ARIZONA SUMMIT ON

ARTIFICIAL INTELLIGENCE

LAW AND THE COURTS

TOPIC: ARTIFICIAL INTELLIGENCE AND DEEPFAKES; PAPER 1

Authored by:

Mark Lanterman

Deepfakes and the Impact of AI on the Courtroom

Mark Lanterman

I. INTRODUCTION

In 2014, Professor Stephen Hawking expressed his weariness over advancements in artificial intelligence, telling the BBC, "The development of full artificial intelligence could spell the end of the human race". Though Hawking himself had a very personal relationship with AI, even making communication possible throughout his battle with ALS, he still feared, "the consequences of creating something that can match or surpass humans."[1] In recent years, AI has only continued to shape how human beings work, learn, and live. The benefits are numerous—from advanced medical technologies to the conveniences afforded by tools such as ChatGPT—and the possibility for new applications seems limitless. The potential is exciting, but equally concerning. Almost ten years later, Stephen Hawking's concerns have resurfaced for many.

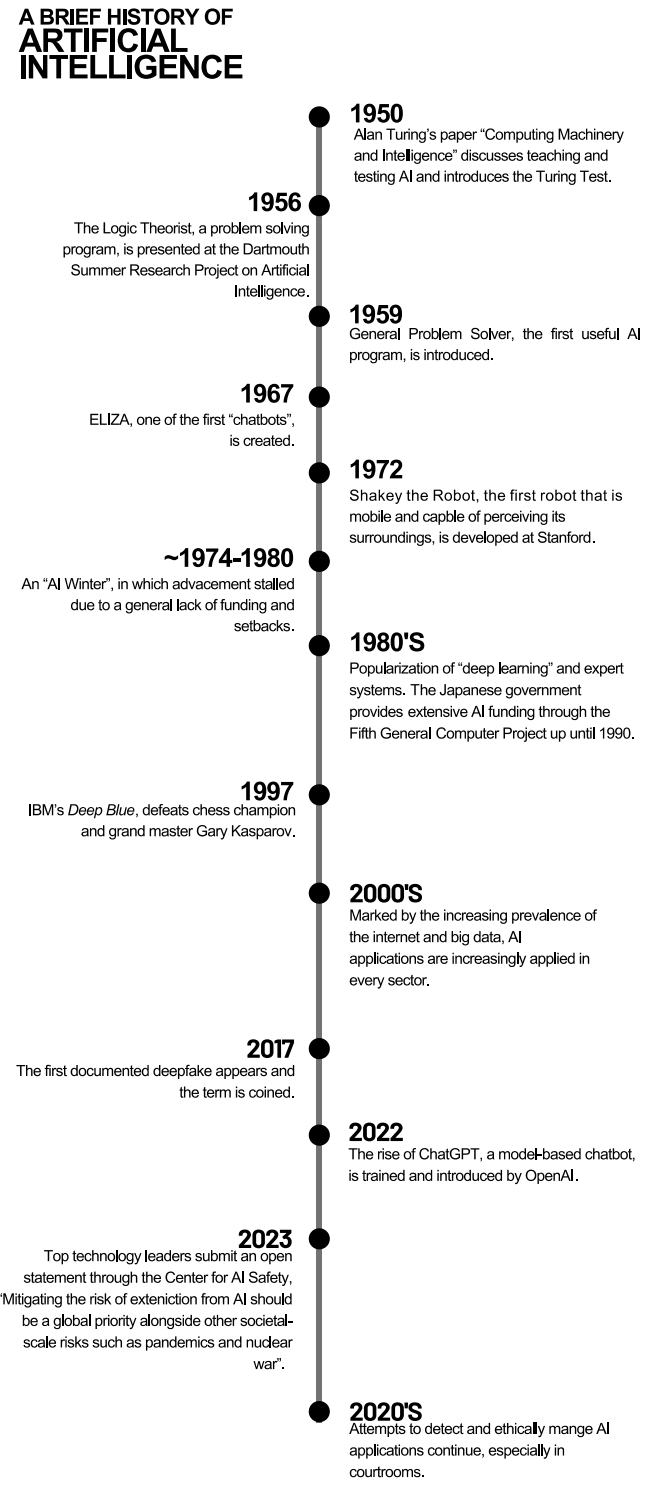
On October 30, 2023, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence was put forth by the Biden Administration.[2] The executive order acknowledges both the risks and manifold benefits of AI technology, as well as the need for establishing governance in managing these technologies as responsibly as possible. It states:

