THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME 26 STLR.ORG NUMBER 2

ARTICLE

JUDICIAL APPROACHES TO ACKNOWLEDGED AND UNACKNOWLEDGED AI-GENERATED EVIDENCE

Maura R. Grossman* & Hon. Paul W. Grimm (ret.)†

Between 2014 and 2024, rapid advancements in computer science ushered in a dramatic new form of technology—Generative AI ("GenAI"). It offered seemingly limitless possibilities for creative applications never before imagined. But it also brought with it a darker side—the ability to create synthetic or "fake" text, images, audio, and audiovisual depictions so realistic that it has become nearly impossible—even for computer scientists—to tell authentic from fake content. Along with this new technology, new terms have been introduced, including "hallucinations" and "deepfakes." The use of GenAI technology has not been limited to computer scientists and IT professionals. It is readily available on the Internet at little or no cost to anyone with a computer and Internet access. It is no exaggeration to say that GenAI has democratized fraud, and that an ever-increasing amount of content on the Internet is now synthetic or AI-generated. Deepfakes have been used for satire and amusement but also to humiliate and destroy the reputations and careers of persons depicted in the fakes, to spread

* Maura R. Grossman, J.D., Ph.D., is a Research Professor in the David R. Cheriton School of Computer Science at the University of Waterloo and an Adjunct Professor at Osgoode Hall Law School of York University, both in Ontario, Canada. She is also principal of Maura Grossman Law in Buffalo, N.Y.

† Hon. Paul W. Grimm (ret.) is the Director of the Bolch Judicial Institute and the David. F. Levi Professor of the Practice of Law at Duke Law School. Previously, he served as a District Judge (and before that as a Magistrate Judge) in the U.S. District Court for the District of Maryland. The authors wish to thank University of Waterloo undergraduate student, Sunny Hu, for her able research assistance and Gordon V. Cormack, Charles L.A. Clarke, and Amy Sellars for their helpful comments on a draft of this article. The views and opinions expressed herein are solely those of the authors and do not necessarily reflect the views or opinions of any institutions or clients with which the authors may be affiliated.

disinformation, to manipulate elections, and to mislead the public. They will most certainly find their way into the resolution of court cases where judges and juries will face real challenges understanding the operations and output of complex AI systems and distinguishing between what is real and what is not.

In this Article, we explore the development of GenAI and the deepfake phenomenon and examine their impact on the resolution of cases in courts. We address the ways in which both known-to-be-AI-generated evidence and suspected deepfake evidence may be offered during trials. We review the research literature regarding the ability of deepfakes to mislead and influence juries, and the challenges with detecting deepfakes that judges, lawyers, and juries composed of laypersons will face. We draw an important distinction between two kinds of AI evidence. The first is "acknowledged AI-generated evidence," about which there is no dispute that the evidence was created by, or is the product of, an AI system. The second is "unacknowledged AI-generated evidence," or potential deepfake evidence, where one party claims the evidence is an authentic representation of what actually happened, and the opposing party claims the evidence is a GenAI-fabricated deepfake. We discuss the application of existing rules of evidence that govern admissibility of evidence and how they might be flexibly applied—or slightly modified—to better address what is at issue with known AI-generated evidence. With respect to unacknowledged AI-generated evidence, we explain the challenges associated with using the existing rules of evidence to resolve the question of whether such evidence should be admitted, and the possible prejudice if it is allowed to be seen by the jury. We describe two proposed new rules of evidence that we have urged the Advisory Committee on Evidence Rules to consider regarding the evidentiary challenges presented by acknowledged and unacknowledged AI-generated evidence, and the actions proposed by the Committee to date. We finish with practical steps that judges and lawyers can take to be better prepared to face the challenges presented by this unique form of evidence.

