

Accountability and Attribution in AI-Generated Content Authentication: Lessons from China's AI Content Labelling Mandate

Wenlong Li¹

¹ Research Professor, Zhejiang University.

Corresponding author. E-mail: fettes.lee@gmail.com

Abstract

This paper interrogates how China's pioneering AI content labelling mandate responds to urgent, transnational concerns of provenance and authentication concerning AI-powered content generation and dissemination. Effective 1 September 2025, this unprecedented regulatory framework compels AI providers, deployers as well as distribution platforms to implement a bifurcated schema of explicit and implicit content labelling, thereby prioritising unambiguous provenance and seamless traceability across the entire lifecycle of AI-generated artefacts.

It is observed that China's regime is aggressively prescriptive, hierarchical and deterrent, positioning itself as a laboratory for rigorous attribution. Essentially, China's approach does not merely mitigate digital deception domestically, but also reconfigures international regulatory paradigms by pairing technical mandates with a robust organisational extensive architecture for accountability. It actively reconfigures platform and provider incentives at scale, creating a testbed for enforcing provenance solutions.

The resonance of these measures is unmistakably global: parallel developments across the world (e.g., the EU's AI Act and emergent statutes in some US states), some of which have seemingly influenced China's development, echo its commitment to provenance, and China's action both reflects and shapes the emergent consensus. Rather than advancing prescriptive solutions or engaging with comparative analysis, this short piece offers a measured examination of China's regulatory experiment as a vantage point for broader reflection. It illuminates the underlying questions and emerging dilemmas in the discourse around AI authentication and attribution and leverages China's leading steps to provoke critical thought about the possibilities and limitations of governance strategies in a rapidly evolving digital environment.

Keywords

China, AI-Generated Content, AI Supply Chain, Content Labelling, AI Regulation.

DOI: 10.65701/q9m2g5d8x1

1 Introduction

The exponential rise of AI-generated content has precipitated a moment of reckoning for global regulatory and ethical discourse. As generative model now generates text, image, video and audio that defy effortless human discernment, the prospect of a reality saturated with indistinguishable synthetic media has provoked urgent demands for both robust content authentication and sophisticated governance architectures. Recent episodes – including high-profile deepfake political interventions that have reshaped electoral narratives (Kapoor and Narayanan, 2024; Labuz and Nehring, 2024) or the viral dissemination of fabricated news orchestrated by advanced AI (Adami, 2024) – have accentuated the need for mechanisms that ensure transparency and accountability in an increasingly synthetic information environment.

The challenge of AI content authentication is further magnified what (AI) ethicists term the ‘many hands’ problem: as AI-generated content initiates within sophisticated technical pipelines and subsequently traverses a myriad of hands – including developers, platform intermediaries, automated distributors, and global user networks – attribution and accountability are dispersed to the point of opacity (Nissenbaum, 1996). The literature on the AI supply chain stresses the fractal complexity of these pathways, cautioning that authentication must grapple not only with technical provenance, but also with the unpredictable and collective agency embedded in a polycentric, trans-platform content ecosystem (Cobbe et al., 2023a). Crucially, the challenge of content authentication extends beyond the purview of direct supply-side actors. As AI-augmented content migrates through social networking sites, news aggregators, messaging services, audiovisual streaming platforms, and the expansive terrain of user-generated communities, the dissemination mechanism become radically polycentric. This diffusion does not merely obscure provenance but multiples vectors for manipulation and abuse, thereby intensifying the risks posed by harmful or deceptive content.

Imagine that a wellness influencer launches a series of AI-generated explainer videos touting a new dietary supplement. As the content spreads, marketing agencies splice in fabricated user testimonials and doctor endorsements produced by text-to-video AI systems. The videos then move downstream: affiliate marketers employ additional automated tools to dub the material into several languages and inject trending but wholly unfounded health claims tailored to different markets. On social media and messaging apps, virality takes over, leading to the remixing of the original videos into short clips and meme, many stripped of context or transformed to highlight unproven cures. Influencers and micro-celebrities join the trend, further amplifying exposure with personalised stories, while e-commerce vendors automate links to dubious supplement sellers beneath each post. This scenario isn’t far from reality and begs a host of persistent questions: how should liability be assigned across the tangled web of originators, curators, distributors and users? What regulatory and technical mechanisms might reliably reconstitute provenance when the chain is so easily broken? Can any system of accountability remain effective in a landscape defined by constant remixing, opacity, and the expanding capacity of both people and artificial agents?

The intricacy of this scenario, marked by relentless remixing and indeterminate lines of agency, starkly reveals the profound uncertainties and unresolved questions that contemporary governance regime must now confront. The European Union’s Artificial Intelligence Act (AIA) stands as the first comprehensive attempt to address these concerns, introducing binding obligations such as the disclosure of synthetic media and machine-readable standardised labels for AI-generated content (EU, 2024). Across the Atlantic, the United States presents a markedly fragmented landscape featuring a mosaic of state-level statutes (notably in California (California, 2024), New York (N.Y., 2024) (US, 2025) and Washington (Washington, 2025) among a few others⁶³) still under deliberation, some ill-fated federal bills (US, 2023b)⁶⁴, and voluntary industry commitments encouraged by the then Biden administration (US, 2023a)⁶⁵. In contrast, China has instituted perhaps the most far-reaching and rapidly

⁶³Other developments at state levels include Wisconsin (Senate Bill 644, mandating disclosures in political ads if they contain AI-generated synthetic media), Tennessee (Senate Bill 2431, requires disclosure for AI-generated content in Tennessee using another’s likeness or falsely attributed authorship), Oklahoma (Senate Bill 746, mandating disclosure of AI-generated content in political ads, requiring a statement indicating AI involvement in creation). Notably, House Bill 2094 from Virginia that would have defined certain types of AI as “high risk” and placed transparency and reporting requirements on their usage was vetoed by their governor. see Andrews (2025)

⁶⁴The Federal Communications Commission (FCC) announced a proposal on 25 July 2024 on Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements, which would require broadcast stations to disclose AI-generated content through on-air announcements, defining such content as images, audio or video created using computational technology. see Commission (2024)

⁶⁵Following the Biden administration’s soft-handed regulatory approach to AI—including Executive Order 14110, AI policies in the United States has fundamentally shifted since Trump’s return to office; the Trump administration rapidly rescinded key Biden-era directives and, under its new AI Action Plan, is now pursuing an aggressive program of deregulation. See House (2025).

implemented interventions to date, enacting mandatory requirements for the tracing and labelling of AI-generated content. Despite its historical and global significance as the first of its kind, developments in China remain under examination within much of western policy discourse. It is precisely this overlooked Chinese trajectory, with distinct legal architecture and implications for global governance, that forms the focus of this paper. To be specific, this short piece seeks to clarify China’s rapidly evolving statutory and administrative regime of AI content authentication. It does not attempt to perform a comprehensive comparative analysis, and any reference to transnational borrowing or divergence serves to situate China’s trajectory within the increasingly interconnected global evolution of AI governance.