> *Artificial Intelligence must be safe and secure. Meeting this goal requires robust, reliable, repeatable, and standardized evaluation of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use.… Testing and evaluations, including post-deployment performance monitoring, will help ensure that AI systems function as intended, are resilient against misuse or dangerous modifications, are ethically developed and operated in a secure manner, and are compliant with applicable Federal laws and policies.* ***Finally, my Administration will help develop effective labeling and content provenance mechanisms, so that Americans are able to determine when content is generated using AI and when it is not.***

Courts are being called upon to address AI in multiple forms; from developing standards and policies for using generative AI tools such as ChatGPT in writing court documents to identifying a potential deepfake submitted into evidence. While the executive order of October 2023 puts forth a goal of enabling Americans to be able to immediately "spot" a product of AI, technologies that would allow for this instant identification with complete accuracy are not

[1] https://www.bbc.com/news/technology-30290540

[2] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/



currently available. Though this objective may be achieved at some point in the future, it is important that courts prepare themselves for addressing current issues involving AI. Depending on what technologies are developed and implemented in the future, such as watermarking or labeling systems, it will still be important to have protocols in place for instances in which the veracity of digital evidence remains contested.

In particular, courts need reliable methods to manage deepfake technology, especially as it pertains to detection and in addressing the "deepfake defense". This paper will provide a brief history of advancements in artificial intelligence and deepfake technology, an overview of some of the issues that these technologies present in court, and a proposal for how to best address deepfakes given current technological limitations.

II. BACKGROUND

Well before Stephen Hawking's comments, A.M. Turing's 1950 paper, "Computing Machinery and Intelligence" discussed approaches to teaching and testing machines, though resources and knowledge at that time were not sufficient to begin pursuing AI in earnest. As computing technologies developed, and were less expensive to utilize, so too did advancement in artificial intelligence. Marked by numerous setbacks and the need for computing systems to evolve first, the journey to deepfake technologies and applications such as ChatGPT has been a long one. From

the science fiction fantasies of the early 20th century to today, artificial intelligence has taken up a notable position in modern consciousness. Though once primarily restricted to the academic community, many AI applications are now commonly available.

ChatGPT, a chatbot developed by OpenAI, is one such example. Released in November of 2022, ChatGPT quickly became a popular topic in almost every sector. Once released, ChatGPT was lauded for its potential benefits and uses, but ethical questions about its development and concerns about safety and security soon steered the conversation. OpenAI explains in its blog, "We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests."[3] From being temporarily banned in Italy to Sam Altman himself, the CEO of OpenAI, admitting to being "a little bit scared of AI",[4] ChatGPT has continued to make international headlines. Within the legal community, problems soon materialized when it came to using ChatGPT in an acceptable way. Many within the legal community are still looking for guidance when it comes to strategically implementing ChatGPT while minimizing the risks. Policies for guiding appropriate use (and when human intervention is necessary to review AI-produced materials) are especially necessary for lawyers tasked with the responsibility of safeguarding their clients' information.

A New York lawyer used ChatGPT to create a legal brief, which was discovered after cited cases were shown to be fabricated.[5] He explained that he had been unaware that ChatGPT could create false information, and expressed remorse for not verifying that the content it produced was accurate. This incident demonstrated the need to create standardized practices for ChatGPT, and AI more generally, when used for legal purposes. It also showed that in spite of ChatGPT's impressive ability to create believable content instantly, human oversight is still needed to ensure its accuracy. Following this incident, U.S. District Judge Brantley Starr of the Northern District of Texas implemented a policy requiring attorneys to "file a certificate to indicate either that no portion of any document they file was generated by an AI tool like ChatGPT, or that a human being has checked any AI-generated text."[6] However, some judges may find this kind of measure to be unwarranted, believing that current standards and ethical responsibilities are sufficient in guiding an attorney's use of AI. In an open letter drafted with the assistance of ChatGPT, Judge Scott U. Schlegel stated his opinion that, "an order specifically prohibiting the use of generative AI or requiring a disclosure of its use is unnecessary, duplicative, and may lead to unintended consequences". Furthermore, he stated that, "Generative AI, much like any tool, is only as effective as the legal expertise guiding it."[7]

In addition to ChatGPT, other types of AI have found their way into the courtroom. While practices are having to be developed to guide how applications such as ChatGPT are used

[3] https://openai.com/blog/chatgpt
[4] https://www.cnbc.com/2023/03/20/openai-ceo-sam-altman-says-hes-a-little-bit-scared-of-ai.html
[5] https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html
[6] https://www.cbsnews.com/news/texas-judge-bans-chatgpt-court-filing/
[7] https://www.judgeschlegel.com/blog/-a-call-for-education-over-regulation-an-open-letter

within the legal profession, the court is being called upon to recognize instances in which AI is being used by litigants to create fake evidence, or as an excuse to weaken real evidence.

