JRC 137136

European Commission

SCIENCE FOR POLICY BRIEF

Generative AI Transparency: Identification of Machine-Generated content

HIGHLIGHTS

→ Generative AI is a cutting-edge field of AI that can create realistic human-like content. Concerns over its misuse and societal implications highlight the importance of identifying machine-generated content.

→ Solutions to identify AI generated content should possess the following four properties: efficiency, integrity of data, robustness to content alteration, and protection against manipulation.

→ Current technical solutions based on metadata, watermarking, fingerprinting or detection do not fulfil these requirements sufficiently for text, audio, image and video content.

INTRODUCTION

Generative AI (GenAI) [1] is a cutting-edge field of artificial intelligence (AI) that has recently gathered considerable attention with the recent advancements of state-of-the-art techniques and the emergence of consumer-facing products such as ChatGPT, Midjourney or Sora. GenAI refers to machine learning models used to generate media content such as text, audio, image or video that mirrors human-made content. The potential of applications spans across various industries. The creativity and reasoning abilities of this technology may affect all intellectual and artistic professions.

Despite the benefits that GenAI offers across various domains, its growing capabilities have sparked concerns about unique AI-specific risks to safety and fundamental rights. GenAI has the potential to support misinformation campaigns and amplify opinion manipulation [2], and to increase the efficiency of fraud by making plagiarism or impersonation more difficult to detect and more efficient [3]. These risks can impact significantly democratic processes [4]. GenAI also blurs the line between human- and machine-created content, triggering new paradigms for jobs, creativity and copyright rules [5].

Policy context

The EU AI Act [6] is a pioneering initiative to propose transparency obligations for generative AI. In its Art. 52(3), it states that “users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful (‘deep fake’), shall disclose that the content has been artificially generated or manipulated”. The European Parliament proposed [7] extending recital 60g to specifically state “generative foundation models should ensure transparency about the fact the content is generated by an AI system, not by humans”. Finally, in the provisional agreement [8], co-legislators agreed to include additional transparency obligations in Art. 50: “Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that 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". Beyond the EU, the question of machine-generated content has also been considered. This includes in the US voluntary commitments by the leading AI companies to develop robust technical means that ensure that users know when content is AI generated [9], and in China numerous requirements on deep synthesis internet services, including the use of technical measures to label content produced or edited by these services [10].

Technological context

The rise of digital technologies over the past fifty years has considerably remodelled the landscape of ownership and copyright. The possibility to duplicate and alter digital assets led to the development of tools to embed hidden information or markings into digital media such as images, audio, videos, or documents (watermarking) or univocally identify them (fingerprinting), to ensure the authenticity, traceability and security of assets [11]. The first watermarking approaches involved simple and visible modifications to images, such as addition of logos or text. Overtime these techniques evolved into watermarks that are imperceptible to human senses and robust against alterations, in both the visual and audio domains. Watermarking approaches have been combined with fingerprinting methods, cryptographic techniques for added security.

Generative AI technology

GenAI can generate any type of data such as genomics data, 3D environments, or tabular data. For the purpose of this brief, the focus is on text, audio image and video, which are the most common types of generated data in research works. The main AI techniques behind GenAI are Transformer-based large language models [12] such as Generative Pre-trained Transformers (GPT) for text generation; convolutional network based techniques [13] such as WaveNet for audio generation; Generative Adversarial Networks (GAN) [14] and diffusion models [15] for image generation. Generation of video [16] or multi-modal content relies on combined techniques.

Scope of the brief

Making Generative AI more transparent and being able to detect and identify machine-generated content is crucial to ensure that the confidence in digital technology and media will remain intact [17], promoting trust on the European digital ecosystem.

This brief aims at reviewing four technical solutions (see Figure 1) to achieve this goal. They are evaluated according to four desirable properties:

1. Efficiency: Reliable identification of generated content, by retrieving information such as the name of the provider, the date of creation, or a digital signature for authentication. The process should require minimum effort and time, and remain consistent over time.
2. Integrity of data: Preservation of the integrity of the content, i.e., limited degradation or distortion of the original data.
3. Robustness to content alteration: Preserved efficiency when the content is subject to changes or alterations that are foreseeable and do not affect the synthetic nature of the content, nor alter the overall appearance or interpretability of the content (e.g., the brightness of an image or the volume of an audio).
4. Protection against manipulation: Ability to withstand any modification that is intended to manipulate the information used for identification purposes, either to change the identifying elements (tampering) or to remove the information (removal).