It is argued that China’s AI content labelling mandate stands as an unprecedented statutory response to the complex, polycentric reality of AI-generated content. In essence, it seeks to proactively “stitch together” a continuous chain of accountability, binding service providers, distribution platforms and even end-user activities with an auditable compliance framework even as technical fragility persists. At the core of this regime is the transformation of “collective responsibility” that shifts the enforcement bottleneck from upstream providers to content distribution infrastructures, leveraging platform centrality in digital content flows as a regulatory fulcrum (Mühlhoff, 2025; Taylor, 2024). The stability and persistence of provenance metadata – central to China’s regulatory architecture – remain vexed by technical and economic realities. Hence, its effectiveness will likely turn on the regime’s ability to close technical loopholes, balance innovation with oversight, and coordinate standards both domestically and internationally. In practice, perfect compliance is improbable, making broader, layered approaches to accountability and risk mitigation essential. The Chinese mandates anticipate these weaknesses by criminalising removal or alteration of markers and prohibiting the distribution of tools for metadata erasure⁶⁶, but the technical arms race between robust provenance and circumvention tools is likely to persist, posing persistent threats to operational scalability. By channeling legal duties through a dense regulatory nexus – where platforms are accountable not only for detecting and flagging but also for reinforcing and tracing every instance of AI-generated content – the regime imposes significant compliance demands and operational costs, particularly for smaller developers or cross-border actors. The regime’s success ultimately hinges less on perfect technical integrity than on the visibility, deterrence, and enforceability of its mandates. Large commercial providers and mainstream platforms are more incentivised to comply, but enforcement against smaller developers, open-source deployments, and gray-market actors remains a critical bottleneck.

This article is structured as follows. Following the introduction, Section II elucidates the conceptual and technical architecture of AI-generated content authentication, interrogating the technological complexities that render the tracing and substantiation of synthetic media particularly challenging. Section III pivots around China’s trailblazing regulatory intervention in AI content labelling, offering a precise synopsis of the pertinent statutory and administrative obligations. Section IV critically assesses how its regime has grappled with the intricacies of provenance verification and authentication while exposing its blind spots, structural vulnerabilities, and regulatory lacunae. The final section situates these developments within the wider transnational and global regulatory landscape, drawing connections to ongoing international momentum and highlighting their prospective significance for the evolving terrain of AI governance.

2 AI-Generated Content Authentication: A Primer

Artificial intelligence-generated content authentication is undergoing a rapid metamorphosis, driven by the meteoric rise of synthetic media and an ever-more pressing need to safeguard digital trust. At its heart lies an intricate interplay of provenance, authenticity, and watermarking—concepts inseparably bound in their mission to combat digital deception. Content provenance emerges as the “digital paper trail” of our era: a meticulously curated ledger chronicling a digital asset’s origin, creative process, and every intervening alteration (Werder et al., 2022). It weaves together details of authorship, chronology, locale, and the myriad transformations a piece of content endures as it traverses digital ecosystems. Provenance extends traditional notions of authenticity into the digital realm, ideally leveraging technological methods – such as cryptographic signatures and tamper-resistant metadata – to help users verify that certain content is genuine and unaltered (Longpre et al., 2024b; Pan et al., 2023). Authenticity, in this context, is the degree to which digital material can be reliably traced to its original source and verified as untampered. Provenance data underpin this authenticity, offering assurance against digital manipulation,

⁶⁶Art 10, 13 MLAIGSC.

misinformation and the increasing sophistication of synthetic media generated AI (Bereskin, 2023; Feher, 2025; Schick, 2020). A case in point is the Coalition for Content Provenance and Authenticity (C2PA), which stands as an open, cryptographically assured standard for tracking digital content origins and edits. Its Content Credentials bind provenance data to content using digital signatures, and any change severs this cryptographic bond, making tempering immediately evident across platforms and throughout the content lifecycle.⁶⁷

A distinction must be made among the various labels and markers deployed within authentication frameworks. Occasionally, the term “watermark” is indiscriminately applied to any conspicuous textual overlay—such as “DRAFT” stamps or “Not for Release” banners—that signal provisional status or restricted access (Council, 2024a). Yet, in the specialised context of digital authentication, watermarking takes on a more nuanced meaning. It signifies the deliberate embedding of subtle signals or patterns within digital content—crafted to be virtually invisible or inaudible to the unaided senses (Srinivasan, 2024). Unlike overt overlays or pictorial icons, these sophisticated marks are engineered to elude ordinary perception, surfacing only through the use of specialised algorithms or proprietary decryption keys. This paper embraces both visible and veiled modes as integral to contemporary authentication strategies and confines the term “watermarking” to its most precise sense: the surreptitious infusion of hidden, machine-detectable markers, intended exclusively to establish provenance and safeguard authenticity.

A multitude of techniques—none of which serve as a universal panacea—exist for content authentication. A time-honoured mainstay (and also the linchpin of China’s approach as revealed later) is metadata: embedded digital records chronicling a file’s origin, authorship, and processing history (Cheney et al., 2009; Deelman et al., 2009; Simmhan et al., 2005). Yet, relying on metadata is fraught with technical and organisational vulnerabilities (Hart and de Vries, 2017). Many mainstream social media, content-hosting, and file-sharing platforms routinely excise metadata (such as EXIF data or creator tags) when users upload or download material, citing privacy, file size and interoperability concerns.⁶⁸ Further, metadata can be manipulated or forged by adversaries and even well-intentioned users, thanks to the ready availability of metadata editing tools. While advanced metadata standards or cryptographically signed provenance records offer a promising remedy to achieve tamper resistance or evidence, their effectiveness hinges on seamless, consistent support across a chaotic digital landscape—otherwise, provenance can be irretrievably lost as content circulates between systems adhering to divergent standards or formats.

Advanced statistical watermarks present a sophisticated alternative to conventional metadata by weaving subtle markers directly into the very structure of digital outputs—manifesting, for instance, as unique pixel distributions in images or as distinctive token patterns within text (Srinivasan, 2024). These covert techniques, while ingenious, are not immune to tampering. Common manipulations such as format conversion (for example, switching from PNG to JPEG for images, or re-encoding textual data) can easily disrupt or completely obliterate the fragile statistical fingerprints that constitute these watermarks (Ren et al., 2024). Furthermore, adversaries with a vested interest in erasing provenance can deliberately employ strategies like image augmentation, paraphrasing, or noise injection—essentially randomising or mimicking watermark characteristics to obfuscate the content’s origin or to spoof detection mechanisms (Xu et al., 2024). The absence of universal standards further compounds these vulnerabilities: many watermarking schemes are proprietary and tailored to specific models, so even robust watermarks may fail if every participant in a distribution chain is not equipped for their detection and validation. Open-source models further pose a structural challenge, as technical protections can be swiftly disabled by users with minimal sophistication (Srinivasan, 2024). Therefore, watermarking is thus adept at “raising the bar” for evaders, primarily within closed, commercial ecosystems, but cannot be seen as a fail-safe option.

⁶⁷C2PA provides an open technical standard enabling digital media (such as images, video, audio, and documents) to carry cryptographically signed provenance metadata, known as Content Credentials. These credentials record who created the content, when, how, and with what tools, as well as subsequent edits or incorporated elements, all in a tamper-evident, verifiable structure. The standard does not act as DRM; rather, it promotes transparency regarding content origin, edit history, and chain of custody. Removing or altering these credentials is detectable because cryptographic hashes would fail to validate. C2PA is designed for compatibility with common metadata standards (like IPTC, XMP, EXIF) and can be used offline, with only a minimal increase in file size. The standard and its Trust List ensure that only credentials issued by certified authorities are considered trustworthy. For more, see the C2PA FAQ: <https://c2pa.org/faq/> (accessed 2025-07-31).