I. INTRODUCTION: THE PATH FROM GENERATIVE AI TO DEEPFAKES

In 2013, the first author of this paper (Grossman) was a speaker at a bench and bar conference sponsored by the Tenth Circuit Court of Appeals. One of the justices of the U.S. Supreme Court attended the event as the justice assigned to that Circuit. During a cocktail reception, the author, using her cell phone—discreetly, so she thought—snapped a few photographs of the justice before being approached by the U.S. Marshals Service. Apparently, the justice in question did not appreciate being photographed holding an alcoholic beverage and the marshals requested that the author delete the photos she had taken in exchange for an opportunity to have her photograph taken with the justice, sans beverage. Obviously, the author dutifully complied with the marshals' request.

By the time a decade had passed, however, there was technology readily available to anyone with a computer and Internet access, that could not only create a highly realistic photo of the justice holding an alcoholic beverage, but also a video of that same justice appearing to be stumbling drunk at the same 2013 reception, and there was no U.S. marshal that could do anything to prevent that fake video from being disseminated.[1]

[1] *See* Ian Sample, *What are deepfakes - and how can you spot them?*, GUARDIAN (Jan. 13, 2020), https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them [https://perma.cc/5T3G-63SM] (describing how deepfake technology can easily depict a famous person doing or saying something that they never did).

How did we get there in less than 10 years? In retrospect, the answer is fairly straightforward. In 2014, deep-learning algorithms referred to as "Generative Adversarial Networks" ("GANs") were introduced by Ian Goodfellow and his colleagues.[2] GANs provided an adversarial training framework—using two competing algorithms—for generating synthetic content.[3] One algorithm—the "generator"—created new content, while a second—the "discriminator"—evaluated that content against "real" data in an effort to distinguish the two.[4] The discriminative network provided the generative network with iterative feedback, improving its ability to produce output that more closely mimicked real data.[5] This approach advanced the production of realistic images, audio, and video, laying the groundwork for modern generative AI.

In 2017, Ashish Vaswani et al. further introduced a groundbreaking architecture called "Transformers," detailed in their seminal paper *Attention is All You Need*.[6] Unlike prior recurrent neural networks, Transformers employed a "self-attention mechanism" that allowed them to dynamically capture long-range dependencies and relationships between input sequence elements.[7] Transformers gained widespread adoption due to their superior performance at tasks such as natural language processing ("NLP") and their efficiency in processing data in parallel, which also made them suitable for long sequences.[8] Around the same time, Diffusion Models emerged as a powerful alternative to GANs, using iterative refinement (i.e., "denoising") to generate high-fidelity, stable, diverse images.[9] Together, these three innovations—GANs and Diffusion Models (which excelled at high-quality image generation and modeling complex data distributions) and Transformers (which were proficient at handling sequential data, particularly for language modeling)—revolutionized generative AI, enabling sophisticated multimodal systems capable of creating text, images, and video with unprecedented realism and complexity.[10]

---

[2] Sunil Dangi, *The Evolution of Generative AI Models: From GANs to Transformers*, MEDIUM (Dec. 21, 2023), https://medium.com/@sunil.dangi/the-evolution-of-generative-ai-models-from-gans-to-transformers-853aafda017d [https://perma.cc/QL8C-N36H].

[3] *Id*.

[4] *Id*.

[5] *Id*.

[6] Ashish Vaswani et al., *Attention Is All You Need*, 2017 PROC. 31ST CONF. ON NEURAL INFO PROCESSING SYS. 6000; *see also* Amanatullah, *Transformer Architecture explained*, MEDIUM (Sept. 1, 2023), https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c [https://perma.cc/8VPR-QQEJ] (providing a simplified explanation of the transformer architecture model).

[7] Vaswani, *supra* note 6, at 6006-07.

[8] Capital One Tech, *Transformer Model in NLP: Your AI and ML Questions, Answered*, CAPITAL ONE (Nov. 8, 2023), https://www.capitalone.com/tech/machine-learning/transformer-nlp/ [https://perma.cc/3WBP-X78M].

