Generative AI and the legal system

Maura R. Grossman, JD, PhD, is a Research Professor in the Cheriton School of Computer Science, cross-appointed to the School of Public Health Sciences, at Waterloo; an Adjunct Professor at York University's Osgoode Hall Law School; and an affiliate faculty member of the Vector Institute for Artificial Intelligence. She is also Principal at Maura Grossman Law, an eDiscovery law and consulting firm in Buffalo, New York.

Maura is known widely for her work on technology-assisted review (TAR), a supervised machine-learning method that she and her colleague, Cheriton School of Computer Science Professor Emeritus Gordon V. Cormack, developed to expedite the review of documents in high-stakes litigation and in systematic reviews for evidence-based medicine.

More recently, Maura has turned her attention to the implications of generative AI on the legal system. With retired District Judge Paul W. Grimm, Cheriton School of Computer Science Professor Dan Brown and undergraduate research assistant Molly Xu, she has coauthored a paper titled "The GPTJudge: Justice in a Generative AI World" that outlines new challenges that generative AI presents to the bench and bar.

What follows is a lightly edited transcription of a Q&A interview that explores the issues discussed at greater depth in the paper.

What is generative artificial intelligence?

Generative AI (or GenAI) is a specific subset of artificial intelligence that creates new content in response to a prompt provided by a human. The prompt can be multimodal, as can be the output. Someone can type a question and get a text answer, or they can create a voice clip, an image or a video — really any combination depending on the system used for both the input and the output.

Gen-AI creates content that is increasingly difficult to differentiate from content that humans create. The generated text reads well, the photos look authentic, the audio files sound real, and the videos look convincing.

Probably the best known of the various GenAI systems is ChatGPT, a chatbot released by OpenAI in November 2022, based on a large language model (LLM) that generates fluent text. ChatGPT can carry on a dialogue-style conversation where someone asks it a succession of questions to which it provides human-like responses. People can also ask ChatGPT to draft a homework assignment for them or make their resume more succinct and professional or describe a particular job and ask the chatbot to tailor their resume to that job's specific requirements. Likewise, they can ask ChatGPT to summarize the main points in a book chapter in language that's understandable to an eight-year-old. Or to write a rap song about Taylor Swift in the style of Eminem. ChatGPT can perform any or all these tasks quite well.

ChatGPT doesn't just create content in response to a prompt. It can generate content following specific guidelines — including replicating a particular style or tone — and this is among the reasons why it is increasingly difficult to distinguish its output from that of a human.

How do these generative AI models work?

LLMs use probabilities and statistics to predict the next word in a response based on their training data, which for the most part is content scraped from the Internet. The models may see thousands upon thousands of examples of sentences about a certain topic and the output they deliver is based on what has been most commonly written. That may be accurate, biased, or completely wrong, but the LLM will look for what has been most commonly said in its training data and replicate that.

Many of these tools are trained using two algorithms that compete with each other, a process developed in 2017, referred to as generative adversarial networks or GANs. A generative network creates content while a discriminatory network critiques that content and compares it against real content. The generative algorithm gets better as it gets more feedback from the discriminatory algorithm. These two networks work together such that as the discriminator improves, so too does the generator. That's what makes it so hard to distinguish human from AI-generated content.

As GenAI advances, it will become harder to know which output was created by GenAI and what was created by a human. It will blur the lines between real and generated content, between truth and falsehood, between fact and fiction. This has profound implications for the justice system.

Can AI evidence be admitted in court?

Let me back up a bit first to answer your question. AI evidence suffers the same challenges as most scientific evidence does in that the technology on which it is based can be beyond the ken of the judge or jury, so the parties have to bring in expensive experts to assess and explain it. With evidentiary material that's suspected to be a deepfake — AI-generated content that's not real — parties will need to bring in AI forensic experts. So now the courts are not just dealing with, say, a complicated product liability case where the judge or jury needs to understand the workings of a particular machine or a device to determine liability, but the authenticity of the factual evidence itself will be in question.

