"Is this AI generated?" The challenges of regulating and implementing AI-Generated content labelling in the EU and China

Jufang Wang

Table of Contents

Executive Summary

As of 2026, the blurring of digital reality has reached a critical point. The proliferation of AI-generated content (AIGC) is widely seen as threatening the "epistemic security" of our societies—the shared ability to agree on basic facts and what is real or true. In response, a wave of regulations around the world has emerged in recent years, led by the EU *AI Act* and China's *Measures of Labelling AI Generated Content*. However, implementing these regulations faces some major challenges, including regulatory ambiguities, technical limitations, and cross-border interoperability. This policy report examines and compares the regulatory approaches and implementation challenges in the EU and China, the two primary frontiers for AI governance.

1. Introduction

With the rapid advancement of generative AI tools, the line between AIGC and reality has been increasingly blurred. In February 2026, when ByteDance released its video-generation model Seedance 2.0, some commentators and users noted that its generated videos were "too realistic" —a development that could deepen the crisis of trust in content. This distrust issue is already evident. For instance, amid the United States/Israel's war with Iran, social media platforms like TikTok and YouTube have been flooded with videos claiming to show war zones, yet it is often impossible to determine whether the videos are genuine.

This policy report examines how countries and regions are responding to the issue of "epistemic security"—our shared ability to agree on what is true or real—and the challenges they face in doing so. To remain focused, I have chosen to compare two jurisdictions: the EU and China, the two primary frontiers for AI governance. The EU stands as the world's leading regulatory power for digital technologies; due to the extraterritorial reach of its regulations and the "Brussels effect", it is perhaps the most important jurisdiction for observing AIGC transparency. Meanwhile, China, a global AI superpower alongside the United States, has been a pioneer in regulating new and emerging technologies and was the first major jurisdiction to enforce AIGC labelling.

By examining the regulatory approaches and implementation challenges of the EU and China regarding AIGC labelling, this report highlights their key similarities and divergences. While both jurisdictions mandate invisible marking for all AIGC—utilizing machine-readable labels such as metadata and watermarks—they differ significantly in their requirements for visible labels. Consistent with its risk-based and rights-centred framework, the EU AI Act mandates visible labels only for deepfakes and published text of public interest. In contrast, China adopts an all-encompassing approach, requiring visible labels for all AIGC (and suspected AIGC) to maximise traceability and transparency. Furthermore, the two differ on where the technical burden lies: the EU requires AI providers to ensure marks are "detectable by design", whereas China places the primary detection responsibility on online content platforms as the final line of defense.

Despite their differing requirements, the EU and China share common challenges in regulating and implementing AIGC labelling. First, both face regulatory ambiguities and loopholes. While their specific ambiguities differ, these can lead to the under-labelling of problematic AIGC in the EU or low compliance in China. Second, technical limitations persist due to the fragility of metadata and the difficulty of creating robust watermarking. In response, the EU, which demands that technical solutions be "effective, interoperable, robust and reliable", is adopting a staged approach that acknowledges the need for technical feasibility and manageable costs. China, by contrast, utilizes an iterative approach and currently only mandates metadata for invisible marking. A third challenge lies in cross-border interoperability, as diverging labelling standards clash with the largely borderless nature of AI tools and online content.

Regarding its structure, this report first outlines the regulatory frameworks of China and the EU amid the global momentum towards AIGC labelling. It then addresses the three major challenges in detail, comparing the specific approaches of each jurisdiction and their respective implications. Finally, the conclusion summarizes the comparative analysis and offers policy considerations for future governance.

