# Why AI-Generated Content Labeling Mandates Fall Short

By Justyna Lisinska and Daniel Castro | December 16, 2024

**As artificial intelligence (AI) tools become better at creating high-quality content—including text, images, audio, and video—critics worry about potential misuses, such as to spread misinformation, perpetrate fraud, violate intellectual property (IP) rights, and create harmful deepfakes. Some policymakers are proposing a requirement for mandatory labels on all output generated by AI systems so users can distinguish between human-generated and AI-generated content. However, mandatory labeling, particularly through watermarking, is neither a reasonable nor effective solution to the issues policymakers seek to address. Rather than singling out AI-generated content, policymakers should prioritize building trust within the digital ecosystem as a whole.**

## INTRODUCTION

Generative AI (GenAI) enables users to produce high-quality digital content such as images, text, music, and video. This technological advancement has enriched many creative possibilities, boosted workers' productivity, and offered new tools for innovation.[1] Most output from GenAI systems is beneficial and harmless, but some policymakers are concerned about the technology's potential misuse, including to spread disinformation through fabricated content, violate IP rights from AI-generated imitations of existing works, and create harmful deepfakes, such as impersonations used to perpetuate fraud or exploitative content such as unauthorized AI-generated nudes of individuals.[2]

Policymakers have called for mandatory labeling of all AI-generated content; however, this approach has serious limitations. While labeling AI-

generated content, particularly through watermarking, may help users identify some AI-generated material, requiring it for all AI-generated content would be impractical and ineffective because of diverse content, limited resilience to manipulation, and varying regulatory requirements. More importantly, doing so would fall short in addressing policymakers' primary concerns, namely disinformation, IP rights violations, and deepfakes.[3]

This report begins by outlining the main approaches to labeling AI-generated content and then highlights key AI labeling regulations and initiatives from around the world. Following this, the report examines why mandatory labeling, despite its appeal to certain policymakers, is not a good policy option. Finally, it emphasizes the importance of enhancing transparency for all content—whether human- or AI-created—and developing targeted strategies to address the malicious use of GenAI. Instead of mandating technically complex and permanent labels on AI-generated content, this report proposes promoting voluntary labels for all online content through established standards such as the Coalition for Content Provenance and Authenticity (C2PA).[4]

The report provides several recommendations for policymakers to strengthen trust in digital content:

1. Encourage voluntary adoption of adding labels for all digital content through an established industry standard such as C2PA, which embeds cryptographically secure metadata.

2. Launch digital, AI, and media literacy campaigns for users to assess digital content's authenticity and trustworthiness and make informed decisions about the content they consume.

3. Develop targeted responses to problems such as misinformation, IP rights violations, and deepfakes, rather than broadly labeling AI content.

## THREE MAIN APPROACHES TO LABELING AI-GENERATED CONTENT

Digital watermarking, digital fingerprinting, and using cryptographic metadata are three approaches to labeling and verifying AI-generated content. While offering certain benefits, each approach also has limitations that highlight the complexity of reliably identifying AI-generated media in today's digital ecosystem.

### Watermarking for AI

This technique involves embedding a distinct and unique signal, known as a watermark, into the output generated by an AI model (such as text, audio, images, or video). The watermark can be made visible or invisible.[5] The most effective methods for watermarking AI-generated content use invisible signals, detectable only by specialized software. Detecting

whether an invisible watermark exists in a piece of content usually requires using a proprietary tool designed for that specific watermarking system.

Digital watermarking is limited by how much data can be added without reducing content quality. Deliberate or unconscious manipulation of content can also remove the watermark.[6] The techniques used to watermark AI-generated content vary depending on the medium. Consider some of the options:

- **Text:** Uses certain word choices in a block of text.
- **Image:** Embeds invisible data within image pixels.
- **Audio:** Incorporates imperceptible signals in audio tracks.
- **Video:** Hides data within video pixels in every frame.

Unfortunately, none of these techniques are capable of withstanding advanced methods to remove the watermark.[7]

### Digital Fingerprinting

Digital fingerprinting works by generating a unique code (a fingerprint) based on the content itself (e.g., pixels, video frames, text, or audio waveforms, and linking this code to information about the content), such as whether it was produced by AI, the date it was produced, or who produced it. Users can create fingerprints of content they encounter and check to see whether they exist in trusted databases.[8] Digital fingerprinting may fail if the content is substantially altered.[9]

### Cryptographic Metadata

Metadata refers to data about content, such as its creation date and creator.[10] Metadata is often embedded within the files used for common forms of media, such as images, audio, and video. Cryptographic metadata uses cryptographic techniques to secure this information to ensure the integrity and authenticity of the information. Users can check the cryptographic metadata of content they encounter. However, metadata can be unintentionally or deliberately removed.[11]

## THE PUSH FOR LABELING AI-GENERATED CONTENT

The purpose of labeling is to allow users to detect AI-generated content to prevent fraud and deception. It aims to stop the spread of misinformation and disinformation, identify harmful deepfakes, and detect AI-generated content that is trying to be passed off as human-created.

