Institute for Youth in Policy

5214F Diamond Heights Blvd #3055
San Francisco, CA 94109
+1 (715) 469-6884
fellowship@yipinstitute.org
www.yipinstitute.org/fellowship/

Fellowship Capstone | Policy Brief

Fact or Fabrication: AI-Generated Evidence Within Courtrooms

Nikki Wu

I. Executive summary

Artificial Intelligence (AI) advancements are accelerating and it has opened a variety of possibilities within the justice system. However, it has also opened the doors for dangerous vulnerabilities: specifically, the usage of AI-generated evidence. As AI continues advancing, the risk of fabricated evidence entering the courtroom continues to increase as well. Although there are current safeguards in place such as detection tools and authentication technology, these solutions are largely insufficient and leave room for substantial threats to justice as AI continues to rapidly develop. This brief examines and discusses how AI is being used to create false or misleading evidence and the dangers it creates to due process. To protect due process, this brief also examines and proposes new policies, and standards to combat this growing concern.

II. Overview

The integrity of courtroom evidence has been crucial to upholding justice and due process. Without sufficient policy to ensure all evidence that enters the courtroom is proper, we risk eroding the justice system's capability. With this in mind, artificial intelligence has created challenges that the current legal framework is not able to address. With advancements in AI-generated content such as deepfake videos, fabricated audio recordings, manipulated images, and more, the risk of false evidence being entered into courtrooms has become a pressing concern. These fabricated materials have become difficult to detect as advancements within AI are occurring quickly. Thus, making it easier to mislead and influence juries and judges. This growing concern not only threatens and endangers the legitimacy of legal proceedings, but also the foundational rights guaranteed by the Constitution. As AI technologies continue to evolve, the justice system must also adapt accordingly. This policy brief aims to propose solutions to prevent AI-generated content to be used as evidence to mislead judges and juries in our nation.

A. Relevance

AI-generated evidence has been seen to influence legal proceedings; it's becoming a reality that is demanding immediate attention and proactive measures. In a UK child custody case that occurred in 2020, a deepfake audio recording of the father was submitted. The deepfake audio recording portrayed the father speaking a threatening message and was submitted in an attempt to discredit him and create a negative persona of the father. Although the case name and parties have not been publicly disclosed, multiple digital forensic experts confirmed that the recording was indeed synthetic by the source CYFOR. This child custody case serves and

demonstrates how easily accessible and dangerous AI-generated content can be to mislead judges and juries. This potential could well likely compromise the fairness of judicial outcomes. Furthermore, detecting deepfakes and AI-generated content remains a significant challenge.

A study titled, "Warning: Humans cannot reliably detect speech deepfakes," conducted by researchers at the University College London (UCL) and published in PLOS ONE in August 2023, explains critical and ongoing barriers to this challenge at hand. The study included 529 participants who were presented with real and deepfake audio samples in English and Mandarin. The results revealed a pressing concern: deepfake audios were correctly identified only 73% of the time; humans are unable to recognize and distinguish what is AI-generated 100% of the time. This highlights the fact that awareness alone will not be sufficient enough to identify AI-generated audio and it creates a huge risk if additional safeguards are not implemented.

Another recent research study that highlights the danger of AI-generated evidence entering courtrooms is, *The Deepfake-Eval-2024* study. The research study tested the performance of state-of-the-art detection models on real world datasets. The results have raised numerous alarms, the detection accuracy had dropped substantially by nearly 50% for video and audio, compared to previous prior tests. These results further illustrate the potential danger and need for a more realistic evaluation of deepfake detection models.

The combination of AI's advancements and lack of accurate and reliable detection tools pose an unprecedented threat to the United State's legal system. With synthetic audios becoming increasingly realistic as well as just how easy it is to access this tool, the potential to weaponize AI in the courtroom is beginning to grow rapidly. Action within the nation must occur now to combat this growing concern and to protect the integrity of our legal system.

III. History

A. Current Stances

Historically, the justice system places emphasis on the credibility and authenticity of evidence. Physical evidence and sworn testimony were treated as the gold standard in upholding due process and protecting the rights of the defendants. These forms of evidence act as protection against false or misleading information. Additionally, these forms of evidence could be authenticated through legal procedures and experts. Courtrooms are built on integrity, with strict evidentiary rules in place to ensure that only verified and reliable information could influence a legal proceeding's outcome. Without these standards, the courtroom would risk injustice and destabilization.