[9] Ryan O'Connor, *Introduction to Diffusion Models for Machine Learning*, ASSEMBLYAI: SPEECH & TEXT (May 12, 2022), https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/ [https://perma.cc/SHG2-NVCC].

[10] *Id*.; *see also* Dangi, *supra* note 2.

In 2017, a Reddit user by the name of "deepfakes" began to post doctored pornographic clips mapping the faces of celebrities (e.g., Gal Godot, Taylor Swift, and Scarlett Johansson) onto the bodies of pornographic actors.[11] The initial iterations were crude, but with the assistance of new technologies and enthusiastic Internet users all over the world, the application quickly progressed, leading to a massive proliferation of the publication of non-consensual intimate images, frequently referred to as "revenge porn."[12] Deepfakes began as videos that were altered using open-source artificial intelligence ("AI") technologies to stitch together digital artifacts to delight or humiliate. Since their initial introduction, the term has expanded to include virtually any form of fake media, including photographs, voice clones, social media sites, product reviews, etc.[13] The terms "cheap fake" or "shallow fake" refer to fakes that are not created using AI, but rather, are the product of a human hand and so-called non-AI, low-tech editing capabilities.[14] A few years ago, we referred to this technique as "photoshopping."[15] Examples of cheap or shallow fakes include the 2020 videotape of Nancy Pelosi with slurred speech—intended to make the politician appear either demented or drunk[16]—and the 2024 Princess Kate debacle—intended to make the royal look better than she may have appeared at the time due to ongoing cancer treatment.[17]

---

[11] Gemma Askham, *Are Deepfakes the New Revenge Porn?*, BBC (Apr. 25, 2018), https://www.bbc.com/bbcthree/article/779c940c-c6c3-4d6b-9104-bef9459cc8bd [https://perma.cc/LA3X-YVCY]; *see* Gabe Regan, *A Brief History of Deepfakes*, REALITY DEFENDER (June 1, 2024), https://www.realitydefender.com/blog/history-of-deepfakes [https://perma.cc/S6ED-GJPF]; Meredith Somers, *Deepfakes, Explained*, IDEAS MADE TO MATTER (July 21, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained [https://perma.cc/77MH-3V5M].

[12] *See* Chance Carter, *An Update on the Legal Landscape of Revenge Porn*, NAT. ASS'N. OF ATT'YS. GEN. (Nov. 16, 2021), https://www.naag.org/attorney-general-journal/an-update-on-the-legal-landscape-of-revenge-porn/ [https://perma.cc/YFH2-H3NW].

[13] *See* Somers, *supra* note 11.

[14] Hyosun You, *What Are Cheapfakes (Shallowfakes)?*, SAMSUNG SDS (May 23, 2022), https://www.samsungsds.com/en/insights/what-are-cheapfakes.html [https://perma.cc/P5KF-ARMG]. The newest term added to the deepfake lexicon appears to be "softfakes," that are media modified to render a political figure in a more appealing light, typically to humanize them. MARIAGRAZIA SQUICCIARINI ET AL., UNESCO, SYNTHETIC CONTENT AND ITS IMPLICATIONS FOR AI POLICY: A PRIMER 29 & n.29 (2024), https://unesdoc.unesco.org/ark:/48223/pf0000392181 [https://perma.cc/WY7G-V4FB]. Examples include the digital resurrection of a deceased former Indonesian president to endorse a candidate running for office, or voice clones intended to make candidates appear as if they speak multiple languages. *Id.*

[15] Herbert B. Dixon Jr., *Deepfakes: More Frightening than Photoshop on Steroids*, 58 JUDGES' J. 35, 35 (2019).

[16] *Id.* at 35.