2. The global rush to label AIGC

In recent years, labelling AIGC has moved from a voluntary ethical practice guideline to a mandatory legal requirement in jurisdictions including China, the EU, India and South Korea. In China, since 1 September 2025, a new framework—the *Measures of labelling AI Generated Content* ("Measures") and the accompanying mandatory national standard GB 45438-2025—has been in full effect, supplementing earlier relevant laws and regulations from 2023. China's approach is particularly comprehensive in its distribution of responsibility. It focuses heavily on *generative AI service providers* and *online content platforms* (such as WeChat and Douyin, called "Internet information content propagation service providers" in the Measures). The former are required to embed *implicit* labels at the source or both implicit and explicit labels for *deepfakes* (i.e., realistic depictions of persons, places, objects, or events), while the latter are expected to act as the final line of defence by adding *explicit* labels for AIGC (Articles 4-6). Additionally, China requires *content creators* to declare AI use (Article 10) and *app store providers* to review the labelling documentation of AI applications before they are listed (Article 7). This creates a full AIGC "traceability chain" that spans from the moment of generation to the point of user consumption. The compulsory standard specifies how AI providers and online platforms should comply with the Measures in detail. For example, while allowing some flexibility for where to place the explicit labels, the standard requires that the height of labelling text for images and videos should be at least 5% of the shortest side of the frame and the label display duration for videos should not be less than 2 seconds.

The landmark *EU AI Act* will enter full force for AIGC labelling on 2 August 2026. It differentiates between *providers* and *deployers* of AI systems that generate or manipulate content. AI providers must ensure that their outputs are marked (using invisible labels) in a "machine readable and detectable" format, and their technical solutions must be "*effective, interoperable, robust and reliable*", as far as being "technically feasible" (Article 50 (2)). AI deployers (limited to businesses and professional creators) must disclose the AI-generated or manipulated origin of *deepfakes* and published text with the purpose of informing the public on matters of public interest ("*text of public interest*", Article 50 (4)). For deepfakes that are evidently artistic, satirical or fictional, the labels can be nonintrusive. The disclosure obligation for text of public interest can be exempted if the content has undergone human review or editorial control and a natural or legal person takes editorial responsibility for it and logs the process. In addition, Very Large Online Platforms (VLOPs, like TikTok and YouTube) have a "systemic risk" obligation under the Digital Services Act (DSA). They are not only required to label, but must detect and mitigate, risks like AI-driven disinformation including deepfakes (Article 35, DSA). Failing to label viral deepfakes can result in fines up to 6% of global

turnover. In March 2026, the EU published the second draft of the *Code of Practice on Transparency of AI-Generated Content* ("Code") and is now working against a tight timeframe to finalise it by June 2026. The Code is a guiding document for AI providers and deployers to demonstrate compliance with the marking and labelling obligations set in the EU AI Act.

While this policy report focuses on the regulatory approaches and implementation challenges of China and the EU, it is worth noting that many other countries or local authorities have also enacted their regulations on AIGC labelling. For instance, in February 2026, India amended its *Information Technology Rules*, requiring that all AI-generated or modified content that appears "real, authentic or true" must be clearly identified. Starting in early 2026, South Korea, with its *AI Basic Act*, mandates clear labelling when AIGC, including generated advertisements, is difficult to be distinguished from reality. As of writing, the U.S., where many leading developers of Large Language Models (LLMs) and global social media platforms are headquartered, has no single federal law mandating the labelling of AIGC, leaving states to pass their own regulations. The state of California has passed the *AI Transparency Act,* which requires covered AI providers (over 1 million monthly users) to embed invisible metadata into AIGC and to provide a free, public detection tool to verify whether a piece of content is generated by that specific AI system from 2 August 2026 (harmonised to the same enforcing date as the EU AI Act).

Despite varying political and cultural contexts, global AIGC labelling laws and regulations are converging on a "dual-layer" transparency model. Most regulatory frameworks including the EU and China require both explicit or visible/audible labels (such as icons and spoken disclosures) for human viewers and implicit or invisible labels (typically in the form of cryptographically signed metadata, watermarks and fingerprints) for machine detection. Another convergence trend is that most frameworks differentiate between general AIGC and deepfakes, imposing heavier requirements, or only focusing, on the latter.

With the global measures in mandating the labelling of AIGC, the real challenges may lie with the implementation of the regulations. For the EU and China, these challenges mainly include regulatory ambiguities and loopholes, technical limitations, and cross-border interoperability among different marking and labelling standards.