⁶⁸The Social Media Sites Photo Metadata Test evaluated 15 top social media sites, and checked if embedded metadata was retained and displayed on upload to the sites or downloads of various version of the image. The results are displayed at www.embeddedmetadata.org/testresults. Only one social media site, Behance, received favourable results for retaining and displaying embedded data. A few systems retained embedded metadata but failed to use it when displaying metadata on the web site. Ten sites removed at least some metadata when images were downloaded to a desktop environment. <https://iptc.org/news/many-social-media-sites-still-remove-image-rights-information-from-photos/>, see also <https://www.1854.photography/2013/03/study-exposes-social-media-sites-that-delete-photographs-metadata/> or <https://blogs.loc.gov/thesignal/2013/04/social-media-networks-stripping-data-from-your-digital-photos/> See also (Irwin, 2024).

Two additional techniques—post-hoc detection and retrieval-based methods—expand the capabilities for authenticating AI-generated content. Post-hoc detection leverages machine learning to identify subtle statistical anomalies that suggest AI authorship. While such methods are appealing for their universal applicability and lack of dependence on the content’s origin or embedded signals, they tend to have high rates of false positives and negatives, particularly as generative models improve and content undergoes minor edits or paraphrasing, limiting their use mainly to low-stakes environments (Sadasivan et al., 2025; Weber-Wulff et al., 2023). Retrieval-based approaches work by checking suspect content against a database of previously recorded AI outputs. These techniques are most effective when the reference databases are comprehensive and current, yet they face significant privacy and scalability hurdles, especially due to the need for extensive data storage and international governance (Krishna et al., 2023). The Information Technology Industry Council (ITI) also stress the importance of human authentication (a vital fail-safe despite being resource-intensive and subject to bias) while highlighting alternative approaches that leverage blockchain-based tracking and behavioural analysis (Council, 2024b).

Visible signals—while essential for promoting transparency and raising public awareness about AI-generated content (Burrus, Curtis, and Herman, 2024; Wittenberg, Epstein, Berinsky, and Rand, 2024)—are technically better described as explicit content labels or marks rather than watermarks in the strict sense, as now required in new regulations from multiple jurisdictions such as EU, US (some states) and China (Chomanski and Lauwaert, 2025; Fisher, 2024). Notwithstanding the well-versed debates on the efficacy, costs, and unintended consequences of labelling (Gamage, Sewwandi, Zhang, and Bandara, 2025; Jung, Hua, Bao, and Sundar, 2025; Scharowski, Benk, Kühne, Wettstein, and Brühlmann, 2023), these visible labels are similarly fragile. They can be quickly cropped, painted over, or otherwise erased using basic editing tools—often with minimal degradation to the underlying content. Once a labelled item is downloaded, reposted, or digitally repurposed, these explicit marks are frequently omitted, especially as files are converted between formats or screenshot mechanisms are employed to generate “clean” versions with no identifying tags. The distributed and often fragmented nature of the content sharing ecosystem further compounds this fragility: even if explicit labels exist at the point of first publication, their persistence cannot be assured as content migrates across platforms, intermediaries, and devices. In practice, any actor in the distribution chain can intentionally or inadvertently strip labels, rendering visible signals ephemeral.

Therefore, achieving robust provenance and authenticity for digital content is beset by persistent technical, organisational, and legal challenges. All methods—whether provenance tracking, watermarking, or human oversight—face trade-offs, and no solution is fool-proof; adversarial actors continually find ways to circumvent detection, making universal, perfect authentication unattainable. Hence, adopting a context-sensitive strategy that blends different techniques may represent the most realistic path forward. Critically, content authentication remains an adversarial, rapidly evolving field, demanding not just technical progress but also cross-platform cooperation, adaptive regulation, and shared accountability. As regulatory initiatives in China will reveal, sustainable solutions will require coordinated, ongoing adaptation rather than any single, definitive policy or technology.

3 China’s AI Content Labelling Mandate: A Brief Overview

This section presents a succinct, high-level synopsis of China’s Measure for Labeling of AI-Generated Synthetic Content (MLAIGSC), promulgated in March 2025 and in force since September 2025(?). China’s hesitance to promulgate a comprehensive and omnibus legislative regime governing artificial intelligence is increasingly evident, as evidenced by the conspicuous absence of AI-specific regulation from its most recent legislative blueprints.⁶⁹ In lieu of sweeping statutory frameworks, China has opted

⁶⁹The 2025 Legislative Work Plan of the Standing Committee of the National People’s Congress (NPC) sets out comprehensive goals and priorities for the final year of China’s 14th Five-Year Plan. In 2025, the work plan includes completed and ongoing review of legislation in areas such as economics, society, environment, and security (e.g., Infectious Disease Prevention Law, Financial Stability Law), as well as first reviews of 23 new laws covering topics like the Ecological Environment Code, Financial Law, and Social Assistance Law. Preparatory reviews are anticipated for major new domains—including State-Owned Assets Law, Excise Tax Law, and legislation on the healthy development of artificial intelligence and emerging technologies. The policy framework emphasizes Party leadership, legislative quality, people’s participation, and alignment with national strategies. Other features include strengthened constitutional compliance review, international communication of legal texts, and optimization of the socialist legal system—especially targeting digital, ecological, foreign-related, and technological innovation legislation in 2025. See the full 2025 NPC legislative plan: <http://www.npc.gov.cn/npc/c2/c30834/202505/P020250513550316685290.pdf> (accessed 2025-07-31). The 2025 Legislative Work Plan of the State Council of China outlines the main legislative goals and priorities for the year. It focuses on advancing high-quality development, implementing the rule of law, and supporting initiatives in economic reform, social governance, public health, technological innovation, and environmental protection. The plan includes drafting and amending key laws and regulations related to digital economy, social security, ecological conservation, and risk management. Emphasis is placed on aligning legislation with strategic national objectives, enhancing public participation in the legislative process, and optimizing the legal

for a stratified approach—deploying a series of subordinate, mandate-oriented regulatory instruments that nonetheless wield substantial regulatory force. Enacted through the concerted action of a consortium of regulatory heavyweights—including the Cyberspace Administration of China, the Ministry of Industry and Information Technology (MIIT), the Ministry of Public Security, and the National Radio and Television Administration—this mandate signals an unequivocal intent to blanket the entire media and digital ecosystem, embedding compliance across diverse platforms and contexts. In recent years, China has developed a consistent pattern of imposing technology-specific regulations—progressing from algorithmic recommendation systems, through synthetic media controls, to the contemporary regime governing generative AI. The latest mandate on AI content labelling not only amplifies the horizontal sweep of regulatory oversight but also telegraphs a calculated bid for leadership in the global contest to frame, and perhaps dictate, the norms and protocols of next-generation AI governance. Far from deriving solely from indigenous ingenuity, China’s evolving governance template is, at least in part, derivative—inflected by transnational currents, with discernible conceptual borrowings from ongoing legislative debates within the United States, even as its legislative machinery outpaces the more deliberative tempos characteristic of Western counterparts (Li and Chen, 2024). China’s pragmatism manifests as a deliberate strategy, whereby emergent sectoral rules serve not only immediate domestic imperatives but also China’s overt aspiration to consolidate normative influence over global AI standard-setting. This ambition has, in recent years, translated into a concerted effort to assert China’s prominence well beyond academic discourse and into the international architecture of AI governance—most conspicuously crystallised in the trajectory of its AI Action Plan, which stands in notable juxtaposition to the contemporaneous, albeit more fragmentary, policy overtures of the Trump Administration in the United States (Guardian, 2025).

