BRIEFING

European Parliament

Generative AI and watermarking

SUMMARY

Generative artificial intelligence (AI) has the potential to transform industries and society by boosting innovation, empowering individuals and increasing productivity. One of the drawbacks of the adoption of this technology, however, is that it is becoming increasingly difficult to differentiate human-generated content from synthetic content generated by AI, potentially enabling illegal and harmful conduct.

Policymakers around the globe are therefore pondering how to design and implement watermarking techniques to ensure a trustworthy AI environment. China has already taken steps to ban AI-generated images without watermarks. The US administration has been tasked with developing effective labelling and content provenance mechanisms so that end users are able to determine when content is generated using AI and when it is not. The G7 has asked companies to develop and deploy reliable content authentication and provenance mechanisms, such as watermarking, to enable users to identify AI-generated content. The EU's new AI act, provisionally agreed in December 2023, places a number of obligations on providers and users of AI systems to enable the detection and tracing of AI-generated content. Implementation of these obligations will likely require use of watermarking techniques.

Current state-of-the-art AI watermarking techniques display strong technical limitations and drawbacks, however, in terms of technical implementation, accuracy and robustness. Generative AI developers and policymakers now face a number of issues, including how to ensure the development of robust watermarking tools and how to foster watermarking standardisation and implementation rules.

IN THIS BRIEFING

EPRS | European Parliamentary Research Service

Author: Tambiama Madiega
Members' Research Service

PE 757.583 – December 2023

EN

Generative AI: Need for transparency

'Generative AI' refers to technology designed to generate various types of new content in response to a user prompt. It is powered by general-purpose AI (GPAI) models – also referred to as 'foundation models' – that are trained on a broad set of data and can be adapted by AI developers to perform a wide range of tasks. Generative AI tools such as ChatGPT, GPT4, Midjourney, DALL-E and Bard are now increasingly being used by end-users to generate new content, such as video, audio, text, images, computer code and product designs.

The deployment of generative AI is spreading rapidly, with far-reaching implications in terms of the content used and produced by machine-learning processes. Generative AI is trained on machine-learning models called 'large language models' (LLMs), using a massive volume of data including publicly available data. This raises concerns about the unauthorised exploitation of datasets. Furthermore, generative AI is open to misuse and can potentially lead to plagiarism, privacy issues and AI hallucination phenomena, i.e. when false or incorrect information is provided by AI in a very convincing manner. The impact of generative AI on intellectual property and the risk that such tools can produce copyright infringing outputs has long been documented. Finally, generative AI can be used for a range of harmful purposes. Deep fake content (i.e. manipulated or synthetic audio or visual media that seem authentic) is increasingly indistinguishable from human-generated content, and is fuelling disinformation and misleading content.

The need to differentiate AI-generated synthetic content from human content has become a key policy issue as AI-generated text, images and videos become more prevalent and realistic. Recent studies show that human communication is increasingly intermixed with language generated by AI and that people are increasingly unable to detect AI-generated content.[1] Against this background, policymakers and AI practitioners are reflecting on how companies developing AI-generated content should increase the transparency and accountability of generative AI outputs and help users discern the difference between human-generated and AI-generated content. A range of approaches are being tested to trace how AI content is generated and to document its provenance. These include content labelling, the use of automated fact-checking tools, forensic analysis – which examines content for any inconsistencies or anomalies that indicate manipulation, and watermarking techniques.

AI watermarking techniques

AI watermarking is a process of embedding into the output of an artificial intelligence model a recognisable and unique signal (i.e. the watermark) that serves to identify the content as AI-generated. In practice, AI watermarking creates a unique identifiable signature that is invisible to humans but algorithmically detectable and that can be traced back to the AI model. Different watermarking techniques have been developed for text, image, video and audio content.[2]

In practice, watermarking has two separate phases: marking and identification of the watermark. Watermarks need to be created during the model training phase by teaching the model to embed a specific signal or identifier in the content generated. Then, after the AI model has been deployed, specialised algorithms can detect the presence of the embedded watermark in order to identify specific content as AI-generated. Well-designed watermarking should enable AI-generated content to be detected and its provenance identified.[3]

A number of private companies are developing AI watermarking tools. Google is testing a digital watermark called SynthID to identify images created by AI. Microsoft has pledged to watermark AI-generated images and Meta recently announced that its plan to embed an invisible watermark in its text-to-image generation products to enhance transparency.