3. Regulatory ambiguities and loopholes

The EU's draft Code has been responsive to feedback from the AI Industry, academics, policy experts, and member states. For instance, in the second version of Code, the EU had completely removed the taxonomy distinguishing the AI–generated content from AI-assisted content, as the line between them can be too blurred. However, some regulatory ambiguities are embedded in the AI Act itself.

First, the EU AI Act's transparency obligations are notably narrow, mandating visible labels only for deepfakes and text of public interest. While this reflects the EU's risk-based philosophy for AI regulation, it assumes a baseline level of "AI literacy" that does not exist

uniformly across society. A "generic" AI-generated image—such as a fictional historical scene or a manipulated animated character—may be recognised as "synthetic" by a tech-savvy adult but could easily deceive vulnerable groups like children or the elderly. By only requiring visible labels for high-risk categories, the Act risks creating an information environment where a massive amount of generic AIGC remains unlabelled for human viewers, which may cause psychological or financial harm for many users. While a choice by the EU regulators, this still constitutes a regulatory loophole from the point of view of end-users, especially some specific demographics.

Second, the Act defines a "deployer" as a person or entity using AI in a professional capacity, exempting "personal non-professional activity" (Article 3(4)). This creates a significant enforcement vacuum. In today's "creator economy", the line between a hobbyist and a professional is far from being absolute. In practice, there will be the threshold issue: say, if a creator has 5,000 or more followers but does not formally monetise their account, are they "non-professional"? This leads to the risk that high-reach accounts could circulate unlabelled deepfakes or AI-generated political commentaries under the guise of "personal expression". On some platforms like TikTok (content distribution is mainly interest-based, rather than followers-based), even if a creator has no followers at all, it is still possible that their content reaches a huge number of viewers. While platforms are expected to fill this gap under the DSA by detecting and marking such content, technical limitations mean that a big volume of problematic synthetic content will likely remain unlabelled.

The exemption for AI-generated text of public interest that has undergone human review is perhaps another exploitable regulatory loophole in practice. While a human or entity must take "editorial responsibility" and document the process for accountability, there exists a major verification hurdle. For established media organizations, documentation is usually a standard operating procedure. But for the millions of "news-adjacent" influencers, there is no cost-efficient way for regulators (or platforms) to verify their editorial logs. In practice, inspection of such documentation is likely to be reactive rather than proactive. Unless a piece of content goes viral and causes measurable harm or triggers a lawsuit, a creator's claim of "human review" will effectively go unchallenged.

In addition, the judgement about what constitutes "text of public interest" remains a matter open to judgement and thus may create another loophole for transparency implementation. The EU regulators can rely on existing media laws to decide what constitutes matters of public interest, which usually includes news on current affairs, elections, and public health, but the classification in practice won't be so straightforward. For instance, an influencer posting AI-generated health advice might view it as a personal lifestyle tip, while a regulator might see it as a public health matter requiring disclosure. This leads to the situation that creators may not even realise that they are violating the Act until they are facing enforcement actions. For years, a similar regulatory dilemma has been observed in China, where online content authorities do not allow accounts owned by non-media organisations to publish news on current political affairs (时政新闻), but the breaches of this stipulation are wide-spread on

Chinese platforms partly due to the vague nature of what constitutes "current political affairs".

In contrast to the EU's risk-based approach, China has adopted an all-encompassing transparency model. By placing legal obligations on every link in the chain—AI providers, online platforms, app stores, and content creators—and requiring both implicit and explicit labels for all AIGC, China has removed the regulatory ambiguities on what to visibly label and closed the "professional vs. amateur" regulatory loophole that exists in the EU. However, China's regulatory ambiguities lie in the disconnection between the detailed requirements for labelling AIGC and the vagueness of its punitive measures. Like China's many other low-level regulations enacted by government departments (different from laws passed by China's legislature), the *Measures* only stipulate that breaches will be handled by "relevant regulatory authorities based on relevant laws and regulations" (Article 13). China's relevant laws or regulations usually stipulate a relatively low fine (such as a few thousand dollars), unless for data security or monopoly cases. In comparison, in the EU, the penalty for failing to label AIGC or disclose the AI use can reach up to €15 million or 3% of the total worldwide annual turnover for the preceding financial year, whichever is higher (for SMEs and Startups, whichever is lower).