To start with, China’s AI content labelling mandate operates a dual system that requires both explicit (visible) and implicit (metadata-based) labels for AI-generated content⁷⁰; however, the regulatory framework currently provides no substantive provisions for tamper resistance or verifiable evidentiary mechanisms at the quasi-legislative level, with technical specifications relegated to standard-making processes that accompany the mandate.⁷¹

China’s 2025 AI content labelling mandate, in contrast to the earlier 2023 “Deep Synthesis” regulation⁷², explicitly prioritises metadata-based provenance tracking as an essential mechanism for authentication and traceability of AI-generated content, moving beyond the previous implicit framing.⁷³ The new regulations require all qualifying AI-generated content to contain both visible labels and embedded metadata, with the latter detailing creator and content attributes directly within the file’s structure, thereby strengthening technical transparency and attribution throughout the content’s lifecycle.

Digital watermarking is formally encouraged as a recognised authentication tool—but the 2025 rules remain conspicuously silent on the relationship between watermarking and metadata provenance, such as whether service providers can meet compliance using only watermarking or whether metadata-based tracking is strictly mandatory.⁷⁴ No regulatory guidance is offered on possible interoperability or substitution between these two approaches, leaving it uncertain if a provider may rely entirely on watermarking-based authentication without maintaining robust metadata provenance tracking within files.

framework for modernization. For details, see the full work plan: https://www.gov.cn/zhengce/content/202505/content_7023697.htm (accessed 2025-07-31).

⁷⁰Art. 3, Measures for Labeling of AI-Generated Synthetic Content, defines two mandatory labeling types: explicit labeling—marks added to AI-generated or synthetic content in a manner easily perceptible to users (e.g., via text, audio, or graphics); and implicit labeling—technical markers embedded in the file data, not easily perceived by users. See: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm, translation at China Law Translate: <https://www.chinalawtranslate.com/en/ai-labeling/> (accessed 2025-07-31).

⁷¹China has developed a comprehensive and mandatory national standard for AI-generated content labeling (GB45438-2025), which is set to take effect alongside a legal mandate. This standard is supplemented by technical guidelines such as TC260-PG-20233A and TC260-PG-20252A, providing detailed specifications and implementation pathways. As the effective date nears, additional guidelines continue to be published, focusing on the technical realization of invisible labeling in metadata across text, image, audio, and video content, as well as on security and technical inspection procedures. In contrast, efforts to create a national digital watermarking technical standard remain provisional; current drafts are conceptual and under public consultation, not yet reaching the operational detail or enforceability of the content labeling standards.

⁷²The “Provisions on the Administration of Deep Synthesis Internet Information Services,” adopted on November 25, 2022, establish requirements for labeling synthetic media, managing technical standards, and clarifying legal obligations for internet information services that use deep synthesis technologies in China. See the official text: https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm. For an English translation, see China Law Translate: <https://www.chinalawtranslate.com/en/deep-synthesis/> (accessed 2025-07-31).

⁷³Art 5 MLAIGSC

⁷⁴Art. 5 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) encourages service providers to add forms of implicit labeling, such as digital watermarks, within generated synthetic content.

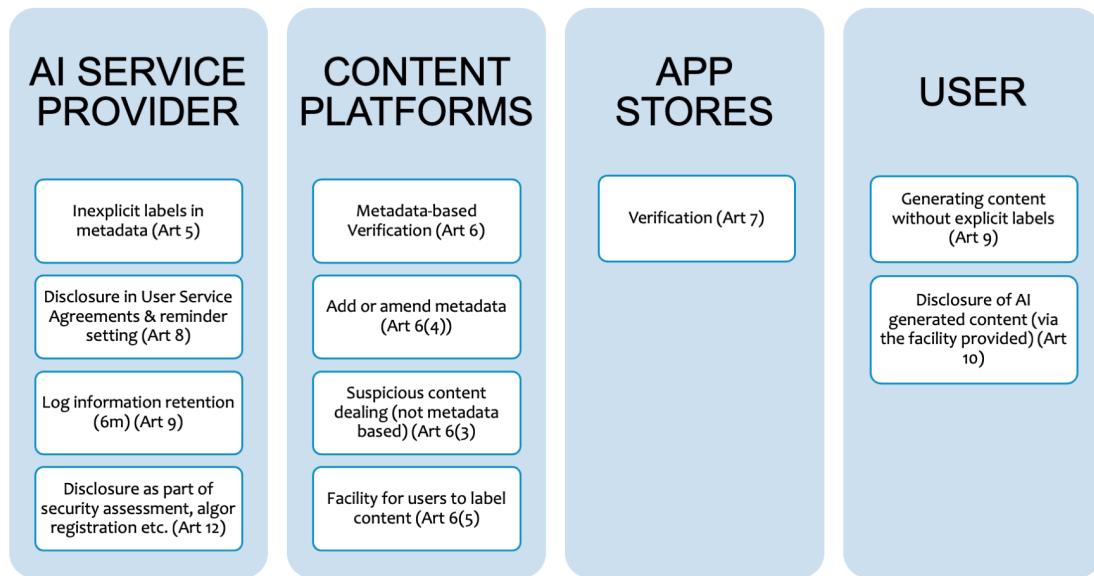


Fig. C1: Responsibility delineation across the content generation and dissemination chain under China’s content labelling mandate

Most notably, China’s AI content labelling mandate establishes a multi-layered system of attribution, assigning escalating obligations to all actors within the AI content pipeline—AI model developers, deployers/service providers, platforms, app stores, and end users—each interlocked to ensure continuous provenance and compliance.

AI service providers, including both developers and deployers, must embed both explicit (visible) and implicit (metadata) labels in all AI-generated content. The mandate bakes these requirements into their products “by design,” disclose labelling practices in user agreements, and maintain generation logs for regulatory traceability.⁷⁵ Platforms, serving as enforcement intermediaries, are mandated to detect, classify, and supplement labels for any AI-generated content they host or distribute, even where upstream attribution may be missing or incomplete, and must embed their own metadata to reinforce or restore provenance throughout the content lifecycle.⁷⁶ App stores and distribution platforms are required to vet generative AI tools for labelling compliance as a condition of approval, screening for functional content labelling before allowing distribution to users.⁷⁷ End users must not remove, tamper with, or obscure any explicit or implicit content labels, and must actively declare whether content uploaded to platforms is AI-generated, utilising platform-provided labelling functions; however, when users request AI-generated content without explicit (visible) labels, service providers may accommodate this only after clearly assigning the user explicit labelling obligations and usage responsibilities through user agreements, and must, in such cases, retain relevant user and transaction logs for at least six months

⁷⁵ Arts. 4–5, 8 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) require service providers to add explicit labels to generated synthetic content if their service falls under Article 17(1) of the Provisions on the Administration of Deep Synthesis Internet Information Services, and to embed implicit identifiers within content metadata as prescribed by Article 16 of those Provisions. Providers must also clearly describe the labeling methods and specifications in user agreements, ensuring users are properly informed of the requirements. Notably, China’s regulations use a broad definition of “service provider,” making no distinction between actor roles in the AI supply chain—unlike the EU AI Act, which differentiates between “providers” and “deployers.” As a result, a wide range of actors in China are subject to comprehensive transparency, safety, and accountability requirements regarding AI-generated content labeling.

