

Second, horizontal boundary ambiguity further complicates responsibility allocation. The rapid diversification of generative services—from standalone apps to embedded functions, and from on-device deployment to cloud-based APIs—makes it increasingly difficult to identify the appropriate duty bearer [9]. According to the Coase Theorem, unclear delineation of rights leads to sharply increased transaction costs. In API-based operational models, for example, risks may originate from foundational models, intermediary layers, or application layers, complicating risk identification and attribution [10]. When applications integrate third-party large-model APIs, it remains unclear whether labeling responsibilities should rest with the API provider or the application developer. In addition, the Labeling Measures focus primarily on fully AI-generated content, whereas much real-world output is human–AI hybrid. Without clearer standards for hybrid creation, responsibility for labeling remains contested, further widening loopholes and uncertainty.

3.2 Technical Dilemmas: Detection Accuracy, Label Persistence, and Fragmented Standards

3.2.1 Technical Limits Undermine Detection Accuracy

Detection and verification tools are inherently imperfect and cannot guarantee accuracy. For dissemination platforms, comprehensive detection of AI-generated content is costly and prone to false positives. As AI technologies iterate rapidly, AI-created content has become increasingly difficult to distinguish from human-created content, further amplifying identification challenges. Existing AI detection tools are neither accurate nor reliably robust. According to an OpenAI report released in 2023, the company's AI Text Classifier achieved an accuracy rate of only 26%, while falsely classifying approximately 9% of human-written texts as AI-generated; was later discontinued [11]. Related research likewise observes that AI-content detection tools perform

inconsistently across models and contexts, reinforcing concerns about reliability.

Against this backdrop, the Labeling Measures require dissemination platforms to carry out verification duties at scale. In practice, the combination of technical uncertainty and compliance pressure may push platforms toward over-inclusive labeling to reduce regulatory risk. This, however, increases the likelihood that human-created content will be mislabeled as AI-generated, potentially harming creators' rights and interests.

3.2.2 Technology Cannot Ensure Label Persistence

Explicit labels can be removed with simple edits, such as cropping, masking, or overlaying. The Labeling Measures rely on implicit labels to support traceability, yet current watermarking and metadata-based techniques remain vulnerable [12]. Current robustness techniques still lag behind generation technologies. Robustness often lags behind generation and editing capabilities, and routine transformations—such as recompression, resizing, filtering, or screen recording—may degrade or erase embedded signals. Adversarial manipulation can also target metadata structures, including forging or altering key fields, which undermines attribution and traceability [13]. Once implicit labels are stripped during dissemination, the verification obligation imposed on dissemination platforms by Article 6 of the Labeling Measures becomes practically unenforceable. In other words, while the law presupposes that labels should exist, the technical realities of dissemination make labels prone to disappearance. This tension between normative efficacy and technical capacity renders the labeling regime difficult to implement in practice.

3.2.3 Fragmentation of Labeling Standards

AIGC dissemination is inherently cross-border in nature, yet the AIGC labeling regime established by the Labeling Measures and its supporting technical standards faces constraints in global applicability. First, the Labeling Measures and the supporting standard GB 45438-2025 adopt a specific JSON-structured metadata format, whereas internationally the more widely used framework is the C2PA standard. The two differ significantly in structural design, required fields, and verification mechanisms. These discrepancies may lead to incompatibility, duplicative labeling, and user confusion.

In addition, the final version of the Labeling Measures removed the clause in the draft-for-comments stating that the Measures would not apply where services were not provided to domestic users, signaling an intention by regulators to expand the scope of application and regulatory boundaries. However, legislative and regulatory approaches to the governance of overseas services remain underdeveloped, leaving a normative gap that is ill-suited to addressing the challenges posed by cross-border flows of AIGC content.