Unsurprisingly, China's ambiguity and leniency on fines have led to inconsistent enforcement and low compliance. In January 2026 (over 3 months after the Measures took effect), a Chinese media report noted that while major AI providers and online platforms had already implemented the labelling obligations, the report discovered a lot of unlabelled AIGC and an emerging shadow market for label-removal services. Citing a legal expert, the media report pointed to the vague punitive standard and limited detection technologies as the two main reasons that led to the wide-spread breaches. In February 2026, the Cyberspace Administration of China (CAC) announced that over 13,400 non-compliant accounts were penalised and over 543,000 pieces of illegal and non-compliant information were removed. Non-compliant examples listed by the CAC included: AI-generated sensationalised stories to drive traffic, the unauthorised production of celebrity deepfakes for commercial gain, the dissemination of AI-generated misinformation such as fake "fire incidents" that disrupted social stability, and the malicious alteration of animated characters that harmed the psychological well-being of minors.

4. The technical limitations of AIGC marking and detection

While the EU and China have adopted different approaches towards visible labelling, both require invisible labelling (i.e., marking) for all AIGC. China's regulations and standards place the obligation of marking using metadata on AI service providers and detection on online content platforms, without explicitly requiring robustness of the marks or the accuracy of detection. In contrast, the marking and detectability obligations fall on AI providers in the EU. Under Article 50(2) of the EU AI Act, AIGC technical solutions must be "effective, interoperable, robust, and reliable", to the extent technically feasible. More specifically, according to the Code, "effective" means high detection accuracy with minimal false

positives or negatives; "interoperable" requires marks being read across all distribution channels and technical environments; "robust" requires marks resistant to common alterations or adversarial attacks; and "reliable" means marks delivering consistent performance across various content modalities (text, image, video, and audio) (Code).

One major challenge for marking and detection is that metadata can be easily stripped. For instance, taking a screenshot of or cropping, or compressing an AI-generated image would remove its metadata. Apart from accidental removals of metadata, content creators may have the incentives to avoid the disclosure of AIGC for commercial, aesthetic or reputational reasons, given that content without AI labels usually appears more authentic or reliable. Additionally, malicious actors, such as those running coordinated disinformation campaigns, may strip labels to spread synthetic propaganda. To combat the fragility of metadata, China's *Measures* explicitly prohibit any organisation or individual from "maliciously deleting, tampering with, forging, or hiding content identification" (Measures, Article 10). Despite this, enforcement of AIGC labelling in China has proven to be a "cat-and-mouse" game. As mentioned above, a shadow market, where accounts actively sell services and tutorials designed to strip AIGC identifiers, has emerged in China. As Chinese leading legal expert Linghan Zhang observed, establishing a comprehensive content labelling system is a long-term systematic project and it is necessary to adopt a "step-by-step" approach. Digital watermarking and other technical solutions that are highly resistant to tampering and compatible across platforms have yet to be introduced in China.

The EU has taken another approach to address the fragility issue. Its Code of Practice requires *multi-layer* marking. Beyond metadata, watermarking will also be compulsory for AI system providers. Watermarks are difficult to remove, as they are imperceptible patterns directly embedded in the content, rather than just in the file header as metadata. Other technical solutions, such as fingerprinting, remain optional in the EU. Additionally, the Code also requires signatories to include in their relevant documentation a prohibition for the intentional removal of, or tampering with, the invisible marks by deployers or any other third party (Measure 1.2). However, even with these measures present, there is no guarantee that most AI providers in the EU can meet the "effective, interoperability, robust and reliable" criteria for marking and detection. This is why the AI Act and Code repeatedly state that all requirements are limited by "the state of the art" and need to take the "costs of implementation" into consideration (the cost factor is especially important for SMEs).