Benefits of AI watermarking

Content authentication and data monitoring

Traceability of generative AI is a key to ensuring a trustworthy environment and identifying the provenance of the data used in the production of an AI model. AI companies are engaged in finding ways to differentiate human-generated and AI-generated content. The effectiveness of some approaches developed by the industry to trace AI-generated content – such as labelling – has been called into question. Open AI had to withdraw from the market a classifier it had trained to distinguish between human and AI-generated text written because of its low rate of accuracy.

Against this backdrop, AI watermarks techniques can serve to establish content authenticity and to perform content authentication. The techniques can also be used in the media sector for data monitoring – automatic registering and monitoring of broadcast radio programmes to ensure that royalties are paid to the rights-holders of the broadcast data.

Indicating authorship and protecting copyright

One of the most pressing challenges today is to work out how copyright rules should address generative AI. Given the uncertainty surrounding data training, OpenAI introduced a compensation programme called Copyright Shield to cover legal costs for copyright infringement suits filed against its customers regarding output generated by its AI tools in the US. Watermarks can help address this issue and enable online content to be traced back to a specific creator. The technology can help creators protect their content and track down copyright infringers more effectively, discouraging the unauthorised use of copyrighted material.

Preventing the spread of AI-generated misinformation

Watermarking AI-generated content also offers a helpful way to identify the origin of AI-generated disinformation. Media and news organisations, including online platforms, can use AI watermarks to indicate to readers that a piece of content was created using AI.[4] Watermarks are also useful when it comes to the authentication of media content and for flagging harmful AI outputs, such as fake news and deepfake videos.

AI watermarking limitations and drawbacks

A number of studies warn that state-of-the-art AI watermarking techniques display technical limitations on many counts. The tools display a number of drawbacks, regarding for instance:

- **Technical implementation**: AI companies face issues in creating watermarking, e.g. there are limited ways to add a marker to a text without changing the underlying meaning. AI-text detectors can also be biased against non-native writers of English. Watermarking techniques are not standardised and a watermark generated by one technology may not be readable by a system using different technology.
- **Accuracy**: AI-text detectors can result in false positives – incorrectly identifying human-created content as the product of AI – and therefore are often not reliable in practical scenarios.
- **Robustness**: studies show that both invisible and visible text-based and audiovisual watermarks can be manipulated, removed or altered (through backdoor attacks), shedding doubt on the authenticity of the content. LLMs are vulnerable to spoofing attacks, i.e. when an attacker (adversarial human) generates a non-AI text that is detected as AI-generated. If humans can infer hidden LLM text signatures and add them to human-generated text to be detected as text generated by the LLMs, this risks causing reputational damage to the LLMs' developers.

AI regulation and watermarking

China

China has already taken steps to ban AI-generated content without watermarks. The Cyberspace Administration of China released a range of requirements to label and watermark AI-generated content in August 2023.[5] Service providers of generative AI are required to watermark text, images, videos, and other content generated by their generative AI services. Three types of watermarking requirement are imposed.[6] Generated content is subject to an 'explicit watermark', i.e. a prompt text indicating 'generated by AI' must be visible, without affecting user usage. AI-generated images, videos, and audio are subject to an 'implicit watermark', i.e. technical tagging including at least the name of the service provider, imperceptible to humans but technically detectable by means of an interface or other tools. Finally, AI-generated content saved as files should display metadata for identification.

United States

United States (US) President Joe Biden recently signed an executive order on AI requiring the US administration to develop effective labelling and content provenance mechanisms so that end users can determine when content is generated using AI and when it is not. The US administration has been tasked, inter alia, with identifying the existing standards, tools, methods, and practices for authenticating content and tracking its provenance, and techniques for labelling synthetic content, such as watermarks. There are also calls on Congress to pass legislation on AI-generated content in 2024 to ensure the watermarking mechanism rules are enforced.

