Knowledge Centre Data & Society
Artificial Intelligence Flanders

HOW WATERTIGHT ARE WATERMARKS?

METHODS AND LEGISLATIVE INITIATIVES AROUND LABELING AI-GENERATED CONTENT

Knowledge Centre Data & Society (2024). How watertight are watermarks? Methods and legislative initiatives around labeling AI-generated content. brAInfood of the Knowledge Centre Data & Society. Brussels: Knowledge Centre Data & Society.

In recent months, AI tools have made significant progress in generating high-quality text, images and videos that are (nearly) indistinguishable from human-generated content.

As a result, AI-generated content is increasingly finding its way to a wide audience, through online platforms such as social media but also through traditional media forms like news broadcasts. The creation of content such as "deepfakes" is blurring the line between reality and fiction. This raises concerns about the possible manipulation of information and its impact on public opinion. Moreover, it is feared that such content may lead to abuse, such as spreading fake news or damaging one's reputation. Therefore, it is crucial to transparently communicate the use of AI in generating content to users and viewers.

In this brAInfood, we focus on some of the methods available to identify and verify the origin of AI-generated content, as well as the laws and regulations surrounding this type of content.

METHODS TO DESIGNATE AI-GENERATED CONTENT

For designating AI-generated content, a distinction can be made between, on the one hand, methods that are clearly and immediately observable by humans, on the other hand, methods that are observable (only) by means of (specialized) software and techniques. We give some examples:

Human-observable methods

> Label or disclaimer
A standard way to indicate that content was generated by AI is to include text informing the reader about the software used to generate the image, video or audio.

Waterproof?
However, adding a label or disclaimer is not a watertight solution to stop the spread of misleading content (such as deepfakes) because it is up to the placer and/or creator of the content to add a label with the (correct) information.

> Visible tags
May consist of a (semi-)transparent logo, text, or graphic overlay placed on top of the image or video, thus indicating authorship and/or the software used.

> Audible markings
Aurally indicates that content has been generated or manipulated by AI. These markings may include spoken text or recognizable "jingle" that plays before, after or during the audio clip.

Waterproof?
Both forms of marking (visual and aural) can be removed, modified or added in a relatively simple way. For example, by removing the logo with software (such as Photoshop) or cropping the image so that the logo falls outside the image.

WATERMARK

Machine-observable methods

> Metadata

(Textual) information or data added to a file that describes the characteristics of that file. For example, this can include information about the author, the hardware used (such as a particular model of photo camera), software, etc. So it is data that provides information about other data.

Waterproof?

Consulting metadata in Web browsers or (social media) apps is not easily accessible to (average) users. In addition, it is relatively easy to modify the content of metadata, which undermines its legitimacy. Indeed, someone with the goal of misleading people can simply remove and/or modify the information that informs users about the use of AI in creating the content.

To combat this form of free metadata modification, the 'Coalition for Content Provenance and Authenticity' (C2PA), a coalition of technology and media companies including Microsoft and Adobe, has proposed an open technical standard for tracking the origin and modification history of audio/visual media in metadata. What sets this particular form of metadata addition apart from others is that the metadata is protected from manipulation using cryptographic methods. Adaptations made in software that supports the C2PA standard are described and signed in the metadata file. For software that does not support the standard, modifications in the file are detected and listed by comparing the current version with the previous (officially signed) version. Although this technique is not foolproof (yet?), it is a way to inform users in a relatively reliable and clear way whether and what modifications the file has undergone.

> Digital watermarking

This is the hiding of information in innocent-looking objects. In images or video, for example, this can be done by adjusting the color (intensity) of certain pixels in a way that is imperceptible to humans. In audio, this can be done, for example, by adding a low or high-frequency tone (inaudible to humans). For example, a company offering/creating AI-generated content may choose to give certain pixels in an image a specific color, marking the image and making it traceable as "AI-generated.

Waterproof?

When a company works with fixed patterns that are added in the same way and place every time, sooner or later they will be traceable and ways can be developed to remove them, negating their intent.
The latter can be combated by randomizing the marking. The most promising watermarking techniques that apply this and are currently available are "statistical watermarking" and "machine learning-based watermarking. These advanced watermarks contain subtle patterns that apply markings to the file using assignment keys and algorithms. These markings can only be deciphered and detected by tools using the same keys used to mark the file.

LEGISLATION

The rapid rise of generative AI has prompted policymakers to develop regulation to manage innovation. Below, we provide a brief (non-exhaustive) overview of some of the measures policymakers have taken recently regarding methods of reporting AI-generated content.

> EU AI Regulation: The AI Regulation imposes various obligations on providers and deployers of AI systems to enable the detection of AI-generated content (excluding use for personal activities). This refers to both AI systems that generate "synthetic content" and systems used by deployers to form "deepfakes." The two obligations are not exclusive. The information below must be communicated to the natural person in a clear, accessible and distinguishable manner, upon their first interaction with the content.

- Providers of AI systems, including GPAI models and systems, that generate synthetic audio, video, images or text must label the content in a machine-readable format and detectable as artificially generated (or manipulated). There is an exception for AI systems that perform an assistive function for "standard" editing (e.g., spell check) or that do not substantially alter input data.
- Deployers of AI systems that can generate or manipulate deepfake images, audio or video must (clearly and distinguishably) disclose that the content was generated or manipulated by AI. It seems clearer from the recitals that these markings must be visible to humans. For a deepfake that is part of an artistic work, the notification can be made in another appropriate way that does not interfere with the enjoyment of the work.
- Deployers of AI systems that generate or manipulate text to publish for the purpose of informing the public about matters of public interest must disclose that the text was generated or manipulated by AI, except if the text is human-reviewed or editorially controlled, and if someone has editorial responsibility for the text.

The Knowledge Centre Data & Society has created prototype disclaimers of the transparency requirements for AI systems that generate deepfakes. You can find all the information about them here.