III. THE DEEPFAKE

According to the Department of Homeland Security's paper, "Increasing Threat of Deepfake Identities", "Deepfakes, an emergent type of threat falling under the greater and more pervasive umbrella of synthetic media, utilize a form of artificial intelligence/machine learning (AI/ML) to create believable, realistic videos, pictures, audio, and text of events which never happened. Many applications of synthetic media represent innocent forms of entertainment, but others carry risk."[8] Deepfakes are created using readily available deepfake technology; they are completely manufactured and do not incorporate existing media. Though sometimes made for the purposes of entertainment, they are also frequently used as a method for spreading misinformation.

In addition to deepfakes, shallow fakes can be similarly deceiving. Though the terms are often conflated, shallow fakes use basic editing techniques and software tools to alter existing media, for example by slowing down parts of a video or selective splicing. With one small edit, an entire video can be altered to give a drastically different perspective than its original. This type of modified digital content may be simpler to create than a deepfake, thus making them more common. However, since they are made from an existing source, they may be less challenging to identify. Deepfakes remain difficult to distinguish from authentic content, even for experts. As they are entirely generated using AI technology, several different measures may be needed to make a determination as to whether a piece of evidence is a deepfake.

In one UK case, a shallow fake almost had a critical impact on a child custody case. "A woman said her husband was dangerous and that she had the recording to prove it. Except, it turned out she didn't. The husband's lawyer revealed that the woman, using widely available software and online tutorials, had doctored the audio to make it sound like his client, a Dubai resident, was making threats. . . [and] by studying the metadata on the recording, his experts revealed that the mother had manipulated it."[9] In this instance, a third-party expert was required to analyze the evidence in question and provide insight into its origin. Though the evidence in this situation was shown to be a shallow fake, it is likely that harder-to-identify deepfakes will only continue to proliferate and complicate proceedings.

Still, at the time of writing, many believe that the risk of deepfakes being submitted into evidence is a less pressing threat than that of its reversal—the deepfake defense. Capitalizing on the uncertainty and mistrust characterizing the "misinformation age", a new tactic has arisen among litigants when presented with strong evidence. "That's not me; it's fake. Prove it's not." Though it may seem a weak defense at face value, it can deplete resources, fatigue juries,

[8] https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf
[9] https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes

and generally prevent a case from moving forward. Depending on the circumstances, it may be difficult to make up for lost time and restore confidence in the evidence as presented.

As both situations continue to play out in the courtroom, courts should be well-equipped to address them. Though judges may not necessarily be directly responsible for identifying deepfakes or altered media, it is important that judges use available measures to gather contextual information and uphold admissibility standards for digital evidence in making authenticity determinations.

IV. THE PROBLEM

Deepfakes are easily generated, easily shared, and can easily fool even the most trained eye. The term deepfake was coined in 2017 after the appearance of what is commonly accepted as the first deepfake; since then, they have become a hallmark of current trends in AI. It should be noted that the best and most convincing deepfakes may require more advanced equipment, processing abilities, and training; however, producing a deepfake is now easier than ever as new tools are introduced to the market. Voice deepfakes (or vocal cloning) can also be eerily convincing. Using AI technology, individuals' voices can be replicated and used to make new recordings.

Deepfakes can pose a two-fold problem in the courtroom. Either deepfakes are admitted as evidence having been maliciously produced by litigants or the deepfake defense will be thrown out indiscriminately to weaken legitimate evidence.