[17] Simon Jenkins, *The Moral of Kate's Picture-editing Debacle Is Simple: The Royal Family Should Tell All*, GUARDIAN (Mar. 11, 2024), https://www.theguardian.com/commentisfree/2024/mar/11/kate-picture-editing-royal-family-tell-all-secrets [https://perma.cc/45QT-YMGL]; Natalie Viebrock, *How Kate Middleton's Photoshop Fail Unveils the High Expectations of the Digital Age*, QUEEN'S J. (Mar. 22, 2024), https://www.queensjournal.ca/how-kate-middletons-photoshop-fail-unveils-the-high-expectations-of-the-digital-age/ [https://perma.cc/DHF4-U8EK]; Victoria

Deepfakes often involve the use of more than one algorithm. First, a series of photographs (or a photograph and a video) of two people are run through an AI algorithm referred to as an "encoder."[18] The encoder isolates and learns the similarities between the two faces and reduces them to their shared common features, compressing the images in the process.[19] A second algorithm, called a "decoder," is then taught to recover the faces from the compressed images.[20] To perform the face swap, the encoded images are fed into the "wrong" decoder; that is, the compressed image of Person A's face is fed into the decoder trained on Person B.[21] The Person B decoder then reconstructs the face of Person A on the body of Person B, with all the expressions and orientation of Person B.[22] Voila, we now have photo or video of Person A doing or saying things that never happened.

Early deepfake videos had telltale signs of inauthenticity: The subject may not have blinked properly, the lip syncing was slightly off, the skin tone could be patchy or too smooth, or there could be flickering around the edges of the transposed faces.[23] Fine details such as fingers, teeth, ears, hair strands, or jewelry often contained errors; additionally, there were often errors such as strange lighting, shadow, or perspective effects (e.g., inconsistent illumination or reflections on the iris or glasses, or faulty physics or geometry).[24] Voice clones had similar telltale signs of inauthenticity such as strange or jarring word choices, inconsistencies in pronunciation and enunciation, a flat, unemotional speaking tone, or strange background crackling or static noises.[25] Today, the technology has improved to the

---

Murphy, *How a Mother's Day Photo Led to a Palace Disaster*, TOWN & COUNTRY (Mar. 12, 2024), https://www.townandcountrymag.com/society/tradition/a60179448/kate-middleton-photoshop-kensington-palace-pr-disaster-analysis/ [https://perma.cc/X5S4-AV73].

[18] Neelam Rawat, *Decoding Deepfake Technology: The Rise, Impact, and Ethical Considerations*, KIET GRP. OF INSTS. BLOG (Jan. 9, 2024), https://www.kiet.edu/blog/department-of-computer-application/decoding-deepfake-technology-the-rise-impact-and-ethical-considerations/ [https://perma.cc/L6BA-QGYY].

[19] *Id*.

[20] *Id*.

[21] *Id*.

[22] *Id*.

[23] Matt Groh, *Detect Deepfakes: How to Counteract Misinformation Created by AI*, MIT MEDIA LAB (Jan. 2025), https://www.media.mit.edu/projects/detect-fakes/overview/ [https://perma.cc/G2HB-BBTC]; Chan Eu Imm, *CNA Explains: How to Spot the Telltale Signs of a Deepfake*, CHANNEL NEWS ASIA (Sep. 14, 2024), https://www.channelnewsasia.com/singapore/deepfakes-cna-explains-ai-how-4603136 [https://perma.cc/E95S-X4QD]; *see also* Negar Kamali et al., *How to Distinguish AI-Generated Images from Authentic Photographs,* ARXIV 1, 15-48 (June 12, 2024), https://arxiv.org/pdf/2406.08651 [https://perma.cc/KQ3B-QNX5] (discussing clues that can be used to distinguish deepfake content, including anatomical impossibilities, stylistic artifacts, and functional implausibilities, among others).

[24] *See* sources referenced *supra* note 23.

