Artificial Intelligence

Keeping Deepfakes Out of Court May Take Shared Effort

Court officials anticipate having messy debates over whether evidence is authentic or fabricated, with deepfakes skewing jurors' decisions and digital forensics analysts helping to find the truth.

Jule Pattison-Gordon

January 24, 2024

https://www.govtech.com/artificial-intelligence/keeping-deepfakes-out-of-court-may-take-shared-effort

Shutterstock

As artificial intelligence evolves and digital fakery becomes both more pervasive and more convincing, court officials are already thinking about how to stop it from unduly influencing legal proceedings.

No solution will be foolproof, but experts say the time has come to start preparing guardrails and considering countermeasures. Members of judicial and tech spaces alike are sounding this alarm about the possibility — and probability — that deepfaked evidence could soon show up in courts. If juries fall for fabrications, they'd base decisions on falsehoods and unfairly harm litigants. And real images and videos could be mistakenly discounted as fakes, causing similar damages.

Evidence must be proven to be more likely to be authentic than not before a judge will admit it for the jury's consideration. That's a new problem in the era of

generative AI, where studies suggest jurors are likely to be biased by video evidence even when they know it might be a fabrication.

“Even when jurors are aware that the audiovisual evidence could be fake, studies have shown, they still tend to align their perception and memories to coincide with what they saw or heard,” said Maura Grossman, research professor in the University of Waterloo School of Computer Science, during an American Bar Association webinar this week.

During the webinar, retired judge Paul Grimm urged judges to schedule hearings well in advance of trial, where parties would have time to bring in experts and hash out disputes over the authenticity of pieces of evidence that are likely to impact a case.

Still, those disputes could be challenging to resolve.

Litigants or courts can get digital forensic analysts to weigh in, but such work, while helpful, can be expensive and time-consuming, said Hany Farid, digital forensics and misinformation expert at the University of California, Berkeley School of Information.

Digital forensic analysts can look for visible and invisible clues in an image. Extra fingers and distorted eyes are hallmarks of obvious deepfakes, but there are subtler signs of sophisticated fabrications, too.

For example, AI often messes up shadows, Farid said. It also tends to fail to make parallel lines in the image properly converge to a vanishing point.

Meanwhile, face-swaps usually only copy a portion of the target’s face — the section from cheek-to-cheek and eyebrow-to-chin — while failing to replicate other distinctive details like the ear. Paying attention to the inner-ear shape and attachment or detachment of earlobes can help when comparing people to potentially fake photos. And faked videos may fail to pair spoken words with the actual mouth shapes needed to pronounce them.

Experts are also preparing for political deepfakes by analyzing the idiosyncrasies of how specific high-profile figures speak and move. Knowing how President Biden moves his head or places emphasis while talking better readies experts to find fake videos of him.

Hany Farid discusses how deepfaked and authentic images display vanishing points, during an American Bar Association webinar.

Other clues are embedded beyond visual and audio perception, and metadata can be revealing. For example, the way an AI system creates a compressed JPEG is different from a camera, Farid said. And machine learning-powered tools trained on real and deepfake images can analyze files for indicative patterns.

Still, such detection efforts only go so far.

“A good forensic analyst has dozens and dozens of techniques, where we can take our time, particularly when it comes to introduction of evidence into courts of law," Farid said. “But it's hard; it takes time. And experts are fairly few and far between.”

That’s where tech companies’ efforts to proactively sort real from fake could make a difference.

Generative AI companies could design systems so that when images are produced, the systems attach watermarks disclosing that it’s fabricated content, or imbed invisible watermarks readable by software. Not all attackers will try to thwart both watermarks, Farid said.

Watermarking has caught federal attention. Last July, the White House announced that seven AI companies had voluntarily promised to develop mechanisms for indicating if content was created by AI. Some federal legislators

also pushed for more binding and broad-sweeping measures, introducing bills that would require AI content to bear watermarks or disclaimers.

Some companies are taking active steps. Companies participating in the Coalition for Content Provenance and Authenticity (C2PA) created icons that content editing and generating systems can embed into metadata. The icon would list details of the content's creation, including the AI tool used to make it.

This month, OpenAI announced it would implement C2PA watermarking on images created by DALL-E 3. Getting other major generative AI companies to sign on too, would go far, said Farid, who anticipated this could happen within months.

As another measure, the companies could take a unique hash or fingerprint of each image they generate and store it. The hash would preserve details about when the image was created and by whom. If such a practice became widespread, courts considering the veracity of an image could ask the companies to check whether it was created by AI.

"Then, when the image surfaces in a court of law or on social media, we can go back to OpenAI and say, 'Is this one of yours?' We can go to Midjourney and say, 'Is this one of yours?'" Farid explained.

Hany Farid (left) and retired Judge Paul Grimm (right) discuss the risks and challenges of potentially deepfaked evidence during the ABA webinar. screenshot

Watermarking doesn't just have to be on fabricated content, either. It can also be used to attest to authenticity. Ideally, manufacturers of cameras, smartphones and other recording hardware would adopt a system that attaches a cryptographic signature to metadata upon creation. The metadata would include

details like time and location, and it would record a history of whether the image was altered or updated.

Authenticity watermarking tools exist now, and the Content Authenticity Initiative (CAI) is one such effort. But the practice is unlikely to become widespread until much of the public adopts mobile devices designed with these features — and people often stick with their old devices for years, Farid said.

“I don't think this is a technological barrier. I think this is an implementation barrier,” Farid said.

He also advocated for making public cameras such as police body cams, dash cams and CCTVs compliant with digital watermarking.