What Lies Ahead for Generative AI Watermarking

Anonymous Authors[1]

Abstract

This position paper outlines the challenges that watermarking faces in ensuring the transparency and traceability of generative AI. While the robustness is seen as the main technical challenge, we argue that there are more significant unanswered questions such as how to use and trust the detection outcomes or how to ensure interoperability between actors. These crucial questions are currently missing from the watermarking community and from the public discourse.

1. Introduction

Digital watermarking conceals information directly into the content itself, *e.g*. in pixels of an image. Watermark decoders or detectors are specific algorithms extracting the watermark signal even if the content has been modified to some extent. It is a mature technology that remains unknown because its first requirement is imperceptibility. Millions of people are daily exposed to watermarked content in: photos of the news industry (web or print) to identify the source photo agency; audio and video of Digital Cinemas or Video On Demand portals to combat piracy; or audio of TV broadcasts for audience measurement.

Meanwhile, identifying content provenance is increasingly important since AI-generated content is used for swaying public opinion, fraud, or impersonation at a higher scale and more convincingly than even authentic content (Spitale et al., 2023). Governments are getting hold of the issue through new regulations that impose watermarking as a technical means for transparency and traceability (USA, 2023; Chi, 2023; Eur, 2023). For instance, the Californian act on watermarking would require model providers to "*place imperceptible and maximally indelible watermarks containing provenance data into synthetic content*" (California State Leg., 2024). Several key players, like Google, Meta, and OpenAI, have already started applying it at scale. For many, the primary concern is robustness: malicious actors might attempt to remove the watermark. This issue is frequently cited as the barrier to implementing watermarking for detection purposes (Christodorescu et al., 2024; Knibbs, 2023; Harris & Norden, 2024). Improving robustness is challenging, but it should not be a case of not seeing the forest for the trees.

2. Why Watermarking is Robust Enough

2.1. Methods for content provenance

Passive methods. *Forensics* methods employ binary classifier spotting hidden traces of generated content (Mitchell et al., 2023; Corvi et al., 2023). They already fall short in effectiveness compared to watermarking techniques (Sadasivan et al., 2023; Saberi et al., 2024), and will get worse as generative models get bettter.

Fingerprinting (or copy detection) stores hashes of all the AI-generated content in a database, which is then queried to determine whether a piece of content is a copy of a registered generation. This problem is not solved at scale, especially in an adversarial setting (Douze et al., 2021).

Active methods. *Visible watermarks* are straightforward and widely recognized, but they are also easy to remove or tamper with (Dekel et al., 2017), making them less reliable.

Cryptographic metadata embed digital signatures and certificates within the metadata. The Coalition for Content Provenance and Authenticity (C2PA) and the International Press Telecommunications Council (IPTC) have recently proposed two standards. The upside is that forging fake cryptographic signatures is extremely hard, however the metadata are often removed during re-uploads or screenshots. A study by Imatag (2018) shows that only 3% of images on the Internet come with copyright metadata.

2.2. Main advantages of invisible watermarking

First, it is an active method of intentionally injecting traces into content, whence the greater robustness than that of passive methods like forensics (Sadasivan et al., 2023) or fingerprinting (Fernandez et al., 2023c). Removing the watermark is always possible, but this always damages the quality, contrary to visible watermark or metadata erasure.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Second, a sound watermarking design has a low false positive rate. Most importantly, it is provably low (Fernandez et al., 2023a), unlike with forensics and fingerprinting, where the rates are empirically measured. Data provenance detection will soon be tested on millions of pieces of content therefore requiring extremely low false positive rates. This is beyond reach of an empirical validation.

Third, watermarking is versatile, coping with different types of content like LLM-generated text (Kirchenbauer et al., 2023; Aaronson & Kirchner, 2023), audio (O'Reilly et al., 2024; San Roman et al., 2024), or image (Zhu et al., 2018; Wen et al., 2023).

Clarification of the terminology. *Robust* watermarking is designed to withstand alterations, making it suitable for verifying the AI-generated nature of content. However, it is not intended for integrity verification, as it does not break when the content is modified. On the contrary, *fragile* watermarking reveals tampering (Wong, 1998) and is more suitable for authenticating real content, but not for traceability because of its lack of robustness.

