



The advent of generative AI: how can we protect ourselves from misinformation in upcoming elections?

📅 04/06/2024

📌 EU AND COMPETITION, PRIVACY AND CYBERSECURITY, SOCIETY

Jacopo Piemonte
Federico Aluigi

The rapid and remarkable progress in the capabilities of generative artificial intelligence (AI) to create all kinds of media, such as text, images, audio, and video, in response to specific inputs, known as prompts, continues to astonish us. In a short period of time, AI systems have become so advanced that they can produce content virtually indistinguishable from the reality, especially considering the immediacy with which users often engage with social media.

After the fake news at the dawn of the social media era, the probably even more fearsome phenomenon of deep fakes has thus emerged. Technologies now

make it possible to create completely fake videos in which, for example, statements can be falsely attributed to candidates for the elections. This can influence and manipulate the opinions, behaviour, and trends of global society. Hence, there is concern and interest among political leaders, jurists, and legislators in curbing this phenomenon, especially in a record year like 2024, which sees some 2 billion individuals worldwide called to the polls¹.

A pact to save the elections

The topic was the focus of the technology section at the *Munich Security Conference (MSC)*², held in February 2024. It is the world's most important forum for debate and

¹ More information in our previous article at the following [LINK](#).

² For more information see the following [LINK](#).



discussion on security and foreign policy, hosting leading figures from the political, legal, military, civil society, and technology fields.

Among the most important outcomes of this year's MSC is a document titled "*A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*", signed by 20 of the most influential multinationals in the IT industry, including Amazon, Adobe, Google, Microsoft, Meta, OpenAI, and IBM. The agreement is aimed at countering the use of generative AI in ways that mislead or deceive voters by creating "*AI-generated audio, video, and images that falsify or deceptively alter the appearance, voice, or actions of political candidates, election officials, and other key figures in a democratic election, or that provide false information to voters about when, where, and how they can legitimately vote*"³.

The multinationals party to the agreement propose, on the one hand, to cooperate in developing detection technologies that certify the authenticity, veracity, and origin of content. On the other hand, in managing their own platforms, the Big Tech companies aim to use these tools to identify and counter the spread of disinformation, such as deep fakes.

Watermarks between innovation and technological limitations

The objective of the MSC is simpler in theory than in practice since, to date, technologies with a sufficient degree of automation and certainty to deal effectively with the problem do not exist. Among the available technologies, the most promising are so-called watermarks, which consist of a mark capable of unequivocally identifying the provenance of a file, so that the end user can understand its source⁴.

Watermarks can be categorized into two main types: "visible" watermarks and "invisible" watermarks. Visible

watermarks are labels or graphics that are overlaid directly on the content, while invisible watermarks are embedded within it in such a way that they cannot be detected by humans. Examples of invisible watermarks include specific information embedded in file metadata, microscopic pixels added to an image, or inaudible sounds in an audio track. Specific software is always required to detect invisible watermarks.

Technical standards have been developed for implementing watermarks, such as those proposed by the Adobe Coalition for Content Provenance and Authenticity (C2PA)⁵. This group, which includes several major IT companies, requires developers of generative AI software to integrate a reference within the file metadata that is automatically recognizable by the platforms on which the file will later circulate.

However, these technologies are far from being mature and still require further development and improvement. Currently, watermarks can be easily circumvented through actions such as cropping or screenshotting an image, or by using generative AI software that does not adhere to the relevant technical standards. Participation in groups like C2PA is not compulsory, and numerous open-source generative AI software not subject to these constraints are widespread. Consequently, those who wish to disseminate a fake video on the Internet need only avoid adhering to the (non-mandatory) standards mentioned.

Watermark against disinformation: Meta's approach

The good news, however, lies in the willingness among major market players to self-regulate and act in a virtuous spirit of compliance.

In an attempt to lead on these issues, the well-known multinational Meta adopted initial guidelines in February 2024, stipulating that any content manipulated

³ For more information see the following [LINK](#).

⁴ For more information see the following [LINK](#).

⁵ For more information see the following [LINK](#).

by means of generative AI would be removed from its platforms. However, Meta itself realized afterwards that it was probably more appropriate to adopt a less strict and more proportionate approach, recognizing that "*generative AI is becoming a mainstream tool for creative expression*".

