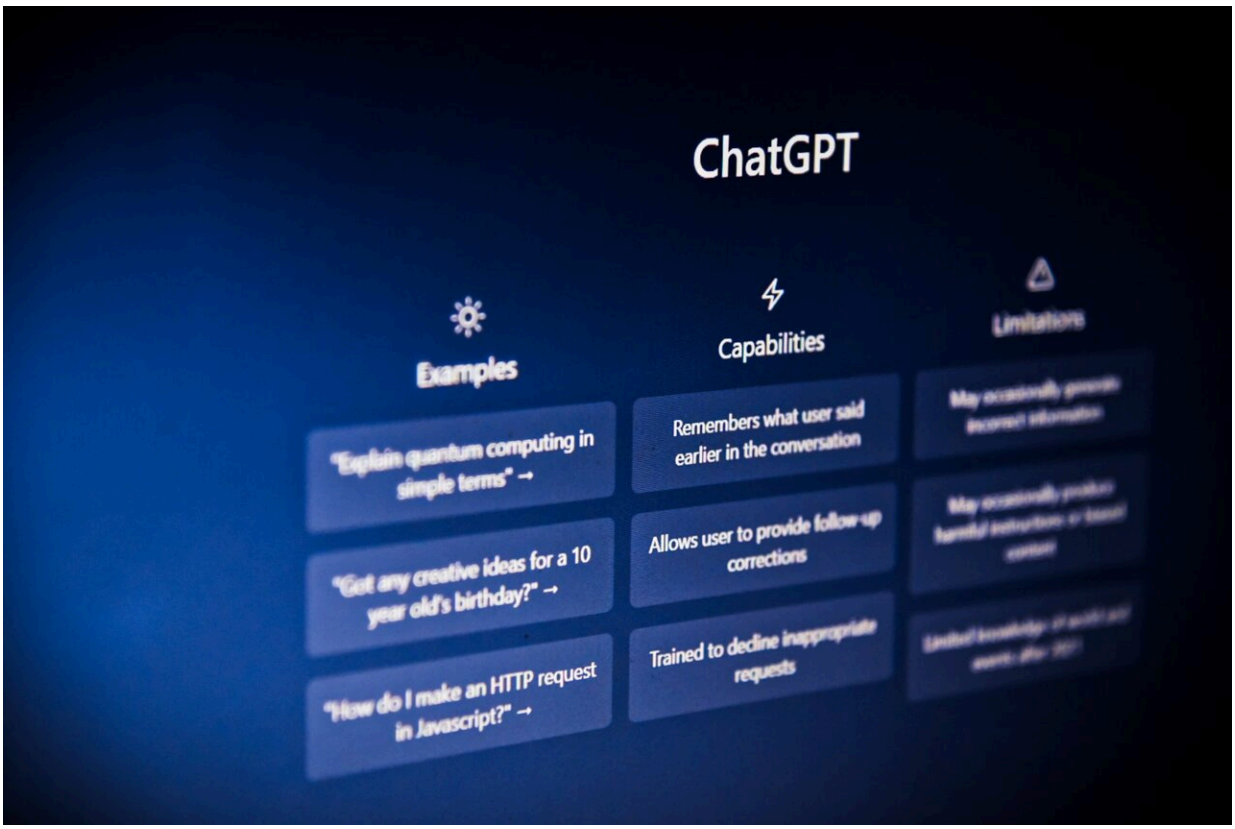
Watermarking ChatGPT, DALL-E and other generative AIs could help protect against fraud and misinformation

March 27 2023, by Hany Farid



Credit: Unsplash/CC0 Public Domain

Shortly after rumors leaked of former President Donald Trump's impending indictment, images purporting to show his arrest appeared

online. These images looked like news photos, but they were fake. They were created by a generative artificial intelligence system.

Generative AI, in the form of image generators like DALL-E, Midjourney and Stable Diffusion, and text generators like Bard, ChatGPT, Chinchilla and LLaMA, has exploded in the public sphere. By combining clever machine-learning algorithms with billions of pieces of human-generated content, these systems can do anything from create an eerily realistic image from a caption, synthesize a speech in President Joe Biden's voice, replace one person's likeness with another in a video, or write a coherent 800-word op-ed from a title prompt.

Even in these early days, generative AI is capable of creating highly realistic content. My colleague Sophie Nightingale and I found that the average person is unable to reliably distinguish an image of a real person from an AI-generated person. Although audio and video have not yet fully passed through the uncanny valley—images or models of people that are unsettling because they are close to but not quite realistic—they are likely to soon. When this happens, and it is all but guaranteed to, it will become increasingly easier to distort reality.