TRANSPARENCY TECHNIQUES FOR GENERATIVE AI

Metadata

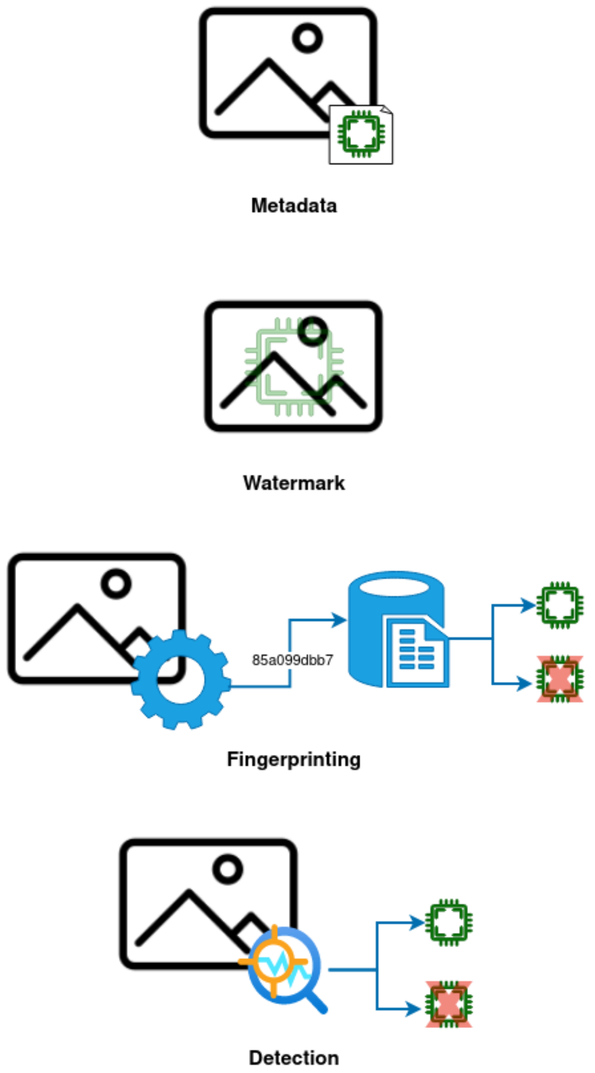
Metadata is data that is embedded in computer files and that provides information about it, such as copyright or ownership details, timestamps, unique identifiers or digital signatures associated to the content.

Identification: Reading metadata is straightforward and requires minimal effort. However, this approach requires using a format that accepts metadata, such as PNG, JPG, MP3 or PDF. While this is common when generating audio or image, it is usually not the case for text generation that returns raw text.

Integrity of data: Metadata does not alter the content and it is stored separately.

Robustness to content alteration: Altering the content does not affect metadata. However, some information in metadata may need to be updated to reflect the alteration.

Figure 1 Schematic description of the four technical solutions analysed to identify generated content.



Protection against manipulation: Metadata can be easily tampered with or simply removed from the file (e.g., using dedicated tools) [18]. Protecting information in metadata against unauthorised alteration could be provided by cryptographic signatures.

Watermarking

Watermarking techniques embed metadata as invisible or barely perceptible markers into the content. Watermarking can be done during the generation of the media content, or afterwards as a post-processing step. Some approaches also attempt to incorporate watermarking at the training stage, so that the GenAI system inherently produced watermarked content [19].

Identification: Watermarking requires specific tools to verify authenticity, detect tampering, or prove ownership. This is still an open research topic in the scientific community with promising results for all types of modalities [20]–[23].

Integrity of data: Watermarking techniques alter the content. However, the watermark can be designed to have minimum impact on the quality of the content, particularly for some modalities like images.

Robustness to content alteration: Watermarking can be sensitive to modifications of the content, which may reduce or prevent the identification [24].

Protection against manipulation: Intentional manipulation of data to remove or alter watermarks in content is possible and has been demonstrated in scientific works [25]. Additional layers of protection (e.g., encryption) can be considered to limit this risk.

Fingerprinting

In the context of transparency of GenAI, fingerprinting consists in generating and storing in an external database a unique identifier for the generated content, known as fingerprint or hash.

Identification: The process of identification consists in calculating the fingerprint of the generated content to be identified and comparing it with a list of known fingerprints [11], [26].

Integrity of data: Fingerprinting does not alter the content, unless it is explicitly stored as a watermark.

Robustness to content alteration: Fingerprinting can be sensitive to modifications of the content, which may lead to a different fingerprint and a wrong identification.

Protection against manipulation: Intentional modifications can lead to different fingerprints, even without visible changes of the content [27].

Detection

AI-based detection tools are built using machine-learning classification techniques and trained on human-made and machine-generated content [28]–[30]. They can be applied to any type of data, provided sufficient examples exist to train detectors.