Apart from marking, detecting AIGC can be another technical challenge for AI deployers and online platforms. China's *Measures* require online platforms to conduct AIGC detection and implement a three-tier labelling system: (1) *AIGC*: When machine-readable metadata is detected; (2) *Suspected AIGC*: When platform algorithms detect "AI-like" traits despite a lack of metadata; (3) *Possible AIGC*: When a user proactively declares the content as AI-generated but no technical mark is present (Articles 12-14). Chinese online content platforms like Douyin and WeChat thus act as the "frontier guards" of AIGC detection and labelling. To demonstrate compliance, there is a possibility that platforms may over label "suspected AIGC". In the EU, the detection burden is inverted. The Code requires signatories of AI

providers to provide an interface or a publicly available detection tool free of charge for downstream users to verify whether content has been generated or manipulated by their AI system (Code, Measure 2.1). For online content platforms in China or the EU, detecting AIGC without invisible labels (intentionally or accidentally removed) is a challenging task, as AI detection tools are notorious in generating false positives or negatives. Also, every time an AI detector is built to find "AI artifacts" (like weird fingers or inconsistent shadows), creators may use AI to fix those specific flaws.

5. The cross-border interoperability challenge

Another hurdle for AIGC labelling is the divergent approaches of jurisdictions—for example, China only mandates metadata marking but the EU requires a multi-layer marking solution—and the lack of a unified labelling standard. This creates an interoperability problem in today's largely borderless online content ecosystem. For instance, a Brazilian creator can prompt a Chinese-origin model (e.g., DeepSeek-V3, Qwen 3.6, or SeeDance 2.0) and post the output on YouTube, where it is consumed by a user in the EU. To make this process frictionless, the technical solutions for marking and detecting AIGC need to be interoperable. Given that the EU AI Act has extraterritorial reach, it applies to any AI content that is accessed within the EU, regardless of where the platform or deployer is based. Thus, without a unified international standard or a "universal translator" between these marks, a European platform may fail to recognize a legitimate AI label applied by a tool from China or other countries (and vice versa), leading to unintentional non-compliance or the spread of unlabelled synthetic media.

The EU and China have adopted different standards for invisible marking. This reflects their respective technical solutions and regulatory purposes: China emphasizes the traceability of content, from generation to propagation, while the EU focuses more on the provenance of content, requiring AI system providers to ensure their marks are "detectable by design". Specifically, the EU's Code of Practice demands that technical solutions must be interoperable, "regardless of the application domain, context or content modality formats", with the goal of achieving "full interoperability" (Measure 3.4). Given the lack of a unified international standard, the EU is adopting a staged approach based on the state of the art. For the metadata layer, the Code requires providers to adopt a "secure, tamper-evident, open verification standard", combined with a digital signature with an interoperable content identifier for the other marking layer(s). For the watermarking layer, the EU is requiring providers to encode an imperceptible public mark that links to the metadata identifier. In comparison, China's compulsory standard requires that metadata must include the following information: content nature (i.e., AIGC, possible AIGC, or suspected AIGC), content producer ID, content identifier1 (filled in by AI providers), platform ID, content identifier2 (filled in by content propagation platforms). Despite adopting different approaches to invisible marking, both the EU and China require metadata to include information about provenance and a unique content identifier(s), making mutual recognition of the other's standards possible.

Facing the global interoperability challenge, the industry standard from the Coalition for Content Provenance and Authenticity (C2PA) is emerging as the de-facto international standard for metadata marking. It is an open, industry-led "content credentials" system that attaches cryptographically signed metadata to a file, identifying the model used, the edits made, and whether AI was involved. C2PA is supported by a coalition of "Big Tech" (e.g., Microsoft, Google, OpenAI, Amazon, TikTok, and Samsung) and "Big Media" (e.g., BBC, AP, Sony, Bloomberg, etc.). Notably, some Chinese-origin companies like Huawei and Vivo (smartphone producer) are also C2PA members, indicating that Chinese companies with global ambitions are voluntarily adopting this standard to maintain access to Western markets. However, unless a Chinese tool (like DeepSeek and Baidu) specifically chooses to export its metadata in a C2PA-compliant format, Western platforms like YouTube and Instagram may not be able to read it. Even if metadata interoperability can be achieved, the EU's "full interoperability" goal remains elusive because, as discussed above, metadata is easily stripped.

