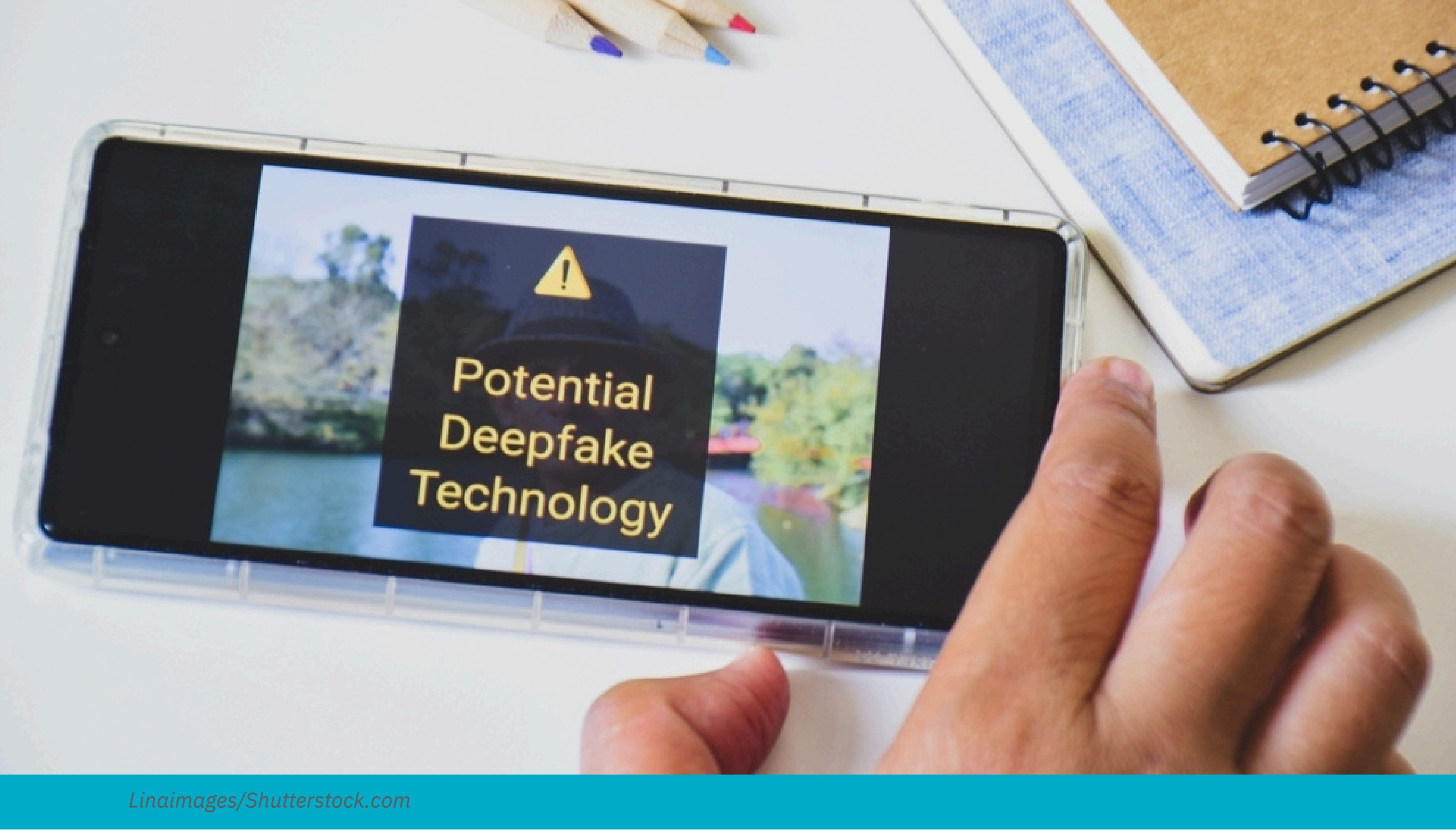


Linaimages/Shutterstock.com

Make Your Mark (On “Deepfakes”)

Wilson Center | Science and Technology Innovation Program

Author: Madison Binder and Elizabeth Newbury

Editorial review provided by: Kara Struckman, Sonja R. O’Brien and Kellee Wicker

This whitepaper was in part informed by interviews with 17 experts. Thank you for sharing your time and expertise.

The interviews were conducted by Kara Struckman and Elizabeth Newbury.