⁷⁶ Art. 6 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) requires online content dissemination providers to regulate the spread of AI-generated synthetic content by: (1) checking file metadata for implicit markers and, if labeled as synthetic, adding prominent notices to alert the public; (2) if no marker is found but the user self-declares the content as synthetic, similarly adding prominent notices; (3) if neither marker nor user declaration exists, but explicit indicators or other signs of generated content are detected, treating the content as suspected synthetic and visibly notifying the public; (4) providing tools so users can declare synthetic content at publication. In cases (1)–(3), dissemination providers must also add information on content attributes, platform name/code, and content ID into the file metadata.

⁷⁷ Art. 7 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) stipulates that application distribution platforms must, during app listing or approval, require app service providers to state whether their applications offer AI-generated synthetic content services. If such services are provided, the distribution platform is responsible for verifying relevant documentation related to synthetic content labeling submitted by the app provider.

to ensure traceability and enforce accountability in line with regulatory requirements.⁷⁸ The reflects a systemic “chain of custody” approach to digital provenance and regulatory compliance, operationalised across all stages from content creation to distribution and end use, thus indicating a new standard of due diligence and potentially raised the bar for best practices.

A defining feature of China’s regulatory paradigm in AI content authentication is its unswerving – and eminently pragmatic – fidelity to metadata-based provenance. This stance is animated by a pointed institutional skepticism regarding the present reliability of avant-garde authentication technologies such as advanced digital watermarking. Consequently, the 2025 mandate vests metadata labelling with cardinal regulatory significance, privileging traceability and operational lucidity over the premature adoption of nascent, insufficiently vetted technical solutions. Yet, the regulatory framework eschews rigidity in favour of strategic openness; it expressly encourages research and experiment with watermarking, intimating a readiness to integrate such innovations as their robustness is empirically demonstrated and their congruence with overarching policy objectives assured. In effect, China’s approach secures immediate regulatory oversight and accountability through the institutionalisation of tried-and-tested metadata methods, while simultaneously preserving institutional latitude to adaptively incorporate more sophisticated mechanisms as technological circumstances evolve.

4 Chains of Accountability with Chinese Characteristics: Promises and Pitfalls

At a pivotal moment when the perils of provenance fragility and the proliferation of untraceable synthetic content threaten the structural integrity of digital trust, China’s regulatory experiment stands as both a crucible and a lodestar for global governance. By weaving together meticulous institutional oversight with a patchwork of technical and organisational safeguards, China signals both a wary pragmatism and a willingness to grapple head-on with the many-hands dilemma that vexes accountability in our era of generative AI. (Jing, 2017). Examining China’s evolving response is not merely instructive but essential, as its efforts to bridge the gaping chasm between regulatory aspiration and technical feasibility illuminate the formidable challenges—and latent possibilities—that confront all nations charting a course through this uncharted regulatory terrain.

In essence, China’s strategy for addressing the inherent weaknesses of metadata-based provenance tracking in the regulation of AI-generated content is distinguished not by a wholesale embrace of cryptographically secure traceability technologies, as epitomised by several US state-level statutes, but by the institution of an organisationally rigorous and multifaceted governance framework. Rather than assuming that technological solutions alone—such as digital watermarking or cryptographic signatures—can resolve the fragility of metadata, Chinese regulators compensate for these vulnerabilities through an elaborate regime of layered institutional and platform-level obligations. This organisational diligence is reinforced by regulatory penalties for non-compliance, persistent audits, and the ongoing development of national standards—all designed to preserve an unbroken chain of accountability even as content circulates across diverse platforms and jurisdictions. (Cobbe, Veale, and Singh, 2023b)

First, China imposes comprehensive and interdependent obligations on all actors in the AI content pipeline—including service providers, platforms, app stores, and users—requiring not only the embedding and maintenance of metadata labels but also continual verification and supplementation at each interface. Platforms and intermediaries are mandated to actively detect, restore, or even replace missing provenance information, using a combination of automated checks, manual audits, and platform-level interventions. Where technical fragility arises (such as the corruption or removal of metadata), regulatory pressure compels these organisations to implement systems for reinforcing and remedying gaps, ensuring that provenance is recoverable and traceability is sustained across the ecosystem. (Longpre, Mahari, Obeng-Marnu, Brannon, South, Gero, Pentland, and Kabbara, 2024a). Importantly, while this regime prioritises organisational diligence and platform capabilities as bulwarks against the fragility inherent in metadata systems, it also consciously reserves regulatory space for the maturation of novel technological solutions, such as digital watermarking, that may in time offer a more definitive fix to the limitations of metadata alone.

⁷⁸Arts. 9–10 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) require users who publish AI-generated synthetic content through online dissemination services to proactively declare and label such content using the provider’s labeling tools. If a user requests the service provider to issue synthetic content without explicit labeling, the provider may do so only after clarifying the user’s labeling duties and responsibilities via user agreement, and must retain related logs—including the identity of the recipient—for at least six months to comply with legal requirements.

Furthermore, this organisational scaffolding is reinforced by a robust enforcement regime: non-compliance with labelling or provenance obligations exposes actors to unspecified regulatory penalties, thereby incentivising diligence and institutional self-policing. Service providers must maintain user logs and supply evidence upon request, while persistent monitoring and compliance audits seek to close potential accountability vacuums.

China’s regulatory architecture also seeks to anticipate the many hands problem—where the diffusion of responsibility between numerous, diverse actors risks causing accountability vacuums. In response, the regulatory and technical approach aspires to “stitch together” an accountability chain that is seamless, even as content migrates across platforms and borders. Recognising the lack of cross-platform coordination, Chinese authorities proactively push for national and industry standards that aim to harmonise provenance protocols and lay groundwork for broader accountability solutions in the future. Therefore, unlike supply chain scholarship that mainly diagnoses how attribution evaporates as technical artefacts, decisions, or operational risks traverses a nested web of actors, each shielded by contracts, technical architecture and institutional boundaries(Nissenbaum, 1996)(Cobbe et al., 2023a), the China’s regime imposes mandatory, persistent labelling at every stage, backed by regulatory penalties for all involved, not just originators. This attempt at operationalising provenance and shared responsibility also starkly contrast with the more voluntary, sectoral, or upstream-focused approaches seen elsewhere, foregrounding both the radical ambitions and new challenges of engineering accountability into the entire lifecycle of AI-generate content.