2.3. Limitations

Attacks. Watermarking is not foolproof (Sadasivan et al., 2023; Saberi et al., 2024; Jovanović et al., 2024). Attacks are roughly categorized based on the attacker's knowledge: *white-box*, full access to the watermarking algorithm and its parameters (*e.g.* model weights); *black-box*, access only to input and output for instance through an API; *no-box*, no knowledge of the system. The effectiveness of an attack generally increases with the attacker's level of knowledge about the watermarking system (San Roman et al., 2024). There is always a trade-off to consider: an attack may succeed in removing or forging a watermark, but at the cost of degrading the quality of the content itself and of making the attack more detectable.

Whom do we defend against? Most mafia organizations or belligerent countries now have the expertise and resources to train their own generative models. They will include neither watermarking nor metadata. Forensic methods are also doomed to fail due to the lack of such data to train a detector. The aim is rather to dissuade 99% of the population, by making the removal of the watermark complex enough and voluntary – or even criminal by law, as what happened with DRM systems (Wikipedia, 2024). This aligns with the motto "keep honest people honest," which Hollywood popularized in the 2000s about DRMs.

3. The Real Challenges Ahead

Who controls watermarking? While everybody is a priori willing to know when they are interacting with generated content, making watermark detectors publicly available introduces security risks. Open-source detectors can lead to white-box attacks, and API access can facilitate black-box attacks. Currently, no record of watermark detection by anyone other than the generative model's owner exists. This situation, where the model provider is both judge and jury, is problematic. It would be more trustworthy if watermarking and detection were managed by trusted entities. This raises questions about who these entities should be and how they are governed.

Open-source generative models? They present a unique challenge since they are freely available and usable without post-hoc watermarks (applied after generation). Ideally, models should be trained or fine-tuned to generate watermarked content natively as in (Fernandez et al., 2023b). Determining responsibility in this context is complex – should it be the responsibility of the individual who uploads a model to a platform, or should hosting platforms like GitHub or Hugging Face enforce in-model watermarking? This issue needs clear regulatory guidance and possibly new technological solutions to ensure compliance.

What to do with detection? This question is not clearly addressed by current regulations. The use of watermarks for labeling authentic or fake content on social networks and search engines, as suggested by current texts like 22949.90.3.(a) of (California State Leg., 2024), may lead to a rebound effect. It may conversely exacerbate misinformation by placing undue emphasis on content that is either not detected, generated by unknown models, or authentic but used out of context.

Moreover, detection of watermarks extends beyond individual pieces of content, often involving the aggregation of evidence from multiple submissions linked to a single account. Kirchenbauer et al. (2024) notably showed that watermarked text may be detected even under strong paraphrasing after observing enough words.

Finally, different entities will use different watermarks, leading to a fragmented ecosystem where nobody is responsible for detection. For instance, the music generation startup Suno (2024) watermarks their outputs, but no platforms (Facebook, X, Spotify, Youtube, etc.) actually detect them. Collaborative efforts are needed to establish standards that ensure watermarks are robust, but, most importantly, recognizable across platforms. It should involve regulators, model providers and content hosting platforms.

4. Conclusion

Watermarking is the most viable technology to improve transparency and traceability in AI-generated content, but still faces significant challenges. Its success will rely on robust implementation, but on industry-wide regulation, standardization and collaboration above other things.

References

Chinese ai governance rules, 2023. URL http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Accessed on August 29, 2023.

European ai act, 2023. URL https://artificialintelligenceact.eu/. Accessed on August 29, 2023.

Aaronson, S. and Kirchner, H. Watermarking gpt outputs, 2023. URL https://www.scottaaronson.com/talks/watermark.ppt.

California State Leg. Amendment to california assembly bill 3211. California State Legislature, April 2024. URL https://legiscan.com/CA/text/AB3211/id/2984195. Amended in Assembly.

Christodorescu, M., Craven, R., Feizi, S., Gong, N., Hoffmann, M., Jha, S., Jiang, Z., Kamarposhti, M. S., Mitchell, J., Newman, J., et al. Securing the future of genai: Policy and technology. *Cryptology ePrint Archive*, 2024.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Dekel, T., Rubinstein, M., Liu, C., and Freeman, W. T. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2146–2154, 2017.

