What Lies Ahead for Generative AI Watermarking

Pierre Fernandez [1] **Anthony Level** [2] **Teddy Furon**[†] [1]

Abstract

This position paper discusses the potential of watermarking as a means to improve transparency and traceability in AI-generated content. Although robustness is often highlighted as a major technical challenge, watermarking has undeniable advantages over other content provenance methods, such as forensics or fingerprinting, making it inevitable. However, more significant unanswered questions remain, such as how to use and trust the detection outcomes and how to ensure interoperability between actors. We should prioritize finding both technical and regulatory answers to these questions – currently scarce in the public discourse – rather than focusing on robustness, which is not truly problematic.

1. Introduction

Digital watermarking conceals information directly into the content itself, *e.g.* in pixels of an image. Watermark extractors or detectors are specific algorithms that extract 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 (Goldstein et al., 2023), 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)[1]. 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 main 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.

In section 2, we first give a brief overview of content provenance methods and of watermarking. In section 3, we highlight its technical strengths, such as low, controllable false positive rates and versatility across different content types. We also argue that *watermarking is robust enough* for its intended purpose, that is, deter the vast majority of the population by making watermark removal sufficiently challenging and legally discouraged. In section 4 we then point out other broader challenges that lie ahead, including governance and control issues, the implications of open-source generative models, and the responsibilities of various stakeholders in ensuring effective watermarking.

2. Overview of Content Provenance Methods

It is important to remember that watermarking is only one technical means toward better traceability and should be considered alongside other content provenance methods. We distinguish between *passive* and *active* methods. Passive methods rely on inherent characteristics of the content without altering it, while active methods modify the content to provide precise control and verification.

2.1. Passive methods

Forensics methods employ binary classifier that spot small hidden traces of generated content, such as variation in words probabilities (Mitchell et al., 2023), odd frequen-

[1]Inria Rennes [2]Imatag. [†]Work supported by ANR / AID under Chaire SAIDA ANR-20-CHIA-0011. Correspondence to: <pierre.fernandez, teddy.furon@inria.fr>.

Accepted to the 2^{nd} Workshop on Generative AI and Law, co-located with the International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

[1] We report the role of watermarking in these texts in App. A

cies in images (Corvi et al., 2023) or voice synthesizer artifacts (Le et al., 2023). Relying on these small unknown traces makes the detectors very brittle to shifts in the distribution of content, and make them fall short in effectiveness compared to watermarking techniques (Sadasivan et al., 2023; Saberi et al., 2024). As a key example, state-of-the-art detection methods (Wang et al., 2023) are fooled to random chance simply by compressing generated images with JPEG (Grommelt et al., 2024). For the same reasons, these detectors are likely to get worse as generative models get better and as their artifacts disappear.

Fingerprinting (or copy detection) stores hashes of all the AI-generated content in a database, *e.g.* NeuralHash (Apple Inc., 2021). These hashes are vector representations $\in \{0,1\}^k$ or $\mathbb{R}^k$ usually generated from self-supervised feature extractors (Oquab et al., 2023; Devlin et al., 2018). When a piece of content is queried, we compare its hash to those in the database and determine if it is a copy of a registered generation. At scale, storing the hashes and searching through them is cumbersome, and reverse search must be approximate to be tractable (Douze et al., 2024). Moreover, the feature extractors are not perfectly robust to content modification: for instance, an audio and its $\times 1.25$ speed-up version may have different hashes. These two factors result in errors especially in an adversarial setting (Douze et al., 2021; Papakipos et al., 2022). Another downside is the need of storing the hashes in a database, which makes it harder to share and impossible for open-source scenarios.

2.2. Active methods.

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.

Visible watermarks are straightforward and widely recognized. However, in addition to obviously degrading the quality of the content, they are also easy to remove or tamper with (Dekel et al., 2017), making them less reliable.

Invisible watermarking usually consists of an *embedder* and an *extractor/detector*, respectively responsible for hiding and revealing the watermark information. It is versatile and copes with different types of content. More often than not, for audio (Chen et al., 2023; O'Reilly et al., 2024; San Roman et al., 2024) and image (Zhu et al., 2018; Tancik et al., 2020), the embedder and the extractor are neural networks: the embedder takes a piece of content and a binary message and outputs a slightly altered version, while the extractor takes the content and outputs a binary message. To detect the watermark, we look at the output message and see if there is a match with the embedded one. There are some variants to these methods (Juvela & Wang, 2023; Fernandez et al., 2023b; Wen et al., 2023). Methods for language model-generated text differ a little (Kirchenbauer et al., 2023; Aaronson & Kirchner, 2023). They modify how next tokens are sampled from the language model, resulting in texts that present a different statistics of sequences of words or letters.

3. Why Watermarking is Robust Enough

Because watermarking works by actively modifying content, there is a common belief that these traces can be easily removed. However, this section outlines the technical superiority of watermarking, particularly in terms of detection confidence and robustness. Additionally, when viewed as part of a broader ecosystem that includes detection algorithms, legal frameworks, and social norms, watermarking proves to be robust enough for its intended purpose.

3.1. Invisible watermarking surpasses other methods

Watermarking presents undeniable advantages. First, it intentionally injects traces into content, whence the greater robustness than that of passive methods like forensics or fingerprinting. For instance, Fernandez et al. (2023c) show that a simple image watermarking method achieves $\approx \times 8$ better recall than fingerprinting after crops that keep 50% of the original images; and Fernandez et al. (2023b) compare watermarking to forensics and show that it achieves the same true positive rate (probability of correctly flagging a watermarked piece of content) for a 10 million times smaller false positive rate (probability of wrongly flagging a non-watermarked one), on images that are cropped, resized and compressed. Attacking the watermark is always possible but this always damages the quality, contrary to visible watermark or metadata erasure – note that this is also true for forensics and fingerprinting methods.