Despite the organisational rigour and interlocking regulatory safeguards, China’s provenance regime is by no means free from blind spots. The framework exhibits limited consideration for the threat posed by malicious manipulation—where actors deliberately strip, alter, or falsify metadata and labels, subverting the system’s traceability and impairing detection, or where AI-generated content is disseminated without any labeling via offshore, unofficial, or encrypted channels(Wang, Yan, Zhang, and Zhang, 2021). These are precisely the scenarios where provenance is most critical, yet the regime’s reliance on metadata and platform-based controls may leave it ill-equipped to address.

Another significant blind spot in China’s regulatory framework for AI authentication is the challenge posed by cross-jurisdictional dissemination. When such content moves transnationally—passing from a Chinese platform into Western social networks, or vice versa—the liability structures and technical capacities underpinning provenance often break down, especially considering the rigidity of China’s emergent standards. The fragmentation is further exacerbated by a lack of global standardisation and coordination(Cihon, 2019a): metadata structures and authenticity conventions used in one legal environment may be unrecognised or unsupported in another, while visible indicators (such as overlays or icons) may be ignored or stripped at the point of import. These asymmetries have produced a fragmented global ecosystem for AI content authenticity, where regulatory effectiveness is constrained by the lack of mutual recognition, technical interoperability, and cross-border cooperation. In practice, national mandates cannot, in isolation, address these transnational governance deficits—a multilateral approach is indispensable. While it falls beyond the purview of this paper to undertake a detailed comparison, the juxtaposition of the emerging standards on AI content labelling put forth by China’s TC260 (of the National Cybersecurity Standardization Technical Committee, 2025) and those evolving in Western jurisdictions (Krog, 2025) amply illustrates both the breadth of regulatory ambition and the formidable challenges attendant to harmonising—or developing—an international standard capable of ensuring effective cross-border provenance and accountability for AI-generated content. Given the routine movement of digital content across borders, regulatory fragmentation constitutes a profound and persistent problem(Cihon, 2019b). As China’s experiment reveals, effective global AI governance must prioritise the mutual recognition and interoperability of provenance and authenticity standards, facilitate robust cross-border enforcement mechanisms, and nurture ongoing multilateral dialogue and consensus-building. International developments in this area have lagged the rapid proliferation of national mandates: coordinated solutions for aligning metadata standards, legal definitions, or technical protocols for AI content authentication across borders are still in their infancy. Despite being an apparent global governance problem, there remain scant initiatives within global or sub-global forums that systematically confront the legal and technical interoperability imperative.

While lauded for its thoroughness, China’s regime for AI content labelling may draw criticism for its rigid, prescriptive, one-size-fits-all approach to provenance. The framework and supporting standards mandate uniform technical methods—dual explicit and implicit labelling, prescribed content placements, and highly visible notices in all media types—leaving little to no room for alternative mechanisms or

adaptation to sector-specific needs. This rigidity is especially pronounced in the technical standards that underpin the regulatory rules, which are both detailed and operationally mandatory, shaping not just policy but the minutiae of compliant technical implementation. Further, China’s approach to AI-generated content labelling is marked by a lack of contextual nuance and flexibility. It mandates universal and visible disclosure of artificially generated content, offering little scope for discretion based on content type, intent, or potential for confusion among audiences. This stands in stark contrast to Article 50 of the EU AI Act, which explicitly considers the context and purpose of AI-generated content. For instance, Art 50(4) of the Act permits exceptions or more proportionate disclosure for “evidently artistic, creative, satirical, fictional or analogous” works, ensuring that mandatory transparency is implemented “in an appropriate manner that does not hamper the display or enjoyment of the work”. Unlike the EU’s model—which actively weighs context, intent, and audience expectations before mandating disclosure and allows for tailored implementation—China’s approach notably centres on the risk of “public confusion” as the core trigger for mandatory disclosure. Under Art 17 of China’s AI-generated content labelling mandate, platforms and providers must clearly label any synthetic content that could potentially “cause the public to confuse or misidentify its nature”. Thus, it hinges upon a more rigid and categorical assessment of when content might mislead or confuse the public, with limited space for interpretive discretion or situational nuance. By contrast, the EU’s approach leans towards a more nuanced, context-dependent assessment, enabling regulatory responses to be tailored to specific risks, audiences and environments. Further, the cumulative effect of ubiquitous labelling, provenance tracking, and log retention could produce chilling effects among users. If creators – particularly those working with AI tools in sensitive or contested cultural domains – fear misclassification, surveillance and regulatory sanction, they may begin to self-censor or withdraw content entirely. This is particularly true for hybrid content, where the line between AI-assisted and human-authored expression is internally negotiated and deeply contextual (Department, 2025). Pervasive governance carries the risk of flattening creative nuance and transforming AI-labelled content into a presumptively distrusted or stigmatised category, and regulators must balance the public interest in traceability with intellectual expression, especially in transmedia and user-generated environments (Treiger, 1989).

Lastly, there are other enforcement challenges rooted in the sheer scale, speed, and technical plurality of the AI content generation and dissemination pipelines. Ensuring consistent compliance across millions of user actions and a sprawling landscape of platforms necessitates not only detection infrastructure but human resources for auditing, nuanced legal interpretation, and unprecedented coordination across stakeholders. Although China has previously exhibited the capacity to orchestrate broad-scale regulatory interventions, the integrative complexity of AI-generated content—wherein metadata or labels can be trivially obfuscated, stripped, or manipulated—renders perfect enforcement out of reach. Compliance is thus likely to be selective, skewed toward the most visible actors: major tech platforms and prominent providers will be subject to close scrutiny and rigorous oversight, while content disseminated through informal, encrypted, decentralised, or cross-border networks will habitually escape detection. Further, social media and content sharing platforms—core enforcement nodes within this system—are tasked with proactively detecting and labelling AI-generated material. Yet, their detection capabilities are constantly tested by rapid technological advances, as well as by increasingly sophisticated evasion strategies employed by bad actors. These technical obstacles make enforcement a moving target, demanding ongoing research and development and the continual refinement of classifiers. Moreover, ramping up platform surveillance and detection capabilities can raise further issues regarding privacy, freedom of expression, and fairness. (Srinivasan, 2024). The resultant legal and social questions—of how detection systems operate, to what extent personal data is processed, and who bears responsibility for errors or overreach—remain unresolved. Moreover, enforcement is destined to be patchy, shaped by technical arms races and the persistent tension between regulatory ambition and the practical limits of control in a dynamic, borderless digital environment.

5 Conclusion

China’s evolving legal architecture, despite the imprint of selective borrowing from both European and American approaches, represents a distinctive and under-examined intervention whose implications for global AI governance are profound but not fully mapped. While China’s law is tailored to its domestic information ecosystem and administrative priorities, it catalyses a test case for global AI governance. Its ambition – operationalising provenance and collective responsibility at scale – exposes the tensions

between technological innovation, regulatory urgency, and compliance enforceability. As other jurisdictions weigh similar requirements, the Chinese approach offers both a template and a caution: technical mandates alone will not resolve the epistemic opacity and unpredictable agency that defines digital media ecosystem, and a truly robust accountability architecture remains inherently unfinished – persistently challenged by both the arms race of forgery and detection, and the institutional politics of transnational content flow. In other words, China’s AI content labelling mandate is not simply a regulatory milestone in the history of AI regulation, but a crucible wherein the limits and promise of engineered accountability are being tested – at the very forefront and on a global stage.