3.3 Efficacy Dilemmas: Signal Failure and Labeling Fatigue

3.3.1 Signal Failure: Formal Disclosure Cannot Reliably Block Substantive Risks

The labeling regime is designed to correct information asymmetries through mandatory disclosure; it does not, by its nature, resolve questions of content authenticity. Legislators, operating on a rational-actor premise, often assume that once audiences are conspicuously informed that content is AI-generated, they will activate deliberative judgment and thereby curb the dissemination of misinformation. Empirical research, however, suggests that merely providing an "AI-generated" label does not significantly change audiences' trust in the content or their willingness to share it [14]. As a result, this form of minimal, formal disclosure is unlikely to achieve meaningful risk interruption. A simulated social media experiment in Canada similarly indicates that adding an "AI-generated" label to a content page has little to no effect on users' trust or sharing behavior. Only a full-screen warning—one that requires users to manually dismiss it before continuing—substantially reduces exposure and sharing intentions. Yet under prevailing commercial logics, mainstream platforms are unlikely to adopt such aggressive labeling designs due to concerns about user experience and platform operations [15]. Another experiment focusing on AI-generated news headlines also finds that the effect of an "AI-generated" label in reducing user trust is only about one-third as strong as the effect of a "false" label [16]. This suggests that, rather than relying on a technology-origin label to discount credibility, governance strategies may be more effective if they directly verify the truthfulness of AI-generated content and explicitly label misinformation as such.

3.3.2 Labeling Fatigue

In a high-volume information environment, a uniform and mandatory labeling regime may induce "labeling fatigue" [17]. Labeling fatigue refers to a systematic negative shift—driven by information overload and cognitive confusion—in how users decode label information, form affective responses, and ultimately behave in response to labels [18]. As AI-related labels proliferate across the internet, audiences' sensitivity to such signals may gradually decline; labels may eventually be treated as background noise akin to cookie-consent pop-ups and thus be ignored [19]. Moreover, persistent reminders to be vigilant against deepfakes may foster a generalized skepticism — an attitude that anything could be fake—in which the public becomes inclined to doubt both true and false content. This is often described as the "liar's dividend" effect [20], whereby politicians and public figures exploit an environment of misinformation and distrust to more credibly deny authentic evidence by falsely claiming that genuine information about them is fabricated [21]. This dynamic is likely to degrade the broader information environment. When standards of truth and falsity become blurred, the value of truth itself is weakened. Accordingly, the implementation of labeling rules must guard against a "cry wolf" effect. Overuse and overgeneralization of labeling not only fail to enhance warnings, but may instead habituate the public and produce cognitive numbness, ultimately leading to institutional ineffectiveness.

4. Reform Priorities for Improving Governance of AI-Generated Misinformation under the Labeling Regime

To improve governance of AI-generated misinformation, the AIGC labeling regime should: (i) clarify labeling responsibilities across the content lifecycle and strengthen compliance incentives; (ii) reinforce research and development of labeling technologies and promote the international interoperability of labeling standards; and (iii) establish a tiered and categorized labeling system to mitigate labeling fatigue, thereby fostering a healthy, orderly, and sustainable digital content ecosystem.

4.1 Reconstructing Responsibilities: Incentives and Clear Boundaries

First, strengthen compliance incentives. Although compliance may raise platforms' short-term costs, it can strengthen long-term governance capacity and reputational standing. Likewise, more intensive public enforcement entails higher regulatory costs but can improve deterrence and bolster institutional credibility. Incentives should therefore be calibrated through legal and policy instruments. On the one hand, sanctions for non-compliance should be strengthened to raise the expected cost of violations, ensuring that actors who opportunistically evade labeling bear commensurate consequences. For willful misconduct such as intentionally removing watermarks or forging provenance labels, policymakers may consider remedies beyond ordinary compensation, including punitive damages where available, to ensure that the expected cost of misconduct exceeds any potential gain. In addition, records of malicious violations may be incorporated into the social credit system and industry blacklists, with restrictions on market access used to enhance deterrence. On the other hand, positive incentives for compliant labeling should also be considered. For example, platforms may provide traffic support or other benefits to users who promptly and accurately label AI-generated content.