Identification: The process of identification consists in inputting the content to the detector. However, the current technology for detecting generated content has a high false-positive rate and can misidentify human-generated content [31].

Integrity of data: This approach does not require altering the content.

Robustness to content alteration: Detection is sensitive to strong modifications of the content [32]. They need to be continuously updated to adapt to the new generations of GenAI.

Protection against manipulation: As with any AI systems, evasion attacks can be built to mislead detectors and make them return wrong predictions [27], [32].

OPEN SOURCE

Openness fosters a culture of innovation, allowing developers to iterate on key ideas and progressively develop increasingly advanced systems. However, when GenAI models are open source, removing metadata, fingerprints or watermarks can be as simple as deleting a single line of code, thus facilitating potential malicious uses. If any of the methods used (metadata, fingerprinting, watermarking, or detection) are also open source, malicious actors may analyse the code to figure out ways to circumvent the mechanisms for identifying generated content.

The most robust approaches to these issues are those that integrate identification mechanisms into the generation process, for example, by embedding watermarks in the GenAI models [33] or by watermarking all images in the training dataset so, the GenAI model intrinsically generates watermarked content [19]. Hybrid open-closed approaches can also be implemented, where one end may be open while the other remains closed. For example, open watermarking code and closed watermarking detection, or vice versa.

DISCUSSION

Transparency measures for GenAI are limited by the current state of the art, and no single solution fits all properties that are desirable to identify robustly and reliably generated content (see Table 1).

Incorporating metadata and watermarking into content allow for easier tracking and authentication of machine-generated digital content. Initiatives such as the Coalition for Content Provenance and Authenticity (C2PA) [34] have been leveraged by GenAI providers [35], [36] to add robust identifiers in metadata in content generated by their systems.

Table 1 Comparison of technical solutions with respect to the four desirable properties discussed above. (Green: Covered, Orange: Partially covered, Purple: No covered).

	Efficiency			Integrity			Robustness to content alteration			Protection against manipulation		
	Text	*Audio*	*Image*	*Text*	*Audio*	*Image*	*Text*	*Audio*	*Image*	*Text*	*Audio*	*Image*
Metadata	Green	Green	Green	Green	Green	Green	Green	Green	Green	Purple	Purple	Purple
Watermarking	Purple	Green	Green	Orange	Orange	Orange	Orange	Orange	Orange	Purple	Orange	Orange
Fingerprinting	Orange	Green	Green	Green	Green	Green	Orange	Green	Green	Orange	Orange	Orange
Detection	Purple	Orange	Orange	Green	Green	Green	Purple	Orange	Orange	Purple	Orange	Orange

However, these methods are not foolproof, as they can be altered or removed.

On the other hand, fingerprinting and detection approaches allow distinguishing between genuine and manipulated content, regardless of the presence of metadata or watermarks. However, they still have some drawbacks: fingerprinting requires a dedicated infrastructure for generating and storing fingerprints on a large scale, while current attempts by major GenAI providers to develop AI-based detection tools have proven to be unreliable so far [31].

A better approach would involve the application of a combination of techniques in specific contexts, taking into account technical and legal considerations, including the type of model, the limitations of transparency measures, obligations for providers, current practices of platforms and organisations handling potential generated content. In particular, solutions that rely on digital signatures would require the setup a suitable Public-Key-Infrastructure (PKI) along with the necessary organisational procedures to handle key distribution for the providers. Additionally, technical implementations could be left to providers, or specified in dedicated standards to promote interoperability.

All these aspects should be weighed in and part of a broader governance system of GenAI to ensure the right interplay between all parties. In practice, this present several challenges, particularly for decentralised open source projects and for AI systems that work in edge devices. On the scientific side, further fundamental research is needed to advance the state of the art and develop more reliable solutions. This involves also the exploration of new engineering methods to develop products enabling a more efficient and reliable identification of generated content.

REFERENCES

[1] D. Fernández Llorca, E. Gómez, R. Hamon, I. Sánchez and G. Mazzini, 'Considerations around AI Act related terminology: general purpose AI systems, foundation models and generative AI', Artificial Intelligence and Law, In Preparation, 2024.

[2] 'Forecasting Potential Misuses of Language Models for Disinformation Campaigns—and How to Reduce Risk', Center for Security and Emerging Technology. [Online]. Available: link.

[3] E. Ferrara, 'GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models'. Journal of Computational Social Science, 2024.

[4] V. Wirtschafter, "The impact of generative AI in a global election year", Brookings, 2024. [Online]. Available: link.

[5] B. L. T. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, and E. Gómez, 'Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis', Arts, vol. 8, no. 3, p. 115, 2019.