Throughout the years, courts have been adapting to changing technologies, from the inclusion of surveillance footage to accepting digital records like emails and text messages. However, the advancements of generative AI have introduced growing concern. Today, AI generated content can convincingly replicate real people, voices, events, and more. These fabrications are difficult to detect and existing evidentiary standards are not equipped to properly evaluate their authenticity allowing fabricated content to seep into court rooms and influence an outcome of a trial. This issue is further confounded by the speed at which AI is developing; if sufficient safeguards are not implemented swiftly, AI will

have already done irreparable damage to the justice system.

Legal and civil organizations have begun raising alarms. Organizations such as the Electronic Frontier Foundation and the ACLU are continuing to warn that without oversight and disclosure requirements, AI-generated evidence could compromise courtroom integrity and undermine constitutional protections.

IV. Policy Problem

A. Stakeholders

The primary stakeholders are defendants, attorneys, and judges within the justice system. Defendants of all kinds face the risk of wrongful conviction or harsher sentencing due to AI-fabricated evidence. Defense attorneys and prosecutors may struggle with ensuring authenticity, while judges must take into account increasingly complex digital evidence.

In addition, forensic analysts, artificial intelligence developers, and legal technology companies are critical stakeholders. These groups are directly involved in either detecting fabricated content or developing tools that can be used in legal contexts. Artificial intelligence developers will also be potentially affected by policy that seeks to regulate AI and its capabilities in generating content that may be indistinguishable to reality. Legislators and government institutions are also responsible for updating legal frameworks to ensure due process is protected in the face of emerging technologies.

B. Risks of Indifference

Failure to address the threat of AI-generated evidence in courtrooms could seriously damage the integrity and stability of the legal system. Defendants may be wrongly convicted or improperly sentenced based on deepfakes or altered audio recordings, leading to breakdowns of justice. Defendants also risk having their character publicly discredited due to AI-generated evidence presented as real. Public trust in the justice system would decline if AI manipulation is not efficiently addressed.

Moreover, without clear regulations, legal professionals and forensic experts may be unequipped to detect and respond to AI-generated materials. This could result in mounting appeals, mistrials, and a backlog in court proceedings. The emotional manipulation of juries and judges through false imagery or voice recordings further complicates judicial neutrality, potentially violating constitutional protections such as the Sixth Amendment right to a fair trial.

C. Nonpartisan Reasoning

AI-generated evidence presents a threat that goes past party lines, affecting civil liberties and judicial fairness. Ensuring that all legal evidence, no matter how it's created, can be verified protects all defendants, whether rich or poor and regardless of race, gender, or political affiliation.

From a budgetary standpoint, proactive regulation and AI-detection protocols can reduce the financial burden of wrongful convictions, retrials, and legal appeals. The financial risk that AI-generated evidence in the courtroom presents far outweighs the cost of implementation of protocol and policy. Ensuring the integrity of courtrooms is not just a moral obligation, it is a necessity for upholding democratic values and due process.

V. Tried Policy

Courts have traditionally relied on Federal Rules of Evidence and expert testimony to validate digital evidence, but these systems were designed before the rise of generative AI. In recent years, some states and agencies have begun exploring policy responses. For example, the National Institute of Standards and Technology has launched initiatives to develop guidelines for detecting deepfakes and AI-manipulated content.

In California, AB 730 was enacted in 2019 to criminalize the use of deceptive deepfakes intended to influence elections. However, no comprehensive federal legislation exists regarding the use of AI-generated evidence in courtrooms. Policies like the Deepfake Task Force Act, introduced in 2023, aim to study the effects of deepfakes across sectors, including law, but have yet to directly implement any policy within the judicial system.

These efforts show growing awareness of the risks that AI poses to the courtroom, but emphasize the gap in proactive and enforceable standards that ensure courtroom evidence remains credible and just. If additional policy and protocol is not implemented, we risk destabilizing the courtrooms.