Some believe that the deepfake defense was made in a case involving Tesla and a wrongful death lawsuit.[10] In 2018, Walter Huang died in a car accident while driving a Tesla vehicle. According to the complaint, the vehicle's Autopilot feature did not function properly, leading to Mr. Huang's fatal car accident. His family contends that Tesla misrepresented the risks of the Autopilot feature technology; a statement made by one of the family's attorneys even states that Tesla is guilty of "beta testing its Autopilot software on live drivers."[11]

Huang's family points to a 2016 video of Elon Musk as proof that Tesla and Elon Musk himself have historically overstated the safety of their vehicles. In one video, Elon Musk can be seen stating during a technology conference, "A Model S and Model X at this point can drive autonomously with greater safety than a person. Right now."[12] In response, Musk's legal team stated that not only does Mr. Musk not remember making that specific claim, but that the video itself could be fake. Simply, given Mr. Musk's fame and notoriety, it is possible that the video may be a deepfake.

---

[10] Sz Hua Huang et al v. Tesla, Inc., The State of California, no. 19CV346663

[11] https://www.forbes.com/sites/alanohnsman/2019/05/01/tesla-sued-by-family-of-silicon-valley-driver-killed-in-model-x-autopilot-crash/?sh=63f0dbfe1c3f

[12] https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused

V. STRATEGIES FOR MANAGING DEEPFAKES

Judge Evette Pennypacker responded, "Their position is that because Mr. Musk is famous and might be more of a target for deep fakes, his public statements are immune".[13] Furthermore, "In other words, Mr. Musk, and others in his position, can simply say whatever they like in the public domain, then hide behind the potential for their recorded statements being a deep fake to avoid taking ownership of what they did actually say and do".[14] In light of this claim, the court had to decide how to proceed:

Confronted with Tesla's refusal to rule out that some clips could be digitally altered deep fakes and therefore not suitable as evidence, the judge came up with an elegant solution: Put the billionaire entrepreneur and artificial intelligence enthusiast under oath and have him testify as to which statements coming out of his mouth are authentic.[15]

To gather contextual information, Judge Pennypacker allowed for an apex deposition[16] of Mr. Musk in order to establish whether or not he had a) attended the functions as portrayed in the footage and b) made the statements in question. This measure was ultimately deemed necessary to determine the authenticity of the recording, likely an unintended consequence of the defense.

On this occasion, the deepfake defense resulted in a need for additional testimony to assist in establishing the veracity of digital evidence presented. Following the court's response, one lawyer representing Tesla stated that the intention was not to claim any videos were deepfakes, but "we raised this idea, this issue, because we're living in a world today where these things exist"[17]. And this is, more or less, the unfortunate heart of the issue. Namely, that the emergence of the deepfake has opened the door to the claim that any piece of evidence, could, in theory, be fake. This court's response illustrates the fact that when dealing with new technologies, the old rules can still apply. Gathering contextual information using available means (i.e. apex depositions) and going to the source are critical steps in minimizing any negative ramifications of the deepfake defense.

When it comes to determining the role and responsibilities of the court in verifying digital evidence, some stress that a judge is only responsible for following the rules of evidence. Judges are not expected to be experts in every issue that may appear before them, which has also been true in matters involving digital evidence. However, as is always the case, judges are called upon to make credibility determinations based on testimony and the facts of a case.

---

[13] https://www.reuters.com/legal/elon-or-deepfake-musk-must-face-questions-autopilot-statements-2023-04-26/
[14] https://news.bloomberglaw.com/esg/musk-likely-must-give-deposition-in-fatal-autopilot-crash-suit
[15] https://fortune.com/2023/04/27/elon-musk-lawyers-argue-recordings-of-him-touting-tesla-autopilot-safety-could-be-deepfakes/
[16] https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-seeing-is-not-believing-authenticating-deepfakes
[17] https://fortune.com/2023/04/27/elon-musk-lawyers-argue-recordings-of-him-touting-tesla-autopilot-safety-could-be-deepfakes/

In Rebecca A. Delfino's paper, "Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery", she submits, "[This article] is the first to propose a new addition to the Federal Rules of Evidence reflecting a novel reallocation of fact-determining responsibilities from the jury to the judge, treating the question of deepfake authenticity as one for the court to decide as an expanded gatekeeping function under the Rules. The challenges of deepfakes—problems of proof, the "deepfake defense," and juror skepticism—can be best addressed by amending the Rules for authenticating digital audiovisual evidence, instructing the jury on its use of that evidence, and limiting counsel's efforts to exploit the existence of deepfakes".[18]

Basic guidelines can help in gathering necessary contextual information.