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.

3.2. Are attacks really a limitation?

Categorizing the attacks. Watermarking is not foolproof (Sadasivan et al., 2023; Saberi et al., 2024; Jovanović et al., 2024). It is subject to attacks that are roughly categorized based on the attacker's knowledge. *White-box* attacks have full access to the watermarking algorithm and its parameters (*e.g*. model weights); *black-box* attacks only have access to inputs and outputs, for instance through an API; and *no-box* attacks do not have any knowledge of the system. The effectiveness and ease of an attack generally increases with the attacker's level of knowledge about the watermarking system (San Roman et al., 2024), the hardest ones being no-box attacks, where there is not even the feedback on if an attack was successful or not.

Most attacks are removal attacks, where the goal is to eliminate the watermark from the content. Watermark forging, where the attacker creates a counterfeit watermark, may pose a more significant problem. Currently, without white-box access to the embedder or extractor, forging a watermark is considerably difficult. 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.

Watermarking keeps honest people honest. Most mafia organizations or belligerent countries have the expertise and resources to train their own models. They will include neither watermarking nor metadata, and forensic methods are also doomed to fail due to the lack of such data to train a detector. Watermarking's goal is not to protect against these cases. Rather, it aims 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), and as is presented in the US Senate "COPIED Act" (see App. A). This aligns with the motto "keep honest people honest," popularized by Hollywood in the 2000s about DRMs.

4. The Real Challenges Ahead

The technical aspects of watermarking, like its robustness to adversarial attacks, are far from being the only considerations to take into account. It is essential to address overlooked challenges that concern governance, control, and, maybe naively, how to even use the detection outcomes.

4.1. Who controls watermark detection?

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. Consequently, no record of watermark detection by anyone other than the generative model's owner currently 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, unbiased entities. This raises questions about who these entities should be and how they are governed.

4.2. Open-source generative models?

They present a unique challenge since they are freely available and usable without post-hoc watermarks (applied after generation). The case in point is Stable Diffusion (Rombach et al., 2021), which was open-sourced in late 2022. Removing the watermark in its source code was as simple as commenting out a single line. Ideally, models should be trained or fine-tuned to generate watermarked content natively as in (Fernandez et al., 2023b; Gu et al., 2023). 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.

4.3. What to do with detection?

This question is not clearly addressed by current regulations. The entire value chain of content distribution (large platforms, social networks, search engines, etc.) should be required to query detectors in order to label their content in some way. For instance, the EU AI Act only provides for codes of good practice for watermark embedding. There is therefore a significant temporal discrepancy: companies will watermark generative models' outputs, but it will be necessary to go to their site to query the detector in the first instance. Content detection will only be incentivized in a second instance, whereas it would have been necessary to deploy detection at the same time and with the same strength as the watermark embedding.

Besides, it is not even very clear how to deploy detection. 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, current regulations lead different entities to quickly develop their own watermarking methods. This results in 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.

5. Conclusion

Watermarking is the most viable technology for improving transparency and traceability in AI-generated content. Its success will rely on robust implementation, but on industry-wide regulation, standardization and collaboration above other things. In short, why worry about whether the watermark is robust enough to adversarial attacks, if nobody other than the company that owns the generative model has the ability to detect it?

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.

Apple Inc. Csam detection - technical summary, 2021. URL https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf. Accessed: Jul. 4, 2024.

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.

Chen, G., Wu, Y., Liu, S., Liu, T., Du, X., and Wei, F. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

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.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

Grommelt, P., Weiss, L., Pfreundt, F.-J., and Keuper, J. Fake or jpeg? revealing common biases in generated image detection datasets. *arXiv preprint arXiv:2403.17608*, 2024.

Gu, C., Li, X. L., Liang, P., and Hashimoto, T. On the learnability of watermarks for language models. *arXiv preprint arXiv:2312.04469*, 2023.

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.

Juvela, L. and Wang, X. Collaborative watermarking for adversarial speech synthesis. *arXiv preprint arXiv:2309.15224*, 2023.

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.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

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.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 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.

Papakipos, Z., Tolias, G., Jenicek, T., Pizzi, E., Yokoo, S., Wang, W., Sun, Y., Zhang, W., Yang, Y., Addicam, S., et al. Results and findings of the 2021 image similarity challenge. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 1–12. PMLR, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2021.

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.

Tancik, M., Mildenhall, B., and Ng, R. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

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].

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 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].

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. Edited and Deepfaked Media Act ("COPIED Act")

The recent Content Origin Protection and Integrity from Edited and Deepfaked Media Act ("COPIED Act"), introduced in the Senate on July 10, 2024, prohibits the manipulation or disabling of AI origin information, which is intended to protect the authenticity and ownership of digital content. However, the addition of a watermark is left to the user's discretion, so it would not be automatic. This may be due to the fact that the bill is also focused on respecting copyright, and the watermark would serve to identify what is in the training databases of GenAI (and therefore a user who does not wish to watermark their content would not claim copyright over it). On the other hand, the bill prohibits the removal of watermarks, which is a world first.

A.4. 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.5. 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.6. 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.