Say I claim that my supervisor left me abusive or inappropriate voicemail messages and here's a recording of the message he left as evidence that the events happened exactly as I say. In the past, this kind of case did not

normally require an expert to assess the evidence. As long as I could show that the recording was more likely than not to be my supervisor's voice, the evidence came in. But now experts will be needed in virtually every case because my supervisor can argue that I manufactured the voicemail message and the court will need to determine if the evidence is real or fake.

If a police officer uses a radar gun to assess a driver's speed, we know enough about that technology that a case disputing a speeding ticket doesn't typically need a radar expert to explain how the device works and if it's valid and reliable. But that's not the case with new technologies like AI.

Say a human resources professional used AI to determine who to interview for a job and the majority of the people excluded appear to belong to a certain racial or ethnic group. If this went to court, the HR professional's employer would likely need to bring in the system's developers to explain how the AI model was trained, what algorithms it used, what training data was used, how it was tested and validated, and so on, to show that the system was not discriminatory.

In this example, there's no question whether the evidence presented to the court is real or fake. The issue is, has the AI been tested and validated sufficiently to know it's valid, reliable, unbiased, and works as intended. That's probably the most straightforward issue to deal with. We're trying to determine if the AI is working as it should be and this is no different from any other technical or scientific evidence; we already have the legal tools for that.

Where we have a new challenge is with deepfakes. While there have always been cases of forgery or other manufactured evidence, we are operating at an entirely different level now because the tell-tale signs of a deepfake — for example, that the hands have six-fingers — are no longer there because the technology has improved so much.

Does GenAI provide new challenges to the justice system?
Yes, with GenAI we have a different issue altogether. We need to determine if purported deepfake evidence should be admitted in civil and criminal trials. It makes a huge difference whether or not the judge or jury gets to consider that evidence in deciding a case.

Part of the challenge for the justice system — at least the system in the United States with which I am most familiar — is that the standard for admissibility is low. Evidence simply has to meet a preponderance standard — meaning that the evidence is more likely than not what I say it is, the odds being slightly more than that of a coin toss.

With deepfake evidence that's a very low bar. Someone can play a recording of your voice in court and I can testify that I've spoken to you many times and know it's you in the recording because I recognize your voice. That's enough for such evidence to be admissible in a U.S. court. But just because it sounds like you doesn't mean the recoding is of something you said. With an AI voice-cloning tool anyone could generate an audio clip that sounds like you for free in minutes. This is the issue.

There's not a perfect fit between the current rules of evidence and the structure of the U.S. and Canadian justice systems for this new kind of evidence, which will increasingly show up in the courts. You say this is a real audio recording and I say it's not; it's a deepfake. That defence — referred to as the liar's dividend — will be made more and more often. Everything now has plausible deniability.

If the defendant's lawyer can show that at the time the alleged offending voicemail message was left on my phone (according to its electronic date and time stamp, referred to as metadata) that my supervisor was having open heart surgery, it's unlikely that the person in the voicemail was him. That should be enough to challenge the claim, but what happens if the evidence doesn't have metadata or the metadata has been altered somehow?

Anyone can claim any evidence is fake, so alleged deepfake evidence must meet other requirements. The person asserting the deepfake defence will have to show with some quantum of proof that the evidence is indeed a deepfake. This may be extraordinarily challenging given how convincing deepfakes can be. While this issue may seem far-fetched, individuals implicated in the January 6 U.S. Capitol attack claimed that the person in the social media or surveillance videos used against them were not actually them. As people become more aware of how easy it is to manipulate and generate audio and visual evidence, defendants will use that skepticism to their benefit.

There's a U.S. evidentiary rule, known as Federal Rule of Evidence 403, that provides general guidance to lawyers and trial judges that might be useful in the alleged deepfake context, but evidence rules are not designed and promulgated for any particular type of technical evidence. The rule has been used cautiously and infrequently to deny the admissibility of evidence that might be unduly prejudicial, confusing, or likely to mislead.

My coauthors and I have argued that we should ease the strict requirements of that rule such that if a party can make a sufficiently credible showing to challenge certain evidence as deepfake, the judge should look at whether that evidence is the only evidence to prove the claim and whether admitting it might be outweighed by other factors. Is there other corroborating evidence? Did six people also hear my supervisor say what was in the voicemail or has that supervisor written similar things in other emails that were sent to other employees? If so, that's a very different set of facts than if the deepfake evidence is the only evidence the court has available to it and we have competing facts about whether or not it is authentic.