Policymakers worldwide have proposed various forms of mandatory AI labelling, often through watermarking requirements. The following are key regulations and initiatives:

## European Union

The AI Act, formally adopted by the EU in March 2024, requires providers of AI systems—including those creating synthetic audio, image, video, or text content—to label AI-generated or altered outputs in a machine-readable format. Providers must use "effective, interoperable, robust, and reliable" technical solutions, taking into account the unique characteristics of each content type, implementation costs, and the latest technical standards.[12] Examples of technical solutions include watermarks, metadata tags, cryptographic methods for verifying content provenance and authenticity, logging mechanisms, and digital fingerprints.[13]

## United Kingdom

The Artificial Intelligence Bill, introduced by the House of Lords in May 2024 under the Conservative government, sought to establish regulations for AI. Although Parliament did not pass the legislation, it included provisions requiring businesses supplying AI-powered products or services to provide clear and unambiguous labeling, including health warnings and opportunities for users to give or withhold informed consent.[14]

## China

The Regulation on the Management of Deep Synthesis of Internet Information Services, which passed in January 10, 2023, mandates that service providers watermark AI-generated content, including text, images, and videos.[15] It requires companies using "deep synthesis" technology (e.g., AI-generated videos, voices, or images) to add a visible mark (watermark) to the content they create or edit.[16]

## United States

There are both federal and state policies around labeling AI in the United States, including the following:

- **President Biden's Executive Order:** President Biden's 2023 executive order on AI mandates that the Department of Commerce create guidelines for labeling AI-generated content. AI companies will use these guidelines to develop labeling and watermarking tools, which the White House aims for federal agencies to adopt.[17]
- **OMB M-24-18**: In September 2024, the U.S. Office of Management and Budget (OMB) issued Memorandum M-24-18, titled "Advancing the Responsible Acquisition of Artificial Intelligence in Government," which specifies that federal agencies purchasing tools to use enterprise-wide must require vendors to implement watermarks, cryptographic metadata, or other similar technical solutions in order to identify the content as AI-generated, link it to the specific model used to create the content, and allow tracing of its origin and edits.[18]

- **California AI Transparency Act:** This legislation, signed into law by Governor Newsom in September 2024 and going into effect on January 1, 2026, requires providers with over one million monthly users to offer three main services:
  - **Free detection tool:** Users can upload content (e.g., images, videos, or audio) to check if it was generated or altered by the provider's AI system.
  - **Visible watermark:** Users have the option to add a clear, visible, and permanent label to any AI-generated content, thereby indicating that it was created or modified by AI.
  - **Invisible watermark:** Every piece of AI-generated content will include a hidden, unremovable label (i.e., a watermark). This label will contain the provider's name, the AI system version, the date and time of creation or modification, and a unique identifier.[19]
- **Advisory for AI-Generated Content Act**: Senator Pete Ricketts (R-NE) introduced S. 2765 in 2023, legislation that requires providers to include a watermark on all AI-generated material, indicating that the content was created by an AI system. The bill, which did not advance out of committee, would have required the Federal Trade Commission to develop and enforce AI watermark standards.[20]

## LIMITATIONS OF WATERMARKING AND LABELING APPROACHES

Watermarks cannot reliably identify AI-generated content for a number of reasons. First, they are easily stripped away, particularly in adversarial settings where actors actively seek to evade detection. Second, inconsistent international laws on watermarking mean that AI-generated content from countries without labeling rules can be circulated without any labels. Bad actors can also use unregulated or open source AI models to produce AI-generated output without watermarks. In this way, the absence of a watermark offers no assurance that content is human made.

Labeling AI-generated content does not address policymakers' more fundamental concerns, such as misinformation, IP rights violations, and harmful deepfakes. The emphasis on detecting AI-generated content with watermarks risks creating a misleading divide between AI-generated and human-created content, overlooking that both sources can produce either beneficial or harmful material.