Introduction

There are currently few possible solutions to stop the spread of deepfakes and the harms they perpetrate. We explored these mitigation techniques by conducting interviews with leading experts in the field of synthetic media to inform the game *The Deepfake Files*. Out of an array of options, one set of solutions surfaced time and again from interviews was 'marking' synthetically generated media, namely either watermarking or provenance marking. This paper will explore some of the key themes about watermarking and provenance that arose from the interviews, specifically similarities and differences, and detailing the strengths and weaknesses of each solution. It will then outline the necessity of a multipronged effort for any mitigation strategy to be effective, focusing on the need for better education and understanding of these techniques.

Marking techniques frequently emerged as a point of discussion not because of the necessary robustness of watermarking or provenance, but because of their dominance in discourse across stakeholders as a solution to nefarious deepfakes. The Verge called watermarking one of the "most hyped solution[s] to many of the social problems posed by generative AI." Watermarking and provenance marking techniques are rapidly being adopted by tech companies, including Google, Meta, and Microsoft. Recent legislation proposed in Congress, such as the Content Origin Protection and Integrity from Edited and Deepfaked Media Act, highlight both watermarking and provenance marking as a means of mitigating the spread of deepfakes. Despite the attention given to either technique, our interviewees repeatedly highlighted their limitations. One expert even referenced them as "rusty bullets," because they come nowhere near being a silver bullet.

A quick note before going further regarding the term "deepfake." Here, you will see that we use "deepfake" alongside the term "synthetic media." Synthetic media refers to a form of media (pictures, video, audio, text, etc.) that is created at least in-part by artificial intelligence (AI)/machine learning tools. Synthetic media is neither good nor bad. To learn more about this read The Positive Use Cases of "Deepfakes". However, in most contexts, when the term "deepfake" is used, it has a malicious or nefarious association. For the purpose of this paper, when we use the term "deepfake," we are referencing synthetic media that is considered harmful for one reason or another.

Methodology

This paper uses two main sources to understand synthetic content and is part of a larger research portfolio used to create the game, *The Deepfake Files*. First, as previously noted, we interviewed seventeen experts in the field to understand the ways in which deepfakes could be mitigated broadly. These experts included cybersecurity experts, computer science researchers, nonprofit leaders, and government employees and were primarily based in the United States. Interviews were confidential and semi-structured, and lasted approximately thirty minutes each.

In doing this research, we found consistent themes arise, encapsulating ways in which the effects of deepfakes could be mitigated, both in a technical and non-technical capacity. Many of these common themes are reflected in the game, *The Deepfake Files*.

To explore what our experts highlighted, this paper also relies on a second mode of data, namely analysis of peer-reviewed journals, popular press, and similarly vetted resources.

Watermarking

In the broadest technical terms, watermarking is a disclosure method requiring someone to actively apply a marker that 'tags' and identifies a piece of media as synthetic or non-synthetic. Watermarking can be used for two purposes–either to denote that a piece of media is synthetic to some extent or to indicate that a piece of media is fully non-synthetic. Most commonly thought of in reference to visual media (pictures, videos), watermarks can also be applied to pixels or audio-waves that are beyond human perception but can be picked up by other algorithms.

Watermarking is embedded in a piece of media by a watermarking algorithm, and can be visible or invisible. Of the two forms of watermarking, visible watermarking is proven to be the least secure, as it is the most susceptible to manipulation or removal. Invisible watermarking is embedded in a piece of media in a way that is not visually perceptible, such as altering an image's pixels in a way that does not distort the image but can be perceived by a detection algorithm. Invisible watermarking relies on a second detection algorithm which receives a piece of content and checks if it contains a watermark. This method is more secure, but less useful or accessible to a casual user.

Strengths

The primary strength of watermarking, expressed across our interviews, is that it can be an easy way to help people with identification. Watermarking can indicate to a viewer whether content is marked as synthetic or non-synthetic.