[25] Deep Media, *A Comprehensive Guide to Detecting Voice Cloning*, MEDIUM (Nov. 21, 2023), https://medium.com/@deepmedia/a-comprehensive-guide-to-detecting-voice-cloning-825f2738c00e [https://perma.cc/6ZGW-H3KT]; Miguel Jette, *How to Spot Deepfake Audio: 3 Tips for Detecting AI-Generated Speech*, REV (May 5, 2024), https://www.rev.com/blog/how-to-spot-deepfake-audio [https://perma.cc/LN7R-LE47].

point where the distinctions are subtle and often exceed normal human perception, requiring an expert to differentiate AI-generated from authentic content.[26]

What is the significance of this technological revolution to the resolution of contested cases in state and federal court? How should judges and lawyers prepare for the inevitable disputes involving whether relevant and probative—perhaps even determinative—evidence offered by one party to prove its case, is challenged by the other party as fake? Are the current rules of evidence adequate to fulfill the task of sorting out authentic from AI-generated evidence? We explore these and similar questions in this paper, beginning in Parts II to V with a discussion of the AI technology that has made generative AI such an important component of the present litigation landscape. We then turn in Part VI to the existing rules of evidence to explore whether they are up to the task that judges, lawyers, and juries will be required to face. Finally, in Part VII we explore new—bespoke—evidence rules that may help ensure that the integrity of the fact-finding function of our adversarial justice system is preserved. We conclude in Part VIII with practical guidelines for courts to follow in the interim, as reforms to the evidentiary rules are considered.

II. DEEPFAKES IN THE REAL WORLD

Deepfakes have democratized fraud and wreaked genuine havoc. In 2019, the head of a U.K. subsidiary of a German energy company deposited nearly £200,000 into a Hungarian supplier's bank account after receiving what he believed to be a phone call from the parent company's CEO directing him to make the transfer.[27] The voice on the call was a deepfake.[28] In 2024, a Hong Kong finance worker at a multinational firm was tricked into paying $25.6 million to fraudsters following a videoconference with what he believed to be several of his coworkers, including the company's CFO.[29] Other than the victim, all of the participants in the videoconference were deepfakes.[30]

These are not the only contexts in which deepfakes have recently appeared. In December 2023, a recording of a high school principal in suburban Maryland surfaced—and went viral online—in which the principal appeared to be caught

[26] *See* Ann-Marie Alcántara, *AI-Created Images Are So Good Even AI Has Trouble Spotting Some*, WALL ST. J (Apr. 11, 2023), https://www.wsj.com/articles/ai-created-images-are-so-good-even-ai-has-trouble-spotting-some-8536e52c.

[27] Jesse Damiani, *A Voice Deepfake Was Used to Scam a CEO out of $243,000*, FORBES (Sept. 3, 2019, 4:52 PM), https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/ [https://perma.cc/BM7N-876C].

[28] *Id*.

[29] Heather Chen & Kathleen Magramo, *Finance Worker Pays out $25 Million after Video Call with Deepfake 'Chief Financial Officer*,*'* CNN (Feb. 4, 2024), https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/ [https://perma.cc/7WAP-3JBL].

[30] *Id*.

making racist and antisemitic comments.[31] In addition to leaving death threats, angered parents and teachers forced the principal to be placed on administrative leave pending investigation.[32] The principal—through his union representative—vehemently denied making the offensive remarks and insisted that the voice on the recording was not his.[33] He turned out to be telling the truth. Months later it was determined that the recording was made by the high school's athletic director whose contract was not being renewed due to "frequent work performance challenges," including, but not limited to, an improper payment of approximately $2,000 made to the athletic director's roommate.[34] The email disseminating the recording was ultimately traced to a server connected to the athletic director, while forensic experts found "the recording contained traces of AI-generated content with human editing after the fact . . . ."[35] Although the principal remained employed by the Baltimore County Public School System, his reputation was irreparably damaged and he was no longer in charge of the Pikesville High School.[36] Even people aware that the recording was a deepfake remained angry about it because it "stayed with them," reviving their prior experiences of discrimination.[37] On January 9, 2025, the former Pikesville High School principal filed a lawsuit against the Baltimore County Public Schools and certain of its employees as well as the athletic director who created the deepfake audio recording.[38]

In 2021, a mother in Pennsylvania was accused of making and disseminating deepfake images of rival cheerleading teammates of her high-school-aged daughter,