China’s AI content labelling mandate neither sets a ready-made blueprint for other countries nor proposes a universally desirable path. Yet, its pragmatic, organisationally intensive focus—especially on downstream distribution—poses reflections Western governments must not ignore. While liberal democracies often trust technological advancement and authentication tools, they tend to leave a vacuum around regulatory rules at the very moment those technologies mature. Moreover, the industry’s accelerated embrace of open-source models since DeepSeek redefined the competitive landscape further complicates the efforts to authenticate and trace the origins of AI-generated content as it is nearly impossible to enforce standardised authentication requirements at the model or developer level (Danen, 2025). China’s focus that heeds downstream dissemination thus emerges as a pragmatic response to these challenges.

This calculated move by China to foreground AI authentication signals not just a leap into the AI governance arena, parallel to the intensifying AI arms race, but functions as an useful reminder: The challenge is no longer just about inventing new safeguards but about ensuring that rules adapt alongside advances. As the legal and technological landscapes continue to evolve, the world’s response to China’s AI content governance experiment may define the next chapter in the global pursuit of attribution, accountability, and trust.

On the whole, China’s approach is unlikely to gain traction in societies that prioritise individual liberties and pluralism. Still, it brings to light important questions about how global governance might address the seamless spread of AI-generated content across platforms and borders. Rather than aligning perfectly with current debates on AI supply chain accountability, the Chinese strategy partly diverges, engaging more directly with the control and distribution of content on social platforms and challenging us to consider how law can evolve in the face of constant technological change.

6 Acknowledgment

The author would like to express our sincere gratitude to the participants of the Paris Conference for AI and Digital Ethics, as well as the attendees of the SSPS Research Seminar at the University of Lincoln, for their insightful feedback and valuable contributions to the development of this paper.

7 Declaration

The author declares that there is no conflict of interest with respect to the publication of this article.

References

- Adami M (2024) How ai-generated disinformation might impact this year’s elections and how journalists should report on it. <https://reutersinstitute.politics.ox.ac.uk/news/how-ai-generated-disinformation-might-impact-years-elections-and-how-journalists-should-report>, accessed: 2025-07-31
- Andrews C (2025) Virginia governor vetoes ai bill. <https://iapp.org/news/a/virginia-governor-vetoes-ai-bill>, published March 25, 2025. Accessed: 2025-07-31
- Bereskin C (2023) Parliamentary handbook on disinformation, ai and synthetic media. <https://www.cpaq.org/media/sphl0rft/handbook-on-disinformation-ai-and-synthetic-media.pdf>, accessed: 2025-07-31
- Burrus O, Curtis A, Herman L (2024) Unmasking ai: Informing authenticity decisions by labeling ai-generated content. *Interactions* 31(4):38–42. 10.1145/3665321, accessed: 2025-07-31

- California (2024) Sb-942 california ai transparency act (2023–2024). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942, approved by Governor September 19, 2024; Effective January 1, 2026; Accessed: 2025-07-31
- Cheney J, Chong S, Foster N, et al (2009) Provenance: a future history. In: OOPSLA '09: Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications, pp 957–964, 10.1145/1639950.1640064, accessed: 2025-07-31
- Chomanski B, Lauwaert L (2025) Automated propaganda: Labeling ai-generated political content should not be required by law. *Journal of Applied Philosophy* 42(3):994–1015. 10.1111/japp.70002, first published: 24 February 2025. Accessed: 2025-07-31
- Cihon P (2019a) Standards for ai governance: International standards to enable global coordination in ai research & development. Tech. rep., Future of Humanity Institute, University of Oxford, accessed: 2025-07-31
- Cihon P (2019b) Standards for ai governance: international standards to enable global coordination in ai research & development. *Future of Humanity Institute University of Oxford* 40(3):340–342
- Cobbe J, Veale M, Singh J (2023a) Understanding accountability in algorithmic supply chains. In: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, Chicago, IL, USA, pp 1186–1197, 10.1145/3593013.3594073, accessed: 2025-07-31
- Cobbe J, Veale M, Singh J (2023b) Understanding accountability in algorithmic supply chains. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp 1186–1197
- Commission UFC (2024) Fcc proposes disclosure rules for the use of ai in political ads. <https://www.fcc.gov/document/fcc-proposes-disclosure-rules-use-ai-political-ads>, notice of Proposed Rulemaking, FCC 24-74, Docket No. 24-211; Accessed: 2025-07-31
- Council ITI (2024a) Authenticating ai-generated content: Exploring risks, techniques & policy recommendations. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf, accessed: 2025-07-31
- Council ITI (2024b) Authenticating ai-generated content: Exploring risks, techniques & policy recommendations. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf, accessed: 2025-07-31
- Danen V (2025) The power of open collaboration: How open source is shaping the future of ai. <https://www.forbes.com/councils/forbestechcouncil/2025/01/03/the-power-of-open-collaboration-how-open-source-is-shaping-the-future-of-ai/>, published January 3, 2025. Accessed: 2025-07-31
- Deelman E, Berriman B, Chervenak A, et al (2009) Metadata and provenance management. In: Shoshani A, Rotem D (eds) *Scientific Data Management: Challenges, Technology, and Deployment*. Chapman and Hall/CRC, accessed: 2025-07-31
- Department HL (2025) Artificial intelligence and the creative double bind. *Harvard Law Review* 138(6):1585–1704. Chapter II.
- EU (2024) Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), art 50. <http://data.europa.eu/eli/reg/2024/1689/oj>, official Journal of the European Union L, 2024/1689, 12.7.2024; Accessed: 2025-07-31