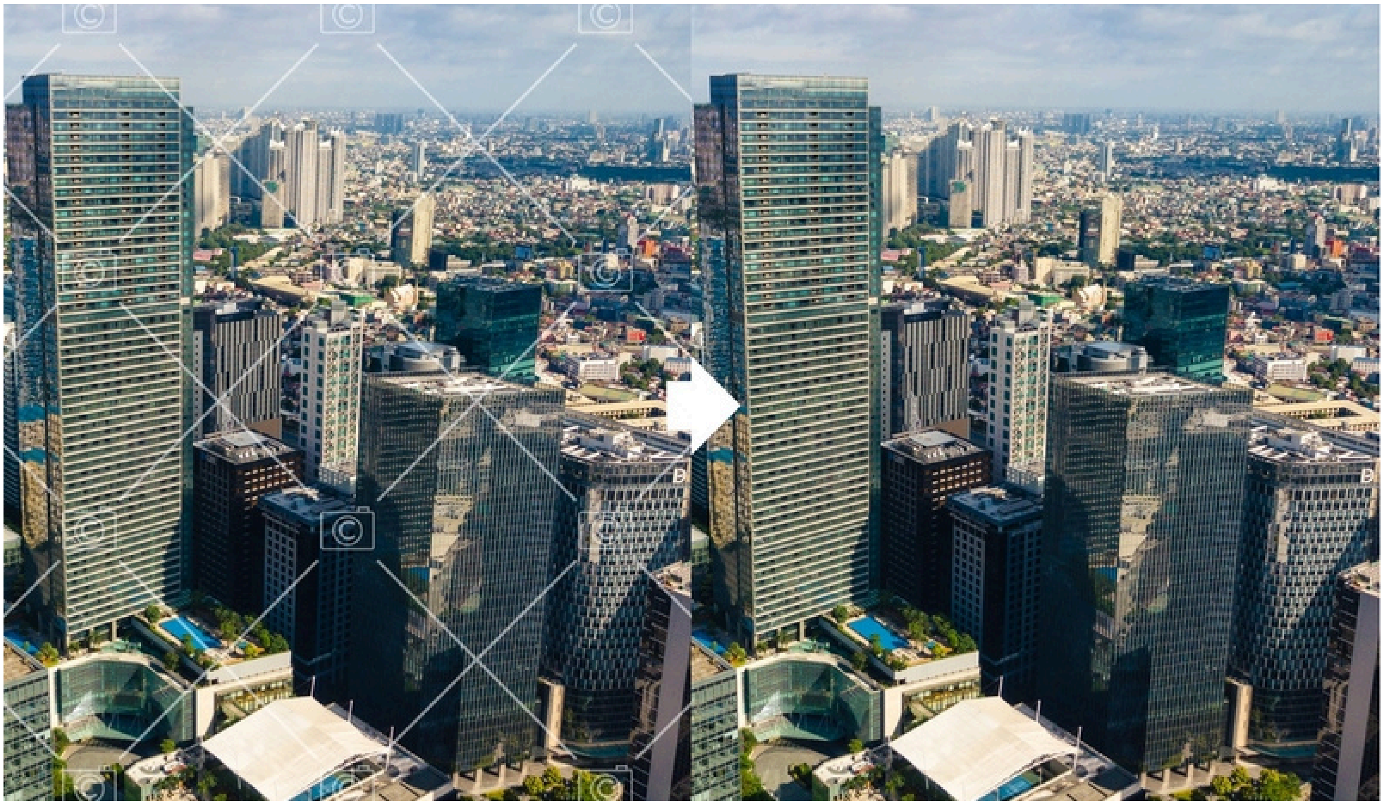
Additionally, watermarking is one of the few technical mitigation techniques we have readily available and is more easily scalable than other strategies, as there are widely available services for watermarking and they can be applied at any point throughout a content's life cycle. Watermarking has received a fair amount of attention, not only in popular media but also in public policy. Governmental bodies from the European Union to China have passed legislation that targets watermarking; countries such as Australia and the United States have also passed or proposed legislation around mitigating deepfakes. Indicating that globally legislators support watermarking as a mitigation strategy. But for all the attention it has gathered, most experts will point to the tactic's significant weaknesses.

Weaknesses

Similar to other mitigation techniques, watermark's weaknesses can boil down to the ease of manipulation, explainability, lack of standardization, and interoperability.

Across interviews, there was consensus that, at least with current technology, we can make it difficult to remove watermarks–but not impossible. In the context of harm reduction approaches, some interviewees felt that the 'average person' might find it difficult to remove a watermark, but it would not deter motivated malicious actors. A simple Google search will turn up tutorials on how to remove digital watermarks. As watermarking can be applied to synthetic and non-synthetic material, it is important to recognize the difference between who bears the burden of labelling; good faith actors vs. bad faith actors. Some experts argue that marking non-synthetic content is more effective as this does not rely on the assumption that individuals creating synthetic material, like deepfakes, will disclose that the content is synthetic.