Douze, M., Tolias, G., Pizzi, E., Papakipos, Z., Chanussot, L., Radenovic, F., Jenicek, T., Maximov, M., Leal-Taixé, L., Elezi, I., et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021.

Fernandez, P., Chaffin, A., Tit, K., Chappelier, V., and Furon, T. Three bricks to consolidate watermarks for large language models. *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023a.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. *ICCV*, 2023b.

Fernandez, P., Douze, M., Jégou, H., and Furon, T. Active image indexing. *ICLR*, 2023c.

Harris, D. E. and Norden, L. Meta's ai watermarking plan is flimsy, at best. 2024. URL https://spectrum.ieee.org/meta-ai-watermarks. Accessed on May 2, 2024.

Imatag. State of image metadata, 2018. URL https://www.imatag.com/blog/state-of-image-metadata-in-2018.

Jovanović, N., Staab, R., and Vechev, M. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*, 2024.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., and Goldstein, T. On the reliability of watermarks for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DEJIDCmWOz.

Knibbs, K. Researchers tested ai watermarks—and broke all of them. *Wired*, October 2023. URL https://www.wired.com/story/artificial-intelligence-watermarking-issues/. Accessed on May 2, 2024.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pp. 24950–24962. PMLR, 2023.

O'Reilly, P., Jin, Z., Su, J., and Pardo, B. Maskmark: Robust neuralwatermarking for real and synthetic speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4650–4654. IEEE, 2024.

Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., and Feizi, S. Robustness of ai-image detectors: Fundamental limits and practical attacks. *ICLR*, 2024.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

San Roman, R., Fernandez, P., Elsahar, H., D´efossez, A., Furon, T., and Tran, T. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, 2024.

Spitale, G., Biller-Andorno, N., and Germani, F. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

Suno. Introducing v3. 2024. URL https://suno.com/blog/v3. Accessed on June 6, 2024.

USA. Ensuring safe, secure, and trustworthy ai. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf, July 2023. Accessed: [july 2023].

Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Wikipedia. Digital Millennium Copyright Act. https://en.wikipedia.org/w/index.php?title=Digital_Millennium_Copyright_Act&oldid=1221667351, 2024. [Online; accessed 5-June-2024].

Wong, P. W. A public key watermark for image verification and authentication. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, volume 1, pp. 455–459 vol.1, 1998. doi: 10.1109/ICIP.1998.723526.

Zhu, J., Kaplan, R., Johnson, J., and Fei-Fei, L. Hidden: Hiding data with deep networks. In *ECCV*, 2018.

A. Watermarking in Recent Drafts (2023-2024)

A.1. White House Executive order

Sec. 3. Definitions.

> *(gg) The term "watermarking" means the act of embedding information, which is typically difficult to remove, into outputs created by AI — including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.*

Sec. 4.5. Reducing the Risks Posed by Synthetic Content.

> *(a) Within 240 days of the date of this order, the Secretary of Commerce, in consultation with the heads of other relevant agencies as the Secretary of Commerce may deem appropriate, shall submit a report to the Director of OMB and the Assistant to the President for National Security Affairs identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for: [...]*
> *(ii) labeling synthetic content, such as using watermarking; [...]*

Sec. 10. Advancing Federal Government Use of AI. 10.1. Providing Guidance for AI Management.

> *(b) To provide guidance on Federal Government use of AI, within 150 days of the date of this order and updated periodically thereafter, the Director of OMB, in coordination with the Director of OSTP, and in consultation with the interagency council established in subsection 10.1(a) of this section, shall issue guidance to agencies to strengthen the effective and appropriate use of AI, advance AI innovation, and manage risks from AI in the Federal Government. The Director of OMB's guidance shall specify, to the extent appropriate and consistent with applicable law: [...]*
> *(viii) in consultation with the Secretary of Commerce, the Secretary of Homeland Security, and the heads of other appropriate agencies as determined by the Director of OMB, recommendations to agencies regarding: [...]*
> *(C) reasonable steps to watermark or otherwise label output from generative AI; [...]*

A.2. California State Legislature AB-3211

The California Provenance, Authenticity, and Watermarking Standards Act is a legislative bill introduced in February 2024, aimed at regulating the use of generative artificial intelligence (AI) to ensure the authenticity and provenance of digital content. This bill would mandate the implementation of watermarking standards to identify synthetic content and require disclosure of content origins to mitigate the risks associated with AI-generated content.