G7

The International Guiding Principles on Artificial Intelligence adopted by the G7 leaders in October 2023 under the Hiroshima Process call on organisations developing and using the most advanced AI systems (including the most advanced foundation models and generative AI systems) to develop and deploy reliable content authentication and provenance mechanisms, such as watermarking, to enable users to identify AI-generated content.

European Union

In the EU, the AI act provisionally agreed in December 2023[7] places several obligations on providers and users of AI systems to enable the detection and tracing of AI-generated content.

Deep fakes and other AI-generated content will have to be labelled as such. Providers of AI systems would have to disclose that the content they provide is AI-generated so that users can make informed decisions on further use. Providers would also have to design their systems in such a way that synthetic audio, video, text and image content is marked in a machine-readable format, and detectable as artificially generated or manipulated.

GPAI/foundation models will have to comply with specific transparency obligations before being placed on the market. They will have to implement measures that respect EU copyright law, applying state-of-the-art technologies, and publish a detailed summary of the copyright-protected content used to train their AI algorithms. In addition, generative AI providers, such as ChatGPT, will have to disclose that their content was generated by AI, and design their models so as to prevent them from generating illegal content. Implementation of these obligations will likely require use of watermarking techniques.

In addition, the 2022 Code of Practice on Disinformation introduces voluntary self-regulatory standards to fight disinformation in the EU. The code stipulates that signatories who develop or operate AI systems and who disseminate AI-generated and manipulated content through their services (e.g. deepfakes) commit to take into consideration the transparency obligations under the

Artificial Intelligence Act. The companies commit to establish procedures to counter AI systems that generate or manipulate content, such as warning users and proactively detecting such content.

Copyright and AI model training

Articles 3 and 4 of the EU Copyright Directive contain a set of exceptions to copyright protection for text and data mining (TDM).

Accordingly, researchers in academic research institutions and cultural heritage institutions are free to use all lawfully accessible works – including copyrighted content available on the public internet – to train machine-learning applications.

Commercial AI developers can only use works that are lawfully accessible and whose rights-holders have not explicitly ruled out use for TDM purposes. Creators and rights-holders who want to control the use of their works can opt out to prevent their works from being used to train AI models or to establish a negotiating position for licensing the use of their works.

The implementation of such obligations could prove challenging in practice. The transparency measures supported by robust watermarking measures would allow rights-holders to rely on the opt-out mechanism envisaged under the Copyright Directive.

AI watermarking implementation and open questions

The limitations of the state-of-the-art watermarking raise a number of questions that generative AI developers and policymakers will have to address to ensure the implementation of requirements on watermarking. Two main issues in particular demand policymakers' attention.

Setting up a robust watermarking environment

A recent report by the Organisation for Economic Co-operation and Development (OECD) recommends obliging all organisations developing a foundation model intended for public use to demonstrate a reliable detection mechanism for the content it generates as a condition of its public release.[8] The detection mechanism could then be made publicly available to allow users to ask whether an AI model had generated a specific item (wholly or in part).

However, watermarking implemented in isolation will not be sufficient. It will have to be accompanied by other measures, such as mandatory processes of documentation and transparency for foundation models, pre-release testing, third-party auditing, human rights impact assessments and media literacy campaigns.[9]

Furthermore, experts stress that more interdisciplinary research is necessary to develop more robust watermarking and AI-content detection techniques.[10] For instance, using traditional 'information retrieval' (IR) methods – by keeping a private log of the foundation-model's generated content and running a detector tool on this private log – may be one way to avoid watermarking evasion strategies.[11] Other experts stress that typical notice and content disclosures are largely ignored by users and believe that the industry should develop AI language technologies that are self-disclosing by design (e.g. producing language that humans intuitively connect to AI sources and avoiding language that people wrongly associate with humanity).[12] Self-disclosing AI machines could be the solution to ensure AI's transparency and accountability.[13]

In addition, the provisions in the Digital Services Act concerning trusted flaggers and a notice-and-action mechanism, could be extended to the domain of generative AI to establish a more effective and decentralised system for flagging and removing illegal content generated by AI systems. Such community-driven oversight would ensure a broader base for monitoring and mandating a quick response to violations highlighted by trusted flaggers.[14]

Watermarking standardisation and implementation rules

There has been a worldwide research effort to find more robust watermarking techniques in recent years.[15] In this context, the long-identified need to foster the standardisation of watermarking techniques[16] must be addressed at EU level, but also at multilateral level, as stated at the G7 forum.