Hence, the revision of the guidelines in April 2024, in which Meta stated that only content violating the platform's terms of use or considered most dangerous for users would be deleted⁶. The new guidelines signal the US company's intention, as of May 2024, to include watermarks on its platforms (*i.e.*, Instagram, Facebook, and Threads) to alert users to content produced by generative AI. Content to be tagged will be automatically identified and flagged by the platform if the generative AI software that produced it adheres to recognized technical standards such as Adobe's C2PA. Otherwise, Meta specifies that users of its platforms will be asked to self-flag such content.

The race to regulate generative AI

Having said all the above, there is no doubt that there is a growing awareness of the need to regulate and monitor the use of generative AI. Indeed, several regulations on this matter have been implemented in different parts of the world.

The most decisive approach seems to come from China, which, as of August 2023, has imposed by law that anyone programming and making generative AI software available to the public must include, on the one hand, a label indicating the origin of the content and, on the other hand, "implicit" metadata and watermarks⁷.

With less timeliness, US President Joe Biden also recently (October 2023) signed an executive order on AI that

requires the administration to develop effective mechanisms for tagging and sourcing content⁸. There are also calls for Congress to pass legislation on AI-generated content in 2024 to ensure that regulations on watermarking mechanisms are enforced.

The European Union has also intervened on this point with the famous AI Act. Article 50 of this legislation requires providers of generative AI systems to ensure that the file created contains a mark detectable by special software that indicates its origin. Additionally, software users are required to report whether the content they are disseminating or using is a deepfake⁹.

Finally, even the United Nations (UN) has spoken out on the subject, with a General Assembly resolution - non-binding - adopted in March 2024 and supported by more than 120 states, which encourages the development of effective, reliable, and accessible technologies to distinguish AI-generated material from human material. On a more general level, the resolution, the first on the subject by the UN, aims to create secure and reliable AI systems that can effectively support global sustainable development¹⁰.

Beyond deception: the necessary safeguards

In conclusion, the advent of generative AI is radically redefining the landscape of digital content production and consumption. Tools such as ChatGPT, DALL-E, and others are enabling users to generate a wide range of content, from images to text, in an increasingly realistic manner. However, this rapid evolution brings with it crucial challenges. The confusion between AI-generated and human-created content raises significant concerns about the spread of misinformation as well as privacy and intellectual property issues.

⁶ For more information see the following [LINK](#).

⁷ For more information see the following [LINK](#).

⁸ For more information see the following [LINK](#).

⁹ For more information see the following [LINK](#).

¹⁰ For more information see the following [LINK](#).

The need for transparency in distinguishing “synthetic” AI-generated content from human content is emerging as a central issue. Policymakers and practitioners are faced with the challenge of identifying innovative solutions to ensure that users can discern what is in front of them, such as videos on social media, in a context where technological resources have not yet reached full maturity. For this reason, it will be crucial


to move towards the ethical use of AI: while embracing innovation, we must not forget the importance of preserving integrity and trust in the digital environment. Through the adoption of transparent approaches, accountability systems, and protection tools such as watermarking, we can guide AI towards a future where creativity can flourish without compromising the honesty and authenticity of content.





Jacopo Piemonte

ASSOCIATE

 j.piemonte@dejalex.com

 +39 02 72554.1

 Via San Paolo 7
20121 – Milano

 +32 (0)26455670


 Chaussée de La Hulpe 187
1170 – Bruxelles



Federico Aluigi

ASSOCIATE

 f.aluigi@dejalex.com

 +32 (0)26455670

 Chaussée de La Hulpe 187
1170 – Bruxelles

MILANO

Via San Paolo, 7 · 20121 Milano, Italia
T. +39 02 72554.1 · F. +39 02 72554.400
milan@dejalex.com

ROMA

Via Vincenzo Bellini, 24 · 00198 Roma, Italia
T. +39 06 809154.1 · F. +39 06 809154.44
rome@dejalex.com

BRUXELLES

Chaussée de La Hulpe 187 · 1170 Bruxelles, Belgique
T. +32 (0)26455670 · F. +32 (0)27420138
brussels@dejalex.com

MOSCOW

Potapovsky Lane, 5, build. 2, 4th floor, office 401/12/9 · 101000, Moscow, Russia
T. +7 495 792 54 92 · F. +7 495 792 54 93
moscow@dejalex.com