- **Watermarking Requirements:**
  - Generative AI system providers must embed imperceptible and indelible watermarks in synthetic content, detailing the content's origins.
  - Watermarks must be designed to be maximally indelible and retain information even if the content is altered.
- **Disclosure and Reporting:**
  - Providers must develop tools to decode watermarks and make them publicly available.
  - Any vulnerabilities or failures in AI systems must be reported to the Department of Technology within 24 hours.
- **Online Platform Responsibilities:**
  - Large online platforms are required to disclose the provenance data of content to users and use advanced techniques to detect unlabeled synthetic content.
  - Platforms must also ensure users disclose if content is synthetic when uploading.
- **Digital Cameras and Recording Devices:**
  - From 2026, new devices sold in California must offer the option to embed authenticity and provenance watermarks.
  - Manufacturers must provide software updates for older devices to enable watermarking if technically feasible.
- **Annual Risk Assessment:**
  - Generative AI providers and large platforms must produce an annual Risk Assessment and Mitigation Report to evaluate the risks and harms associated with synthetic content.
- **Penalties for Non-compliance:**
  - Violations can result in administrative penalties up to $1,000,000 or 5% of the violator's annual global revenue, whichever is greater.
- **Regulatory Framework:**
  - The Department of Technology is tasked with adopting necessary regulations to implement the act and updating them as needed to align with national or international standards.

A.3. EU AI Act

Recital (133)

> *A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. In light of those impacts, the fast technological pace and the need for new methods and techniques to trace origin of information, it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate.*

Recital (134)

> *Further to the technical solutions employed by the providers of the AI system, deployers who use an AI system to generate or manipulate image, audio or video content that appreciably resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful (deep fakes), should also clearly and distinguishably disclose that the content has been artificially created or manipulated by labelling the AI output accordingly and disclosing its artificial origin. Compliance with this transparency obligation should not be interpreted as indicating that the use of the AI system or its output impedes the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter, in particular where the content is part of an evidently creative, satirical, artistic, fictional or analogous work or programme, subject to appropriate safeguards for the rights and freedoms of third parties. In those cases, the transparency obligation for deep fakes set out in this Regulation is limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work, including its normal exploitation and use, while maintaining the utility and quality of the work. In addition, it is also appropriate to envisage a similar disclosure obligation in relation to AI-generated or manipulated text to the extent it is published with the purpose of informing the public on matters of public interest unless the AIgenerated content has undergone a process of human review or editorial control and a natural or legal person holds editorial responsibility for the publication of the content.*

Recital (135)

> *Without prejudice to the mandatory nature and full applicability of the transparency obligations, the Commission may also encourage and facilitate the drawing up of codes of practice at Union level to facilitate the effective implementation of the obligations regarding the detection and labelling of artificially generated or manipulated content, including to support practical arrangements for making, as appropriate, the detection mechanisms accessible and facilitating cooperation with other actors along the value chain, disseminating content or checking its authenticity and provenance to enable the public to effectively distinguish AI-generated content.*

Article 50: **Transparency obligations for providers and deployers of certain AI systems**

paragraph 50(2)

> *Providers of AI systems, including GPAI systems, generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust, and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation, and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards.*

A.4. Chinese Interim Measures on Generative AI

Article 12

> *Providers shall mark the generated content such as pictures and videos in accordance with the "Regulations on the Management of Deep Synthesis of Internet Information Services".*

A.5. Practical Guidelines for Cybersecurity Standards

We refer the reader to the article: Labeling of AI Generated Content: New Guidelines Released in China for a review on the "Practical Guidelines for Cybersecurity Standards – Method for Tagging Content in Generative Artificial Intelligence Services", written in order to implement the requirements of the "Interim Measures for the Management of Generative Artificial Intelligence Services" for identifying generated content.