[6] European Commission, 'Proposal for a Regulation laying down harmonised rules on Artificial Intelligence'. 2021. [Online]. Available: link.

[7] European Parliament, 'Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation [...] laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [...], 2023. [Online]. Available: link.

[8] European Parliament, 'CORRIGENDUM to the position of the European Parliament adopted [...]", 2024. [Online]. Available: link.

[9] T. W. House, 'FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence', The White House. [Online]. Available: link.

[10] E. Hine and L. Floridi, 'New deepfake regulations in China are a tool for social stability, but at what cost?', Nat Mach Intell, vol. 4, no. 7, pp. 608–610, 2022.

[11] L. de C. T. Gomes, P. Cano, E. Gómez, M. Bonnet, and E. Batlle, 'Audio Watermarking and Fingerprinting: For Which Applications?', Journal New Music Research, vol. 32, no. 1, pp. 65–81, 2003.

[12] A. Vaswani et al., 'Attention is all you need', 31st International Conference on Neural Information Processing Systems, Dec. 2017, pp. 6000–6010. [Online]. Available: link.

[13] A. van den Oord et al., 'WaveNet: A Generative Model for Raw Audio', arXiv: 1609.03499, 2016.

[14] I. Goodfellow et al., 'Generative Adversarial Nets', in Advances in neural information processing systems (NIPS), 2014, pp. 2672–2680.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis With Latent Diffusion Models', IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[16] OpenAI, 'Video generation models as world simulators', 2024.[Online]. Available: link.

[17] GPAI, 'State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release', Global Partnership on Artificial Intelligence, 2023. [Online]. Available: link.

[18] Microsoft, 'How to remove metadata from your photos (and why you should)', Microsoft 365. [Online]. Available: link.

[19] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, 'Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data', IEEE/CVF International Conference on Computer Vision, 2021, pp. 14448–14457.

[20] S. Wu, J. Liu, Y. Huang, H. Guan, and S. Zhang, 'Adversarial Audio Watermarking: Embedding Watermark into Deep Feature', IEEE International Conference on Multimedia and Expo (ICME), pp. 61–66, 2023.

[21] Sven Gowal and Pushmeet Kohli, 'Identifying AI-generated images with SynthID', Google DeepMind. [Online]. Available: link.

[22] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, 'A Watermark for Large Language Models', 40th International Conference on Machine Learning, 2023.

[23] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, 'Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images', 37th Conference on Neural Information Processing Systems, 2023.

[24] N. Agarwal, A. K. Singh, and P. K. Singh, 'Survey of robust and imperceptible watermarking', Multimedia Tools and Applications, vol. 78, no. 7, pp. 8603–8633, 2019.

[25] Z. Jiang, J. Zhang, and N. Z. Gong, 'Evading Watermark based Detection of AI-Generated Content'. arXiv: 2305.03807, 2023.

[26] M. Steinebach, 'An Analysis of PhotoDNA', 18th International Conference on Availability, Reliability and Security, in ARES '23, 2023.

[27] L. Struppek, D. Hintersdorf, D. Neider, and K. Kersting, 'Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash', ACM Conference on Fairness, Accountability, and Transparency, in FAccT '22, pp. 58–69, 2022.
[28] R. Tang, Y.-N. Chuang, and X. Hu, 'The Science of Detecting LLM-Generated Texts'. Comm. of the ACM, vol. 67, no. 4, pp. 50-59, 2024.
[29] S. Munir, B. Batool, Z. Shafiq, P. Srinivasan, and F. Zaffar, 'Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models', 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1811–1822, 2021.
[30] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, 'DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature', International Conference on Machine Learning (ICML), 2023.
[31] OpenAI, 'How can educators respond to students presenting AI-generated content as their own? | OpenAI Help Center'. [Online]. Available: link.
[32] M. Saberi et al., 'Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks'. arXiv: 2310.00076, 2023.
[33] P. Fernandez, G. Couairon, H. Jégou, M. Douze, T. Furon, 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models', arXiv: 2303.15435, 2023.
[34] L. Rosenthol, 'C2PA: the world's first industry standard for content provenance (Conference Presentation)', in Applications of Digital Image Processing XLV, SPIE, Oct. 2022, p. 122260P.
[35] OpenAI, 'C2PA in DALL·E 3', OpenAI Help Center. Accessed: Feb. 13, 2024. [Online]. Available: link.
[36] N. Clegg, 'Labeling AI-Generated Images on Facebook, Instagram and Threads', Meta. [Online]. Available: link.

AUTHORSHIP

This policy brief was prepared by, Ronan Hamon, Ignacio Sánchez, David Fernández Llorca and Emilia Gómez.

AUTHORSHIP

COPYRIGHT