In this new world, it will be a snap to generate a video of a CEO saying her company's profits are down 20%, which could lead to billions in market-share loss, or to generate a video of a world leader threatening military action, which could trigger a geopolitical crisis, or to insert the likeness of anyone into a sexually explicit video.

Advances in generative AI will soon mean that fake but visually convincing content will proliferate online, leading to an even messier information ecosystem. A secondary consequence is that detractors will be able to easily dismiss as fake actual video evidence of everything from police violence and human rights violations to a world leader burning top-secret documents.

As society stares down the barrel of what is almost certainly just the beginning of these advances in generative AI, there are reasonable and technologically feasible interventions that can be used to help mitigate these abuses. As a computer scientist who specializes in image forensics, I believe that a key method is watermarking.

Watermarks

There is a long history of marking documents and other items to prove their authenticity, indicate ownership and counter counterfeiting. Today, Getty Images, a massive image archive, adds a visible watermark to all digital images in their catalog. This allows customers to freely browse images while protecting Getty's assets.

Imperceptible digital watermarks are also used for digital rights management. A watermark can be added to a digital image by, for example, tweaking every 10th image pixel so that its color (typically a number in the range 0 to 255) is even-valued. Because this pixel tweaking is so minor, the watermark is imperceptible. And, because this periodic pattern is unlikely to occur naturally, and can easily be verified, it can be used to verify an image's provenance.

Even medium-resolution images contain millions of pixels, which means that additional information can be embedded into the watermark, including a unique identifier that encodes the generating software and a unique user ID. This same type of imperceptible watermark can be applied to audio and video.

The ideal watermark is one that is imperceptible and also resilient to simple manipulations like cropping, resizing, color adjustment and converting digital formats. Although the pixel color watermark example is not resilient because the color values can be changed, many watermarking strategies have been proposed that are robust—though not

impervious—to attempts to remove them.

Watermarking and AI

These watermarks can be baked into the generative AI systems by watermarking all the training data, after which the generated content will contain the same watermark. This baked-in watermark is attractive because it means that generative AI tools can be open-sourced—as the image generator Stable Diffusion is—without concerns that a watermarking process could be removed from the image generator's software. Stable Diffusion has a watermarking function, but because it's open source, anyone can simply remove that part of the code.

OpenAI is experimenting with a system to watermark ChatGPT's creations. Characters in a paragraph cannot, of course, be tweaked like a pixel value, so text watermarking takes on a different form.

Text-based generative AI is based on producing the next most-reasonable word in a sentence. For example, starting with the sentence fragment "an AI system can…," ChatGPT will predict that the next word should be "learn," "predict" or "understand." Associated with each of these words is a probability corresponding to the likelihood of each word appearing next in the sentence. ChatGPT learned these probabilities from the large body of text it was trained on.

Generated text can be watermarked by secretly tagging a subset of words and then biasing the selection of a word to be a synonymous tagged word. For example, the tagged word "comprehend" can be used instead of "understand." By periodically biasing word selection in this way, a body of text is watermarked based on a particular distribution of tagged words. This approach won't work for short tweets but is generally effective with text of 800 or more words depending on the specific watermark details.

Generative AI systems can, and I believe should, watermark all their content, allowing for easier downstream identification and, if necessary, intervention. If the industry won't do this voluntarily, lawmakers could pass regulation to enforce this rule. Unscrupulous people will, of course, not comply with these standards. But, if the major online gatekeepers—Apple and Google app stores, Amazon, Google, Microsoft cloud services and GitHub—enforce these rules by banning noncompliant software, the harm will be significantly reduced.

Signing authentic content

Tackling the problem from the other end, a similar approach could be adopted to authenticate original audiovisual recordings at the point of capture. A specialized camera app could cryptographically sign the recorded content as it's recorded. There is no way to tamper with this signature without leaving evidence of the attempt. The signature is then stored on a centralized list of trusted signatures.

Although not applicable to text, audiovisual content can then be verified as human-generated. The Coalition for Content Provenance and Authentication (C2PA), a collaborative effort to create a standard for authenticating media, recently released an open specification to support this approach. With major institutions including Adobe, Microsoft, Intel, BBC and many others joining this effort, the C2PA is well positioned to produce effective and widely deployed authentication technology.

The combined signing and watermarking of human-generated and AI-generated content will not prevent all forms of abuse, but it will provide some measure of protection. Any safeguards will have to be continually adapted and refined as adversaries find novel ways to weaponize the latest technologies.

In the same way that society has been fighting a decadeslong battle

against other cyber threats like spam, malware and phishing, we should prepare ourselves for an equally protracted battle to defend against various forms of abuse perpetrated using generative AI.

Provided by The Conversation