### Technical Limitations

Watermarking AI-generated content is often seen as a way to ensure that users can always distinguish between human-created and AI-generated content, yet it faces significant technical challenges that limit its

effectiveness. Effectively watermarking AI-generated content remains challenging due to two main reasons:

1. **Vulnerability to manipulation:** Watermarks, whether visible or invisible, are not tamper-proof.[21] Visible watermarks can be easily removed with basic editing, while invisible watermarks—though more resilient—are still susceptible to degradation or removal through advanced techniques.[22]

2. **Lack of standardization:** Since each company can use a unique watermarking approach tied to its proprietary models, cross-platform verification can be challenging. Users need different tools to verify content generated by different companies, complicating the detection process and reducing accessibility. Additionally, without standardized approaches, there could be variations in how accurately and consistently watermarks are detected across platforms.

## Issues Beyond AI-Generated Content

Many of the concerns, such as misinformation, IP rights violations, and content manipulation, that motivate proposals to require labeling of AI-generated content are not exclusive to AI-generated content. Labelling requirements also do little to address the underlying factors that cause these issues.

### Misinformation and Disinformation

Misinformation (unintentional inaccuracies) and disinformation (deliberately false content) can originate from content produced by either AI or humans. Labeling AI-generated content fails to address the deeper causes of misinformation and disinformation, such as individual tendencies to share unverified content. Additionally, labeling AI-generated content does not provide a mechanism for holding users accountable for spreading false or harmful information, nor does it stop them from distorting or manipulating factual content. Without tackling these deeper issues, labeling risks becoming a superficial solution to a far more complex problem.

### IP Rights Violations and Misrepresentation

Labeling AI-generated content to make it distinguishable from human-generated content does not prevent IP rights violations from occurring. For example, if someone uses AI to create an image that closely resembles someone else's copyrighted work without permission, this AI-generated content may be an instance of copyright infringement. But this infringement is no different than if a human artist manually creates content that violated IP law.

Mandatory labeling requirements would also not address concerns that some people will misrepresent AI-generated content as their own work,

such as students passing off AI-generated essays as their own or creators misrepresenting the extent to which they made original contributions to a final work. In each of these cases, there is no reason to distinguish between those individuals who used AI and those who did not, or to expect any solutions that only focus on AI tools to address the underlying motivations for this type of deception.

### Deepfakes

Watermarking deepfakes may help identify some AI-generated audio, photos, and videos, but it does not address the malicious intent behind their creation and spread, especially in political contexts. In the political sphere, bad actors may deploy deepfakes to falsely show public figures saying or doing things they never did, with the aim of swaying public opinion or influencing elections. Even with a watermark identifying the video as AI-generated, these individuals can still spread it to mislead viewers, relying on the likelihood that many will overlook or misunderstand the watermark. Similarly, when the goal of a deepfake is to embarrass or harass someone, even if the watermark successfully alerts most people that the content is generated by AI, the deepfake may still have its intended effect. Moreover, using GenAI is not the only way to produce deceptive media that appears realistic. Individuals can manually create deceptive media, such as a voice impersonator recording fake audio, or use non-AI digital tools, such as photo editing software to create misleading images.

## Misleading Distinction Between AI and Human Content

The focus on AI labeling creates a false distinction between AI-created and human-created content, ignoring the fact that human-created content can be harmful and AI-created content can be harmless. This approach may unintentionally bias audiences against AI-generated content, prompting them to question its reliability purely based on its origin rather than assessing its actual credibility.

For instance, someone might use an AI system to generate a well-researched and accurate report, but an "AI-generated" label could lead readers to refuse to trust the report, assuming it is unreliable solely because it was created by AI. Conversely, human authors could produce misleading or biased articles, yet their work might be trusted simply because it lacks an AI label. This double standard does little to address the real issue: evaluating content on its merit, not its method of creation.

## International Limitations

Certain countries may require watermarking to label AI-generated content, while others may impose no such rules, thus creating a fragmented landscape that undermines global transparency efforts. AI-generated material from jurisdictions without these mandates can easily circulate internationally without labels, blurring the lines between AI- and human-

created content. This inconsistency poses challenges for monitoring and distinguishing content across borders.

For instance, on social media, videos labeled as "AI-generated" may originate from countries with strict watermarking regulations. Meanwhile, similar AI-generated videos from countries without such rules might lack any labels, leading viewers to mistakenly believe that only labeled content is AI generated. This discrepancy not only misguides audiences but also weakens attempts to build trust in AI-generated media.

## POLICY RECOMMENDATIONS

Policymakers should focus on building trust in the digital ecosystem as a whole rather than singling out AI-generated content. Instead of trying to mandate unremovable labels, which are technically unfeasible for all harmful content, policymakers should encourage the use of voluntary labels for safe content. Removing labels from harmful content can put people at risk, as they may unknowingly believe or rely on it. In contrast, removing labels from safe content poses minimal harm to people.