Second, the framework should clarify labeling duties to reduce overlap and boundary ambiguity. Where generative service providers also function as dissemination platforms, the analysis should begin by disentangling and comparing the statutory obligations attached to each role. Under the Labeling Measures, dissemination platforms have three sets of obligations: (i) verification, checking whether content carries implicit labels in its metadata and whether the uploader has disclosed AI involvement; (ii) labeling, applying differentiated labels such as "generated," "possibly generated," and "suspected generated," based on verification results; and (iii) function-provision obligation, providing labeling tools and reminding users to label. By contrast, the obligations of generative service providers are confined to labeling content generated by their own services. They are not required to verify downstream content or assign differentiated labels, and their labeling focuses on production-side provenance rather than dissemination-side information. Where a provider serves both as a generative service provider and as a dissemination platform, duplicative compliance with overlapping duties can waste resources and reduce efficiency. In such cases, overlapping obligations should be satisfied once, rather than performed twice. To prevent lax compliance and facilitate administrative oversight, rules may require dual-role providers to file a record or make a declaration confirming that overlapping obligations have been fulfilled, so that competent authorities can verify compliance. For non-overlapping obligations—such as verification duties, categorical labeling, and the addition of dissemination-related elements—service providers should remain legally required to perform them in accordance with the law. In cases where third-party models are embedded in applications, legislation or regulatory guidance should clarify the boundary of responsibilities between API providers and application developers. For example, foundational model providers may be required to embed implicit watermarks by default in output content, while front-end application developers should be responsible for providing conspicuous user-interface notices, thereby achieving a reasonable upstream–downstream allocation of labeling obligations. With respect to the identification and labeling of human–AI hybrid creations, competent authorities should develop more granular standards and distinguish, at minimum, among three primary categories — purely AI-generated content, human–AI co-created content, and content suspected of being generated or synthesized—so as to apply differentiated labeling obligations. Such differentiation would not only enhance regulatory precision, but also better balance technological development with risk prevention [22].

4.2 Technology Enablement: Robust Labeling, Reliable Detection, and Interoperable Standards

Effective labeling depends on technical capacity. Current systems face two practical gaps: provenance signals can be stripped or forged, and detection remains unreliable in real-world settings. Investment should therefore prioritize (i) more tamper-resistant implicit labels that remain recoverable after common transformations such as compression, cropping, and re-encoding, and (ii) more reliable detection methods that generalize across models and contexts. Research and industry can improve watermark robustness against adversarial manipulation and develop cross-modal techniques that work consistently across text, images, audio, and video. Complementary approaches, such as cryptographic signing and secure provenance registries, may further raise the cost of forgery and manipulation. Detection capacity is equally important for dissemination platforms, which must verify labels at scale. Drawing on California's legislative experience, regulators could impose higher technical requirements on large-scale AIGC providers—for example, requiring them to offer free, publicly accessible AI-content detection tools that both users and platforms can use to identify AI-generated content. Regulators could also lead the development of shared detection services or model repositories that small and medium-sized platforms can access through standardized interfaces, thereby reducing duplicative development costs. As more AI services embed provenance markers by default and detection tools become widely available, unlabeled AI-generated content will have far fewer opportunities to evade identification.

In addition, international mutual recognition and interoperability of labeling standards should be promoted. Given that major global technology companies have jointly

advanced the C2PA content provenance standard, China should participate more actively in international standard governance. While safeguarding national data security, China should improve the compatibility of its domestic labeling specifications with standards such as C2PA. Measures could include reserving space in domestic metadata formats for field mapping and conversion with C2PA, or developing bridging tools to translate provenance information across standards. China could also pilot cooperation with major cross-border content platforms to test how C2PA credentials are recognized when international content enters the domestic ecosystem and to adjust for standard differences in a timely manner. Finally, continued engagement in relevant international standard-setting organizations and fora would allow China to share governance experience, absorb global best practices, and gradually facilitate the convergence of global rules for labeling AI-generated content.

4.3 Mechanism Optimization: Constructing a Tiered and Categorized Labeling System

Given the diminishing signaling value of labels and the risk of audience fatigue, labeling should move from a one-size-fits-all model to a tiered, risk-based design. A practical approach is to calibrate labeling intensity to the risk level of the content.