When evaluating the admissibility of evidence that is potentially deepfake, judges need to avoid the unfair prejudice that can occur if the fake evidence is allowed to be presented to the jury. We know that audio and video evidence strongly influences what people perceive and remember. If you show jurors an audiovisual clip as compared to having them read a witness transcript, they are 650 percent more likely to remember the content in the audiovisual clip. That's just how the human mind works. The court will need to consider the potential risk, negative impact, or problematic consequences that could occur if the evidence turns out to be fake. Of course, the implications could be anything from losing a small sum of money to losing custody of children or one's freedom.

Can lawyers and litigants use ChatGPT to prepare court filings?

Yes, and in the past year both lawyers and self-represented persons, what are known as *pro se* litigants, have used ChatGPT to prepare filings for court. But one big problem is the propensity of GenAI to hallucinate. Briefs have been drafted with citations that sound authoritative but are not real and refer to cases that don't actually exist. In response, we have a frightened judiciary that has reacted by requiring disclosure of the use of AI and certification that the filer has verified the cited sources.

On the positive side, ChatGPT can allow litigants without sufficient monetary resources to hire lawyers to file cases in court, so GenAI can increase access to justice. People who simply can't afford a lawyer or cannot access one for other reasons can now use GenAI to generate customized legal papers specific to their circumstances and jurisdictions.

But on the negative side, someone who is vexatious can now use ChatGPT to file hundreds of complaints against multiple people and flood the courts. Malicious filers also can prepare simultaneous filings in courts around the country, flooding the courts with duplicate, frivolous submissions.

Can judges and their staff use GenAI for research or to draft opinions?
We know of at least three judges who have used GenAI to draft opinions. You might think, what's the problem since GPT-4 has passed the U.S. bar exam? But the concern is that ChatGPT can provide different answers to the same question at different times, not to mention hallucinate completely false information.

We have a justice system in the first place to guarantee people their day in court before a living, breathing, human judge. Using a GenAI system for independent research without informing counsel or the parties and providing them with an opportunity to object or respond to arguments that are not in the record may expose the court to sources of information that have not been put in evidence, or that raise other due process issues. So for now, while the tools can be a useful adjunct, as far as I am aware, the courts in the U.S. and Canada are refraining from using GenAI for decision-making and drafting purposes.

Further reading

Paul W. Grimm, Maura R. Grossman, Gordon V. Cormack. Artificial Intelligence as Evidence, 19 *Northwestern Journal of Technology and Intellectual Property* 9–106 (2021)

Cynthia Cwik, Paul W. Grimm, Maura R. Grossman, Toby Walsh. *Artificial Intelligence, Trustworthiness, and Litigation, Artificial Intelligence and the Courts: Materials for Judges* 1–35. American Association for the Advancement of Science (2022).

Paul W. Grimm, Maura R. Grossman. Artificial Intelligence Comes to Court. *Federal Magistrate Judges Association ("FMJA") Quarterly Bulletin* 4–5, 26 (June 2022).

Paul W. Grimm, Maura R. Grossman, Sabine Gless, Mireille Hildebrandt. Artificial Justice: The Quandary of AI in the Courtroom. *Judicature International* 1–13 (Sept. 2022).

Maura R. Grossman, Paul W. Grimm, Daniel G. Brown. Is Disclosure and Certification of the Use of Generative AI Really Necessary?, 107:2 *Judicature* 68–77 (Oct. 2023).

Maura R. Grossman, Paul W. Grimm, Daniel G. Brown, Molly (Yiming) Xu. The GPTJudge: Justice in a Generative AI World, 23:1 *Duke Law & Technical Review*. 1–34 (Dec. 2023).

Maura R. Grossman, Paul W. Grimm, Cary Colganese. Point/Counterpoint on AI in the Courts: How Worried Should We Be?, 107:3 *Judicature* 41–48 (Mar. 2024).