Before-and-after example of an AI tool removing watermarks from a stock city photo. *(MDV Edwards/Shutterstock.com)*



Additionally, watermarks lack specific explainability on what they are disclosing. While watermarks can identify a piece of media as synthetic, they do not necessarily specify what has been altered. Manipulations may range from fully generating an image, to something as benign as a red eye remover. While both involve the use of AI, watermarks may simply label them as being AI manipulated, without any explainability of the extent.

A lack of standardization also poses a challenge for watermark detection tools. As there is no single standard for watermarking, actors can employ different techniques, which, in turn, require different detection algorithms for authentication. There is variation alone in the use of visible and invisible watermarks, and within those categories there is further variation in how the techniques are applied. For example, in 2023, researchers at Meta released Stable Signature, a new method for watermarking which "leaves a secret binary signature into all images generated by latent diffusion models, like Stable Diffusion." This method is resistant to manipulations like cropping, or changing the color. However, Stable Signature can only be applied to latent diffusion models, and would not be compatible with any version of ChatGPT, for example. Another set of researchers at Hugging Face and Google's Deepmind have released technology called SynthID Text, which applies a watermark to AI-generated text. SynthID uses a type of watermarking that modifies the output of the model by creating a "statistical signature in the generated text while maintaining its quality." This can then be detected by a classification algorithm which determines if text is AI generated or not. However, without standardization not all detection algorithms are trained to detect these signatures and there are many open source models which do not produce any watermark at all and are unlikely to adopt this technology. Moreover, utilizing multiple marking regimes will make it more difficult for people to understand what to look for.

The final weakness stems from challenges with interoperability. As one of our interviewees pointed out, sharing media across social media platforms, websites, and even devices often alters the content in ways that are imperceptible to the human eye but can alter or tamper with watermarks in part or fully negating their effectiveness. For example, simply screenshotting an image carrying Meta's AI generated image watermark bypasses detection. Researchers at UC Berkeley found that after generating an image using Meta's AI, they were able to screenshot the image, upload the screenshot to an AI detection algorithm, which was not able to identify that it was AI generated.

Overall, the sentiment from our experts was that watermarks needed to be part of a broader system of approaches, both technical and cultural. They are worthwhile to pursue, but not a complete solution on their own.

Provenance

While watermarking can be applied at any point in a contents life cycle, provenance marking, instead, involves the creation of a digital ledger at the point of origin, which then tracks the history of the content. Provenance markings are a cryptological marking, and the "digital lineage" which shows the history of the content includes information such as the location and date of creation and any changes made across its distribution. This data is then available to the viewer, similar to metadata, and is meant to provide authenticity by identifying the genesis of the content and all alterations.

Strengths

One of the strengths of provenance marking is the digital ledger. Rather than simply identifying a piece of media as either synthetic or non-synthetic, the ability to follow the chronology of the content allows for the identification of any alterations. This differs from watermarks which simply indicate whether content is synthetic or non-synthetic, while provenance markings allow users to see specifically what alterations were made. It is important to note that alterations are not necessarily synthetic or malicious and could include things like cropping an image or adjusting the saturation. The ability to identify missing or suspicious information in the ledger can be an indicator of manipulation and is an additional strength of provenance markings because it gives the user the tools to see whether an actor tampered with the ledger.

The digital lineage of provenance marking can be important not only in identifying synthetic content, but also in identifying other forms of manipulation. Maliciously motivated actors are already seeking to exploit uncertainty by utilizing real content taken out of context, in other words, altering it without using AI. For example, a speech by Nancy Pelosi was slowed down to make it appear as if she was drunk and slurring her speech. This change did not utilize AI but was a manipulation of the original video. This alteration would still be tracked by the digital ledger.

Beyond individuals being able to trace the lineage of the content they view, provenance solutions could also help digital content hosting platforms prevent the spread of content with missing or suspicious provenance information. Social media companies, cloud services, web hosting services, and news organizations could prevent users from uploading and posting "noncompliant text, photos, and videos that are missing key provenance information" or flag that content as potentially synthetic or manipulated. By doing so, this would help reduce the spread of malicious content, or at the minimum allow users to be notified that content they are viewing may have been manipulated.