From the perspective of content risk, AIGC may be categorized into high-risk, medium-risk, and low-risk tiers. High-risk content includes domains with substantial implications for social stability and national security, such as political elections, public security, and deepfakes. For such content, the strictest labeling measures should apply: conspicuous explicit labels should be placed in prominent positions and reinforced by embedded implicit watermarks, and, where necessary, supplemented by full-screen warnings or pop-up alerts. Medium-risk content involves areas related to the public interest but with manageable risks—such as commerce, finance, education, and healthcare—and may adopt "explicit + implicit" dual labeling while avoiding excessive disruption to user experience (for example, by using less intrusive icons or softer visual prompts for explicit labels). Low-risk content—such as entertainment, artistic creation, and personal life-sharing—may primarily rely on implicit labeling to meet traceability needs, while explicit labeling may be exempted under defined conditions or presented only in a subtle manner, thereby reducing unnecessary interference with user experience.

From the perspective of application scenarios, labeling rules may distinguish among three categories: public dissemination, use within specific groups, and private or personal use. The Labeling Measures already incorporate an incipient form of scenario differentiation—for example, by allowing explicit labeling to be contractually exempted in professional creative contexts—reflecting a context-sensitive regulatory logic. Going forward, scenario-based classifications could be further enriched. For AI content used only within limited or private contexts, the intensity of explicit labeling may be reduced or even rendered non-mandatory; for content disseminated publicly, label conspicuousness and standardization should be enhanced. Moreover, application scenarios may be combined with content-risk tiers to form a "labeling matrix", under which content characterized by "high risk + broad dissemination" would be subject to the highest level of labeling, while low- and medium-risk content or private uses would correspondingly be subject to lower labeling levels. By precisely tailoring labeling rules to different circumstances, such a system would both safeguard the public's right to know and prevent over-labeling, thereby reducing audiences' cognitive burdens.

From the perspective of actor capacity, differences in firm size, technical capability, and market influence should also be taken into account so as to establish differentiated responsibility mechanisms. This approach may draw on legislative experience in California by stratifying AIGC service providers based on scale and influence and assigning differentiated labeling obligations accordingly. The California AI Transparency Act (2024) establishes differentiated thresholds based on user scale. It limits its primary regulatory focus to "Covered Providers" with more than one million monthly active users, requiring them to provide free AI-content detection tools and to embed provenance labels, while exempting small and medium-sized developers that do not meet the threshold. This asymmetrical regulatory model—calibrated to provider scale and market influence—effectively balances the governance benefits of technical regulation against firms' compliance costs.

In China, a similar differentiated scheme could be developed. For ultra-large platforms (e.g., services with hundreds of millions of monthly active users), the highest labeling standards could be required, including dual-track explicit and implicit labeling, provision of publicly accessible detection tools, and periodic submission of labeling-effectiveness assessment reports. For mid-sized platforms, core labeling obligations should be fulfilled, while certain technical details and evaluation frequencies could be moderated. For startups and small enterprises, mechanisms such as "regulatory sandboxes" could provide partial exemptions or guidance-based (rather than strictly mandatory) requirements within an experimental period—protecting innovation while maintaining a baseline level of safety. Through the fine-grained allocation of labeling obligations, such a system would ensure that large platforms do not become regulatory blind spots due to scale, while preventing smaller actors from adopting evasive strategies under excessive compliance burdens, thereby enhancing the overall sustainability and effectiveness of the labeling regime.

Governance of AI-generated misinformation is ultimately a systematic task. Labeling offers a useful but limited tool. By requiring disclosure that content contains AI-generated elements, labeling can reduce information asymmetries and support accountability across production and dissemination. However, as this article has argued, labeling operates as a disclosure mechanism and cannot be equated with a determination of factual accuracy. Its real-world effectiveness also depends on stakeholders' willingness to cooperate, the reliability of technical safeguards, and the complexity of audience psychology and behavior. Addressing these challenges requires action on three fronts: clarifying the allocation of responsibilities among relevant actors at the institutional level, improving the tamper-resistance of labeling and provenance technologies, and adopting a more

fine-grained, tiered, and risk-based labeling framework for AI-generated content. Together, these reforms can strengthen governance capacity while balancing innovation with security and efficiency with fairness.