1. The best defense is proactively upholding authentication standards and the rules of evidence, especially when handling digital media. These measures will best allow for the preservation of original source material, which can be analyzed by third experts should the need arise. When a claim of fake evidence is made, judges can look to how well digital evidence has been managed by both sides as one metric for assessing the likelihood of whether a claim is being made in good faith.

2. Context is key. Additional witness testimony may be required to investigate deepfake claims. Asking specific questions about the evidence at hand, as well as ascertaining how that evidence has been collected, can shape the court's next steps.

3. Third-party forensic experts can be valuable in providing information about a piece of evidence, indicating a probability of its authenticity. A special master appointed by the court can investigate how digital evidence has been handled throughout a case and determine whether best practices have been upheld in the collection, preservation, and analysis of digital evidence. An expert may be able to provide a digital narrative of the evidence in question which may include analyzing original source materials and reporting on any signs of tampering, alteration, or corroborating findings that support claims of inauthenticity. However, it should be noted that is not currently possible to instantly identify a deepfake, or any type of "fake" digital evidence. Can an expert definitively state whether something has been "faked"? Not necessarily. Deepfakes are especially problematic as even technological experts may have difficulty in spotting them. In spite of these challenges, an expert's assessment may be able to supply the court with an additional viewpoint to help inform its own assessment.
4. Expert analysis of digital evidence as well as the gathering of contextual information through deposition and cross-examination can enable the court in its determination (and the assigning of sanctions, if necessary).

[18] https://repository.uclawsf.edu/cgi/viewcontent.cgi?article=4012&context=hastings_law_journal

While costs are a concern when considering utilizing an expert witness, this measure may minimize long term costs incurred due to false claims or the submittal of fake evidence. The burden of the expense can be assigned at the discretion of the judge, perhaps depending on the results of the expert's opinion. Another benefit is that the potential expense may, in fact, deter individuals from making false claims or submitting fake evidence.
Tools designed to detect deepfake technology, though currently at varying degrees of progress and usability, will likely mirror AI in their evolution and development. According to an October 2023 MIT Technology Review article written in response to the goals stated in the Executive Order on AI, "The trouble is that technologies such as watermarks are still very much works in progress. There currently are no fully reliable ways to label text or investigate whether a piece of content was machine generated. AI detection tools are still easy to fool".[19] Part of the evolution of AI is its pursuit of evading detection. At this stage, courts should likely primarily rely on the existing frameworks and systems in place, combined with additional measures to establish context as required.

VI. IN CONCLUSION

Fake evidence is nothing new—but juries existing within a world of "fake news" and readily available, AI technology, is. Courts have to be enabled to manage the new challenges brought about by AI, in the various forms it may appear; from establishing protocols for how materials produced by ChatGPT must be reviewed by counsel, to creating a course of action to manage instances of the deepfake defense. The bad news is that deep fake technology creates undeniable hurdles; the good news is that many of the same protections that existed before for similar issues still apply. And, when in doubt, every tool available should be used to establish context. This may include involving an objective, third party to provide a reliable digital narrative. Though a number of different information-gathering measures may be needed, movement towards improved detection technologies will continue to shape how courts can most efficiently respond.

The legal community should be mindful of the possibility of altered or fake evidence being presented by their clients. Lawyers are never permitted to present evidence that they know for a fact to be false; however, evolving technologies may render more stringent standards necessary.

Ten years ago, Stephen Hawking had clear reservations about the trajectory of artificial intelligence. Nine years later, a statement titled, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war", was signed by hundreds of AI, security, and technology leaders.[20] For some, the risk seems overstated. For others, the warning feels appropriate given current problems. Quietly

[19] https://www.technologyreview.com/2023/10/30/1082678/three-things-to-know-about-the-white-houses-executive-order-on-ai/
[20] https://www.safe.ai/statement-on-ai-risk

progressing, 2023 seemed to be the year when many began to share a common sentiment with Hawking and others throughout the years who have expressed their concerns.

Even OpenAI founder, Sam Altman, urged increased regulation and oversight at a Senate subcommittee hearing in May of 2023.[21] As the October 2023 Executive Order explains:

> *Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.*

Society is undoubtedly having to grapple with balancing the numerous benefits of these technologies with their significant risks. In the courtroom, existing evidentiary rules can form the basis of how deepfakes, and the deepfake defense, are addressed. Calling for additional testimony, and the input of expert witnesses, are measures that can allow the court to gather contextual information in determining the admissibility of evidence.

[21] https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html