- Feher K (2025) Generative AI, Media, and Society. Routledge, accessed: 2025-07-31
- Fisher SA (2024) Something ai should tell you – the case for labelling synthetic content. *Journal of Applied Philosophy* 42(1):272–286. 10.1111/japp.12758, first published: 18 August 2024. Accessed: 2025-07-31
- Gamage D, Sewwandi D, Zhang M, et al (2025) Labeling synthetic content: User perceptions of label designs for ai-generated content on social media. In: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp 1–29, 10.1145/3706598.3713171, accessed: 2025-07-31
- Guardian (2025) China calls for global ai cooperation days after trump administration unveils low-regulation strategy. <https://www.theguardian.com/technology/2025/jul/26/china-calls-for-global-ai-cooperation-days-after-trump-administration-unveils-low-regulation-strategy>, published: 26 July 2025. Accessed: 2025-07-31
- Hart TR, de Vries D (2017) Metadata provenance and vulnerability. *Information Technology and Libraries* 36(4):24–33. 10.6017/ital.v36i4.10146, accessed: 2025-07-31
- House UW (2025) Winning the race: America’s ai action plan. <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>, july 2025. Accessed: 2025-07-31
- Irwin K (2024) Openai adds labels to ai images, but says metadata ‘easily’ removed. <https://uk.pcmag.com/ai/150789/openai-adds-labels-to-ai-images-but-says-metadata-can-be-easily-removed>, published February 2, 2024. Accessed: 2025-07-31
- Jing Y (2017) The Transformation of Chinese Governance: Pragmatism and Incremental Adaption. *Governance: An International Journal of Policy, Administration, and Institutions* 30(1):37–43. 10.1111/gove.12231, 30th Anniversary Essay; First published 20 July 2016
- Jung Y, Hua P, Bao JA, et al (2025) Ai-generated or ai-modified? user reactions to labeling ai use in social media posts. In: CHI EA '25: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp 1–7, 10.1145/3706599.3720264, accessed: 2025-07-31
- Kapoor S, Narayanan A (2024) We looked at 78 election deepfakes. political misinformation is not an ai problem. <https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem>, accessed: 2025-07-31
- Krishna K, Song Y, Karpinska M, et al (2023) Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Neural Information Processing Systems, pp 27469–27500, published: 10 December 2023. Accessed: 2025-07-31
- Krog GP (2025) CEN-CENELEC JTC21 AI Standards: Complete Detailed Overview. [urlhttps://jtc21.eu/wp-content/uploads/2025/06/CEN-CENELEC-JTC21-AI-Standards-Complete-Detailed-Overview.pdf](https://jtc21.eu/wp-content/uploads/2025/06/CEN-CENELEC-JTC21-AI-Standards-Complete-Detailed-Overview.pdf), Comprehensive overview of the European AI standardization initiative and its alignment with the EU AI Act.
- Li W, Chen J (2024) From brussels effect to gravity assists: Understanding the evolution of the gdpr-inspired personal information protection law in china. *Computer Law Security Review* 54:105994. <https://doi.org/10.1016/j.clsr.2024.105994>
- Longpre S, Mahari R, Obeng-Marnu N, et al (2024a) Position: data authenticity, consent, & provenance for ai are all broken: what will it take to fix them? In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org, ICML'24
- Longpre S, Mahari R, Obeng-Marnu N, et al (2024b) Data authenticity, consent, and provenance for ai are all broken: What will it take to fix them? MIT Generative AI Initiative (MIT-GenAI)

10.21428/e4baedd9.a650f77d, published March 27, 2024. Accessed: 2025-07-31

- Mühlhoff R (2025) Chapter 11: Collective responsibility in the ethics of ai. In: *The Ethics of AI: Power, Critique, Responsibility*. Bristol University Press, 10.51952/9781529249262, accessed: 2025-07-31
- of the National Cybersecurity Standardization Technical Committee S (2025) Notice on the issuance of six cybersecurity standards practice guidelines, including the “method for identifying ai-generated synthetic content — implicit metadata identification for text files”. <https://www.tc260.org.cn/front/postDetail.html?id=20250828165129>, document No. [2025] 118, marked “Wangan Mizi”
- Nissenbaum H (1996) Accountability in a computerized society. *Science and Engineering Ethics* 2:25–42. 10.1007/BF02639315
- N.Y. (2024) Senate bill s7592a: Requires disclosure of the use of artificial intelligence in political communications. <https://www.nysenate.gov/legislation/bills/2023/S7592/amendment/A>, 2023–2024 Legislative Session, Sponsor: Senator Jake Ashby; Accessed: 2025-07-31
- Pan B, Stakhanova N, Ray S (2023) Data provenance in security and privacy. *ACM Computing Surveys* 55(14s):1–35. 10.1145/3593294, accessed: 2025-07-31
- Ren K, Yang Z, Lu L, et al (2024) Sok: On the role and future of aigc watermarking in the era of gen-ai. <https://arxiv.org/abs/2411.11478>, accessed: 2025-07-31, [arXiv:2411.11478](https://arxiv.org/abs/2411.11478)
- Sadasivan VS, Kumar A, Balasubramanian S, et al (2025) Can ai-generated text be reliably detected? <https://arxiv.org/abs/2303.11156>, 10.48550/arXiv.2303.11156, accessed: 2025-07-31, [arXiv:2303.11156](https://arxiv.org/abs/2303.11156)
- Scharowski N, Benk M, Kühne SJ, et al (2023) Certification labels for trustworthy ai: Insights from an empirical mixed-method study. In: *FAccT ’23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp 248–260, 10.1145/3593013.3593994, accessed: 2025-07-31
- Schick N (2020) *Deepfakes: The Coming Infocalypse*. Twelve, New York, accessed: 2025-07-31
- Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. *ACM SIGMOD Record* 34(3):31–36. 10.1145/1084805.1084812, accessed: 2025-07-31
- Srinivasan S (2024) Detecting ai fingerprints: A guide to watermarking and beyond. <https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/>, published January 4, 2024. Accessed: 2025-07-31
- Taylor I (2024) Collective responsibility and artificial intelligence. *Philosophy & Technology* 37. 10.1007/s13347-024-00718-y, accessed: 2025-07-31
- Treiger LK (1989) Protecting satire against libel claims: a new reading of the first amendment’s opinion privilege. *The Yale Law Journal* 98(6):1215–1234
- US A (2023a) Fact sheet: Biden-harris administration secures voluntary commitments from eight additional artificial intelligence companies to manage the risks posed by ai. <https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>, published September 12, 2023. Accessed: 2025-07-31
- US C (2023b) H.r.5586 - deepfakes accountability act, 118th congress (2023–2024). <https://www.congress.gov/bill/118th-congress/house-bill/5586/text>, introduced in the House of Representatives September 20, 2023; Accessed: 2025-07-31

- US S (2025) Assembly bill a6540a: Stop deepfakes act (2025–2026 legislative session). <https://www.nysenate.gov/legislation/bills/2025/A6540/amendment/A>, requires synthetic content creation system providers to include provenance data on generated or modified synthetic content; Accessed: 2025-07-31
- Wang X, Yan Z, Zhang R, et al (2021) Attacks and defenses in user authentication systems: A survey. *Journal of Network and Computer Applications* 188:103080
- Washington (2025) Hb 1170 - 2025-26: Informing users when content is developed or modified by artificial intelligence. <https://app.leg.wa.gov/billsummary?BillNumber=1170&Initiative=false&Year=2025>, requires providers of generative AI systems to include detection tools and enable disclosure options for AI-generated or modified content; referred to House Rules Committee as of January 31, 2025; Accessed: 2025-07-31
- Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al (2023) Testing of detection tools for ai-generated text. *International Journal for Educational Integrity* 19. 10.1007/s40979-023-00146-z, published: 25 December 2023. Accessed: 2025-07-31
- Werder K, Ramesh B, Zhang RS (2022) Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems* 13(2):1–23. 10.1145/3503488, published: 10 March 2022. Accessed: 2025-07-31
- Wittenberg C, Epstein Z, Berinsky AJ, et al (2024) Labeling ai-generated content: Promises, perils, and future directions. *An MIT Exploration of Generative AI: From Novel Chemicals to Opera* 10.21428/e4baedd9.0319e3a6, published March 27, 2024. Accessed: 2025-07-31
- Xu R, Hu M, Lei D, et al (2024) Invismark: Invisible and robust watermarking for ai-generated image provenance. <https://arxiv.org/abs/2411.07795>, accessed: 2025-07-31, [arXiv:2411.07795](https://arxiv.org/abs/2411.07795)
- Łabuz M, Nehring C (2024) On the way to deep fake democracy? deep fakes in election campaigns in 2023. *European Political Science* 23:454–473. 10.1057/s41304-024-00482-9