References

[1] Jin Xin. Take Multiple Measures to Prevent the Misuse of AI Technology [N]. People's Daily, 2025-11-28(016).

[2] Zhang Lu. A Preliminary Study on the Governance and Regulation of General Artificial Intelligence Risks: Issues and Challenges Triggered by ChatGPT [J]. E-Government, 2023, (09): 14-24.

[3] Su Yu. Legal Governance of the Risks of False Information Generated by Large Language Models [J]. Global Law Review, 2025, 47(06):36-52.

[4] Wojdynski, Bartosz W.; Evans, Nathaniel J. Going Native: Effects of Disclosure Position and Language on the Recognition and Evaluation of Online Native Advertising [J]. Journal of Advertising, 2016.

[5] Bellaiche, Lucas, et al. Humans versus AI: Whether and Why We Prefer Human-Created Compared to AI-Created Artwork [J]. Cognitive Research: Principles and Implications, 2023.

[6] Joshi, Nikhita; Vogel, Daniel. Writing with AI Lowers Psychological Ownership, but Longer Prompts Can Help [C]Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). ACM, 2024: Article 407.

[7] Qiu Junping, Zhang Tingyong, Xu Zhongyang. Research on Collaborative Governance Strategies for AIGC Misinformation Based on a Tripartite Evolutionary Game [J/OL]. Library and Information Service, 1-15[2026-01-12].

[8] Du Weina. Review and Positioning of AIGC Labeling Obligations [J/OL]. Journal of Beijing Institute of Technology (Social Sciences Edition), 1-15.

[9] Guo Xiaodong. Practical Dilemmas and Path Optimization of China's AIGC Labeling Regime [J/OL]. Studies in Science of Science, 1-15.

[10] Zhang Xin. Industrial-Chain-Oriented Governance: The Technical Mechanisms and Governance Logic of AI-Generated Content [J]. Administrative Law Review, 2023, (06): 43-60.

[11] Chaka, Chaka. Detecting AI Content in Responses Generated by ChatGPT, YouChat, and Chatsonic: The Case of Five AI Content Detection Tools [J]. Journal of Applied Learning and Teaching, 2023, 6:94.

[12] Li Peiyao; Wang Hongtao; Zhang Han. Secure and Robust Watermarking for AI-generated Images: A Comprehensive Survey [EB/OL]. arXiv preprint arXiv:2405.02704, 2024.

[13] Justyna Lisinska, Watermarking in Images Will Not Solve AI-Generated Content Abuse [EB/OL]. Center for Data Innovation, 2024-8.

[14] Kreps, Sarah; Kriner, Douglas. Labeling AI-Generated Content May Not Change Its Persuasiveness [EB/OL]. Stanford HAI Policy Brief, 2023-12.

[15] Angus Lockhart & Christelle Tessono, Human or AI? Evaluating Labels on AI-Generated Social Media Content, The Dais, Mar. 2025.

[16] Altay, Sacha; Gilardi, Fabrizio. People Are Skeptical of Headlines Labeled as AI-Generated, Even If True or Human-Made, Because They Assume Full AI Automation [J]. PNAS Nexus, 2024.

[17] Epstein, Ziv, et al. Labeling AI-Generated Content: Promises, Perils, and Future Directions [J]. Science, 2023.

[18] Wang Jingtao, Lu Guixi. Avoiding "Labeling Fatigue": Constructing a Tiered and Categorized Labeling Rule System for AIGC [J/OL]. Information Studies: Theory & Application, 1-11.

[19] Nouwens, Midas, et al. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence [C]//Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 2020.

[20] Chesney, Robert; Citron, Danielle. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security [J]. California Law Review, 2019.

[21] Hu Yong. AI-Driven Misinformation: The Present and the Future [J]. Nanjing Journal of Social Sciences, 2024, (01): 96-109.

[22] Zheng Zhifeng, Chen Jing. Defining the Subjects of Generative AI Labeling Obligations and Their Paths of Application [J]. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition), 2025, 61(05):69-78.