### Recommendation 1: Adopting Trust Signals for All Digital Content

Policymakers should promote adoption of voluntary solutions that allow users to check the origin and history of digital content, regardless of whether it was created by AI or humans. Standards such as C2PA—which have drawn the support of major industry players including tech companies such as Adobe, Google, Meta, and Microsoft, publishers such as BBC, Financial Times, and Getty Images, and device makers such as Canon and Nikon—allow creators to embed cryptographically secure metadata in digital content to attest to its authenticity.[23] By implementing these standards, content creators, platforms, and end users can maintain a chain of custody that reveals each modification and preserves context. This transparency fosters trust, as it assures users that they are seeing the content as intended and any edits are documented.

**Government agencies should adopt the C2PA standard in their digital media to promote its use and instill public confidence.** For example, by embedding cryptographic metadata in official publications, government bodies can assure the public of the authenticity of these materials, countering potential misinformation. Government adoption of C2PA would also act as a powerful signal to the public and private sectors alike, showing a commitment to transparency. Individuals engaging with government-published content that includes cryptographic metadata would know that it has not been manipulated by a third party, which would strengthen the public's ability to trust digital content distributed online from government sources.

### Recommendation 2: Launch Digital, AI, and Media Literacy Campaigns

**Policymakers should launch digital, AI, and media literacy campaigns** to equip users of all ages with the skills needed to assess the authenticity of digital content, recognize provenance information, and understand why content transparency matters. These campaigns should focus on empowering users to make informed decisions about the trustworthiness of what they see online, enabling them to navigate digital media landscapes with a critical eye and helping to reduce the influence of manipulated information. It should also teach them how to protect themselves and use AI tools responsibly, such as how to seek help if they encounter deepfakes of themselves, how to be aware of AI-enabled phishing attacks, and how to properly use AI tools in academic settings.

Collaboration with tech companies, media platforms, and digital literacy advocates will be critical to these efforts. Through partnerships with media platforms and tech firms, these campaigns can drive the adoption of content authenticity standards, ensuring that end users understand how verification processes work and why they are important, as well as their limitations. Engaging with the platforms will also boost the visibility and accessibility of these tools.

### Recommendation 3: Create Targeted Solutions for Specific Problems

**Policymakers should develop targeted solutions for specific problems rather than mandate labeling of AI-generated content**. From misinformation to deepfake abuse, each issue requires a nuanced approach that addresses the underlying risks without one-size-fits-all measure.

For misinformation and disinformation, digital literacy initiatives, content provenance standards, and platform-driven efforts should be prioritized to combat misleading information. These solutions should focus on equipping users with skills to critically evaluate content, regardless of whether it was AI generated or human made, helping to mitigate the spread of misinformation and disinformation at its source.

In the case of IP rights violations, rather than creating new labeling requirements exclusively for AI-generated content, policymakers should instead enforce existing IP laws for all content. By strengthening enforcement, creators can better defend their work from unauthorized reproductions and misuse, thereby preserving the integrity of their IP.

Deepfakes present a unique and high-stakes challenge that requires specific approaches based on context. For example, policymakers should require political campaigns to disclose whenever they knowingly distribute materially deceptive media that could harm a candidate's reputation or mislead voters. These rules should apply equally to any materially deceptive media, regardless of whether it was created by AI.[24] In other contexts, such as the distribution of fake sexually explicit images or videos

of someone, creating and enforcing anti-revenge porn laws that cover fake media will help curb this harmful behavior.

Each of these targeted solutions aims to address the specific dynamics of digital misinformation, IP rights violations, and deepfake abuse, creating a safer and more trustworthy digital environment.

## CONCLUSION

Mandatory labeling of AI-generated content through watermarking is not the right solution and would not prevent the main concerns that have motivated these proposals by policymakers. A holistic approach to building trust in the entire digital ecosystem is necessary, with targeted solutions for specific problems. Policymakers should adopt trust signals, support content provenance standards, and improve digital literacy to create a more effective and robust strategy for tackling harmful content online.