Weaknesses

While provenance markings have many strengths, it is still an incomplete solution on its own. Key weaknesses of provenance markers include lack of standardization and the ability for manipulation or removal.

Although provenance marking can enable detection mechanisms for content hosting platforms, there are limitations to these tools without standardization or universal adoption. Standardization would allow for greater effectiveness of these tools. However, scalability barriers exist in terms of getting broad industry and international adoption. Initiatives like the Coalition for Content Provenance and Authority (C2PA), the Origin Project, and the Content Authenticity Initiative are working to create industry standards for content provenance.

These joint industry efforts do not seek to pass negative or positive judgements on content but determine simply whether “assertions included within can be verified as associated with the underlying asset, correctly formed, and free from tampering.” Standards would better enable detection tools and interoperability across devices, platforms, and systems. Additionally, standards can make it easier for people engaging with content to understand what they are viewing and know how to verify content.

Provenance markings require much more technical ability to manipulate compared to watermarks, however, some of our interviewees still brought up concerns about the resilience of provenance markers to manipulation from motivated malicious actors. While provenance marking solutions that utilize technology like blockchain to create ledgers of any changes to the original content are more difficult to manipulate–it is not impossible. Truly motivated malicious actors are likely to “improve in their ability to obfuscate AI content detectors and manipulate provenance solutions,” to circumvent protection mechanisms. For example, while making detection tools publicly available, it can help expand their adoption, however, it can also allow malicious actors to adopt methods to avoid detection.

As with watermarking, participants viewed provenance marking as one piece of a broader network of mitigating solutions, rather than a solution on its own. Provenance markers can be an important step to validating authenticity and content origins, however it will not be an effective measure without the support of other tools.

Futuristic digital grid. *(Chris WM Willemsen/Shutterstock.com)*

Education

Our interviewees repeatedly noted the importance of education and building up trusted sources. There are two main aspects of education in this context–first, ensuring that people understand watermarks and provenance marks and how to engage with them and second, building up a society which actively, critically, and continuously engages with the content they consume.

The first aspect involves teaching people how to navigate the provenance and watermarks. The lack of a watermark or digital ledger does not mean that the content is synthetic or nonsynthetic/true or false. The lack of a marking carries no meaning about authenticity. There is simply so much content out in the world that it is impossible for all content to receive a watermark or provenance signal. These tools will not only need to be accompanied with education campaigns to teach users how to interpret watermarks and provenance but also with education initiatives that teach users not to discredit content on the basis of it not having a marking. Even with a marker, users still need to validate content. As this paper has outlined in both the weaknesses of watermarking and provenance marking, these techniques remain vulnerable to manipulation. This is why education is essential.

Education plays an essential role in equipping people with the skills and knowledge to critically engage with the content they are seeing. As there is no silver bullet solution to mitigate deepfakes, it is important to build a society that is actively engaging with the content they consume. As one interviewee expressed, most people make snap judgements about the content they come across. They don’t necessarily dig further or investigate whether it can be authenticated. Accessing provenance data is similar to accessing metadata, which is a task that the “average” individual viewing digital content does not engage in–education programs can help empower people to use tools available to them. Overall, education and building trusted sources is intended to equip people with the skills and tools necessary to navigate the digital landscape and come to informed conclusions about the content they interact with.

Conclusion

While watermarking and provenance markers have a role to play in deepfake mitigation, any effective solution will not work alone. Mitigating techniques are required in conjunction to create what one interviewee referred to as “pipeline responsibility.” Measures need to be taken throughout the life cycle of content–beginning at the training data and AI model stage, and extending all the way to consumers' interactions with content. Layering techniques will allow for the most effective mitigation strategy, leveraging accountability and transparency early on in the digital life cycle and continuing to add measures as the content progresses. Markers are just one piece of the puzzle.

Wilson Center | Science and Technology Innovation Program

Woodrow Wilson International Center for Scholars
One Woodrow Wilson Plaza
1300 Pennsylvania Avenue NW
Washington, DC 20004–3027

The Wilson Center

🌐 wilsoncenter.org
f woodrowwilsoncenter
𝕏 @TheWilsonCenter
◎ @thewilsoncenter
in The Wilson Center

Science and Technology Innovation Program

🌐 wilsoncenter.org/science-and-technology-innovation-program
𝕏 @WilsonSTIP
in linkedin.com/showcase/science-and-technology-innovation-program