VI. Policy Options

Passing a Deepfake Defense Act

A major policy option is the creation of a deepfake defense act. Establishing a federal subcommittee composed of digital forensic experts to verify the authenticity of digital evidence and ensure it has not been generated or altered using AI. This federal subcommittee would operate directly under the Department of Justice (DOJ). Their main task would include: developing standardized forensic procedures and updating and maintaining a federal certified AI detection tool by partnering with research experts. Establishing a federal subcommittee would ensure a more advanced and proactive approach to evaluating AI-generated evidence across courts. This policy addresses the current gaps in the nation's legal standards.

Judicial and Legal Professional Training Programs

Another critical approach is to implement comprehensive training for judges and attorneys. that focuses on detecting and teaching the most recent up to date ways to identify AI-generated content. These programs offered would also provide in-depth analysis of the latest advances in AI generated media and or audio content. Furthermore, participants will learn about limitations of AI-detection tools and potential risks of relying on forensic reports, and methods to integrate technical findings into legal proceedings. Enhancing technical literary and awareness among legal professionals, the program will aim to strengthen critically evaluating digital content entering legal proceedings. Implementing a program mitigates the influence of misleading or AI-generated content and reduces the risk of wrongful convictions.

Federal Investment in AI Detection Research

Another essential way to reduce AI-generated content entering courtrooms is allocating funding to support research and development of advanced AI-detection technologies that can reliably detect deepfakes, synthetic audios, and more. This initiative would partner and collaborate with federal labs, universities, private tech companies, and digital forensic experts, to establish and develop a state-of-the-art detection tool. This

investment would also mean keeping up with advancements of AI-generation. Fundings for this initiative could be from the Department of Justice's research and technology budget and grants from federal science agencies such as the National Science Foundation or the Department of Homeland Security's Science & Technology Directorate. Investing in innovation would lead to the legal system gaining access to more accurate and reliable tools for verifying digital evidence.

VII. Conclusions

The advancement of generative AI into society brings great innovation, but also great risk when used maliciously. Courtrooms must maintain the highest standards of evidence to protect due process, fairness, and trust in the judicial system.

AI-generated content like deepfakes and fake audio recordings present a unique challenge that existing legal frameworks are not yet equipped to handle. Policies requiring disclosure, developing AI detection capabilities, and reviewing evidence through expert panels are essential next steps.

It is imperative that policy is made before AI-generated misinformation leads to irreversible judicial errors. In the words of Benjamin Franklin, "Justice will not be served until those who are unaffected are as outraged as those who are."

Acknowledgment

References

[1] Chesney, R., & Citron, D. (2019). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. *California Law Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954

[2] National Institute of Standards and Technology. (2023). *Guardians of Forensic Evidence: Evaluating Analytic Systems Against AI-Generated Deepfakes*. https://www.nist.gov/publications/guardians-forensic-evidence-evaluating-analytic-systems-against-ai-generated-deepfakes

[3] Hong, D. (2024). *Deepfakes and the threat to due process: Legal challenges in the era of synthetic media*. Cogent Social Sciences, 10(1), Article 2320971.

[4] California Legislative Information. (2019). *AB-730: Elections: deceptive audio or visual media*. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

[5] Congressional Research Service. (2025, June 4). *Regulating artificial intelligence: U.S. and international approaches and considerations for Congress* (CRS Report No. R48555). https://crsreports.congress.gov/product/pdf/R/R48555

[6] CYFOR. (2020). Deepfake Audio Analysis: UK Child Custody Case. CYFOR Digital Forensics Report. https://www.cyfor.co.uk/digital-forensics

[7] University College London. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0288499

[8] University of California, Berkeley. (2025). *Deepfake-Eval-2024: Benchmarking deepfake detection in real-world conditions*. https://www.researchgate.net/publication/389

581656_Deepfake-Eval-2024_A_Multi-Modal_In-the-Wild_Benchmark_of_Deepfakes_Circulated_in_2024

[9] Electronic Frontier Foundation. (2023). AI and deepfakes: Protecting free speech while preventing harm. https://www.eff.org/deeplinks/2023/ai-and-deepfakes

[10] American Civil Liberties Union. (2024). *Free speech and the rise of AI-generated content*. https://www.aclu.org/