The Commission's draft standardisation request mandates the European Standardisation Organisations (CEN-CENELEC) to deliver a series of European standards by January 2025, including on transparency and information provision for AI system users. Questions such as which actors should be involved in the EU standardisation process and how this process will dovetail with standardisation efforts in other regions of the world will need to be factored in.[17] In this respect, the G7 International Code of Conduct for Organizations Developing Advanced AI Systems asks such organisations to cooperate with the standards development organisations (SDOs) to develop interoperable international technical standards for watermarking and rules to help users distinguish content generated by AI from non-AI-generated content.

Lastly, beyond technical standardisation processes, implementation rules will also be critical for implementing the watermarking standards. Questions such as deciding who should have the ability to detect the watermark signals, and decide whether the content is AI-generated and whether it is misleading, will need to be settled. In addition, policymakers and industry players will have to reflect on how best to enforce watermarking in open-source ecosystems where different versions of open-source software can proliferate.[18]

MAIN REFERENCES

European Union Intellectual Property Office, Automated Content Recognition: Discussion Paper – Phase 1 'Existing technologies and their impact on IP', 2020.

Hacker P. et al, Regulating ChatGPT and other Large Generative AI Models, *2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

Henderson P., Watermarks for Generative AI?, *Georgetown Journal of International Affairs*, 2023.

Jakesch M. et al, Human heuristics for AI-generated language are flawed, *Proceedings of the National Academy of Sciences,* Vol. 120, 11, 2023.

Knott A. et al, Generative AI models should include detection mechanisms as a condition for public release, *Ethics and Information Technology*, Vol. 25, 55, 2023.

ENDNOTES

1 See M. Jakesch et al, Human heuristics for AI-generated language are flawed, *Proceedings of the National Academy of Sciences*, Vol. 120 (11), 2023. Participants in the study could only distinguish between human or AI text with 50 to 52 per cent accuracy. This flaw could be exploited by malevolent actors to engage in a range of illegal or harmful activities, including automated impersonation, targeted disinformation campaigns, fraud, and identity theft.

2 For an overview see EUIPO, Automated Content Recognition: Discussion Paper – Phase 1 'Existing technologies and their impact on IP', 2020.

3 See S. Arnett, The Inside Scoop on Watermarking and Content Authentication, 2023.

4 See L. Craig, AI watermarking, 2023.

5 See P. Henderson, Watermarks for Generative AI?, 2023.

6 See also X. Dan and Y. Luo, Labelling of AI Generated Content: New Guidelines Released in China, 2023.

7 See European Commission, Artificial Intelligence – Questions and Answers, 2023. Council, Press release, 9 December 2023; European Parliament, Press release, 9 December 2023; European Commission, Press release, 9 December 2023. The exact agreed text is being finalised. This briefing uses the wording of the EU institutions' press releases and their earlier position papers on general-purpose AI, foundation models and generative AI.

8 See also A. Knott et al, Generative AI models should include detection mechanisms as a condition for public release, 2023.

9 S. Gregory and R. Vazquez LLorente, Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights, 2023.

10 P. Hacker et al, Regulating ChatGPT and other Large Generative AI Models, 2023.

11 See Knott et al above.

12 See M. Jakesch et al above.

13 See P. Kannan, Was this written by a human or AI?, 2023.

14 See P. Hacker et al, Regulating ChatGPT and other Large Generative AI Models, 2023.

15 See for instance Yunqing Zhao, A recipe for watermarking diffusion models, 2023. See also L. Tang et al, Baselines for Identifying Watermarked Large Language Models, 2023.

16 See for instance, F Mintzer et al, Opportunities for watermarking standards, 1998.

17 See C. Perarnaud, With the AI Act, we need to mind the standards gap, 2023.

18 See A. Knott et al above.

DISCLAIMER AND COPYRIGHT

eprs@ep.europa.eu (contact)

www.eprs.ep.parl.union.eu (intranet)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)