## REFERENCES


1. Eric Zhou and Dokyun Lee "Generative artificial intelligence, human creativity, and art," *PNAS Nexus*, vol. 3 2024: 3, https://doi.org/10.1093/pnasnexus/pgae0528; Daniel Sack and Lisa Krayer, "GenAI Doesn't Just Increase Productivity. It Expands Capabilities" (Boston Consulting Group, September 5, 2024), https://www.bcg.com/publications/2024/gen-ai-increases-productivity-and-expands-capabilities.
2. Matt Burgess, "Millions of People Are Using Abusive AI 'Nudify' Bots on Telegram," *Wired*, October 15, 2024, https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/; Kat Tenbarge, "Fake news YouTube creators target Black celebrities with AI-generated misinformation," *NBC News*, January 30, 2024; Jesse Damiani, "A Voice Deepfake Was Used To Scam A CEO Out Of $243,000," *Forbes*, September 3, 2019, https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.
3. Daniel Castro, "Opportunities for APEC to build trust in the digital economy" (Information Technology and Innovation Foundation (ITIF), August 20, 2024), http://itif.org/publications/2024/08/20/opportunities-for-apec-to-build-trust-in-the-digital-economy/.
4. "How it works," accessed November 15, 2024 https://contentauthenticity.org/how-it-works
5. Justyna Lisinska and Daniel Castro, "The AI Act's AI Watermarking Requirement Is a Misstep in the Quest for Transparency" (CDI, July 9, 2024), https://datainnovation.org/2024/07/the-ai-acts-ai-watermarking-requirement-is-a-misstep-in-the-quest-for-transparency/.
6. "Researchers Say Current AI Watermarks Are Trivial to Remove," University of Maryland, October 12, 2023, accessed November 13, 2024, https://www.cs.umd.edu/article/2023/10/researchers-say-current-ai-watermarks-are-trivial-remove.
7. "SynthId," accessed December 6, 2024, https://deepmind.google/technologies/synthid/.
8. "Digital Fingerprint for Content Verification Explained," Score Detect, February 5, 2024, accessed November 24, 2024, https://www.scoredetect.com/blog/posts/digital-fingerprint-for-content-verification-explained.
9. Andy Parsons, "Durable Content Credentials," Content Authenticity Initiative (CAI), April 8, 2024, https://contentauthenticity.org/blog/durable-content-credentials.
10. Gary Kranz, "Metadata," *TechTarget*, https://www.techtarget.com/whatis/definition/metadata.
11. Andy Parsons, "Durable Content Credentials," *Content Authenticity Initiative*, April 8, 2024, https://contentauthenticity.org/blog/durable-content-credentials.
12. Regulation (EU) 2024/1689 of The European Parliament and of The Council of 13 June 2024.
13. Ibid.

14. UK Parliament, “Artificial Intelligence (Regulation) Bill [HL],” https://bills.parliament.uk/publications/53068/documents/4030.

15. China Law Translate, “Provisions on the Administration of Deep Synthesis Internet Information Services,” November 25, 2022, https://www.chinalawtranslate.com/en/deep-synthesis/.

[16] Ibid.

17. The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

18. Shalanda D. Young, “Advancing the Responsible Acquisition of Artificial Intelligence in Government,” Memorandum For The Heads Of Executive Departments And Agencies, September 24, 2024, M-24-18-AI-Acquisition-Memorandum.pdf.

19. Senate Bill No. 942, chapter 291 (2024).

20. S.2765 - Advisory for AI-Generated Content Act (2023).

21. “Researchers Say Current AI Watermarks Are Trivial to Remove,” University of Maryland.

22. Justyna Lisinska, “Audio Watermarking Won’t Solve the Real Dangers of AI Voice Manipulation” (CDI, October 18, 2024), https://datainnovation.org/2024/10/audio-watermarking-wont-solve-the-real-dangers-of-ai-voice-manipulation/; Justyna Lisinska, “Why Watermarking Text Fails to Stop Misinformation and Plagiarism” (CDI, September 18, 2024), https://datainnovation.org/2024/09/why-watermarking-text-fails-to-stop-misinformation-and-plagiarism/; Justyna Lisinska, “Watermarking in Images Will Not Solve AI-Generated Content Abuse” (CDI, August 15, 2024), https://datainnovation.org/2024/08/watermarking-in-images-will-not-solve-ai-generated-content-abuse/.

23. “Membership,” Coalition for Content Provenance and Authenticity, 2024, https://c2pa.org/membership/.

24. Daniel Castro, “Testimony to the Alaska State Senate Regarding AI, Deepfakes, Cybersecurity, and Data” (ITIF, February 2, 2024), https://itif.org/publications/2024/02/02/testimony-to-the-alaska-state-senate-regarding-ai-deepfakes-cybersecurity-and-data-transfers/.

## ABOUT THE AUTHORS

Justyna Lisinska is a policy analyst at the Center for Data Innovation. Previously, she served as a Policy Research Fellow at King's College London, where she developed a policy program for the United Kingdom's largest project on autonomous systems. She also has experience working within the government and with government officials. Justyna holds a Ph.D. in Web Science from the University of Southampton.

Daniel Castro is the director of the Center for Data Innovation and vice president of ITIF. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

## ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

**Contact: info@datainnovation.org**

**datainnovation.org**