The lack of interoperability between the EU and China (and other jurisdictions) creates a massive cross-border identification gap that threatens global information integrity. This friction allows for "misinformation laundering," where deceptive content is generated using a tool from one jurisdiction and then uploaded to a platform regulated in another. Because the hosting platform cannot read the foreign tool's label, the synthetic content may be presented to users as authentic. Without interoperable marking and detection standards, the public will be vulnerable to sophisticated, cross-border manipulation due to the circulation of unverified content.

6. Conclusions

This policy report compared the regulatory approaches of the EU and China in AIGC labelling and examined the challenges both jurisdictions face in implementing the regulations. While noting some convergent trends in AIGC labelling, it examined the divergent regulatory paths between these two major AI governance jurisdictions, including the risk-based vs. all-encompassing approaches on the visible-layer labelling and the different requirements for technical solutions on the invisible-layer labelling. It centred on three types of challenges in implementation: regulatory ambiguity and loopholes, technical limitations, and cross-border interoperability. While these challenges are largely shared by the two jurisdictions, there are differences due to divergent regulatory philosophies. For example, the EU's risk-based approach may lead to more regulatory ambiguity and under-labelling loopholes compared to China’s approach. In contrast, China's all-encompassing approach is intended to achieve high level of AIGC transparency and traceability; but in practice, it may lead to over labelling and potential "label fatigue", and widespread breaches have been reported due to vague punitive measures.

Some of the regulatory "loopholes" identified in this report are, in fact, intended trade-offs between competing values. For the EU, exempting non-professional activity is a choice to protect individual freedom of expression, despite the potential for harm to content viewers. Likewise, the EU avoids a fixed definition of content of public interest to allow contextual

interpretations and preserve future flexibility. For China, the vague punitive measures for breaches and leniency on fines (especially when compared to the EU) are consistent with its so-called "tolerant and prudent" (包容审慎) digital regulatory principle that is expected to encourage technology innovation. However, both the EU and China will need to balance their different goals and find suitable solutions to address some of their loopholes to protect information integrity.

The technical limitations of AIGC marking indicate that the transparency goal is a moving target, necessitating a staged and iterative approach. Technical solutions for marking and detecting AIGC must keep up with the development of generative AI technologies. Both the EU and China have demonstrated that they understand the technical limitations. The inclusion of the "technically feasible" clause in the EU AI Act reflects a grounded realism, acknowledging that even the highest standards must account for the limitations of the state-of-the-art. This staged approach is mirrored in the EU's roadmap for interoperability, which moves from static identifiers to interactive icons as technical capabilities evolve. Similarly, China's current reliance on metadata—while encouraging but not mandating watermarks—is a pragmatic admission of current technical limitations. As watermarking technology matures, it is highly probable that China will shift toward mandatory watermarking to align with EU-level robustness of marking.

Because generative AI models/tools and online content are inherently borderless, the lack of interoperability between labelling standards will affect the effective labelling of AIGC in all jurisdictions. A digital mark applied by an AI provider in China must be seamlessly readable by a platform in the EU if information integrity is to be achieved, especially because many Chinese AI tools are becoming popular globally. Policy efforts should therefore prioritize the creation of "technical crosswalks" between C2PA-based standards that have been adopted by major Western AI providers and platforms and domestic standards developed in China (and other countries). To enhance cross-border interoperability, it is necessary to establish international standards on AIGC labelling or find other solutions, such as a shared database containing all up-to-date marking standards or a technical "bridge" between these divergent standards.

Acknowledgement

The author wants to thank the insightful and helpful feedback from Professor Robin Mansell (LSE) and Professor Ralph Schroeder (Oxford) on an earlier version of the report. She also benefited from the discussion with Ms Claire Milne (independent ICT consultant).