---

[31] Thomas Lake, *A School Principal Faced Threats after Being Accused of Offensive Language on a Recording. Now Police Say It Was a Deepfake*, CNN (Apr. 26, 2024, 2:25 PM), https://www.cnn.com/2024/04/26/us/pikesville-principal-maryland-deepfake-cec/ [https://perma.cc/6F4Y-SJ43]; Marianna Spring, *The Racist AI Deepfake That Fooled and Divided a Community*, BBC (Oct. 4, 2024), https://www.bbc.com/news/articles/ckg9k5dv1zdo [https://perma.cc/9LNV-Z3VC]; David K. Li, *Teacher Arrested, Accused of Using AI to Falsely Paint Boss as Racist and Antisemitic*, NBC NEWS (Apr. 26, 2024, 9:55 AM), https://www.nbcnews.com/news/us-news/teacher-arrested-ai-generated-racist-rant-maryland-school-principal-rcna149345 [https://perma.cc/LFB5-LT9K].

[32] *See* sources referenced *supra* note 31.

[33] *See* sources referenced *supra* note 31.

[34] *See* Lake, *supra* note 31.

[35] *See* Li, *supra* note 31.

[36] *See* Lake, *supra* note 31.

[37] *See* Spring, *supra* note 31.

[38] Christian Olaniran & Stephon Dingle, *Former Pikesville High School Principal Sues Baltimore County Schools over Racist AI case*, CBS (Jan. 9, 2025, 6:26 PM), https://www.cbsnews.com/baltimore/news/pikesville-high-school-principal-sues-baltimore-county-schools-racist-ai-recording/ [https://perma.cc/FMD9-GK6G]; Anna Merod, *Former Principal Sues Baltimore County Schools over Alleged Racist AI Deepfake*, K-12 DIVE (Jan. 10, 2025), https://www.k12dive.com/news/baltimore-county-schools-lawsuit-principal-deepfake/737105/ [https://perma.cc/96GV-FUPG].

showing them partying partially clothed or nude, drinking alcohol, and/or vaping.[39] The teenagers insisted that the images were fake, and the mother who disseminated them—allegedly accompanied by threatening messages, including one urging a teen to commit suicide—was arrested for cyberbullying for sending harassing and incriminating fake images to the teenagers, their parents, and the gym where the cheerleaders trained.[40] While the mother was found guilty of three counts of misdemeanor harassment—for sending five anonymous messages about three teenagers who had posted self-incriminating images online—it turns out that the photos and videos attached to the messages were themselves real. The mother, like the principal in our previous example, was also ostracized by her community and faced death threats.[41] She was unable to return to work and eventually sued the investigating officer, the county police, the district attorney, and others for defamation and violation of her civil rights.[42]

These stories of fraud and the appearance of deepfakes—or alleged deepfakes—on social media are not isolated incidents. In 2023, there was a tenfold increase in deepfakes detected globally across all industries; in North America alone, there was an eighteen-fold increase in deepfake fraud cases.[43] In 2024, 49% of companies in the U.S., U.A.E., Mexico, Singapore, and Germany reported encounters with deepfake scam attempts.[44] Sixty percent of Americans have expressed a significant concern about deepfakes—more than any other AI-related risk.[45]

In the stories depicted above, the accused creators of the digital evidence were eventually exonerated following forensic investigation of the media at issue. But

---

[39] *Mother 'Used Deepfake to Frame Cheerleading Rivals*,*'* BBC (Mar. 15, 2021), https://www.bbc.com/news/technology-56404038 [https://perma.cc/DS3C-U9Q2]; Marlene Lenthang, *Cheerleader's Mom Created Deepfake Videos to Allegedly Harass Her Daughter's Rivals*, ABC NEWS, (Mar. 13, 2021, 6:13 PM), https://abcnews.go.com/US/cheerleaders-mom-created-deepfake-videos-allegedly-harass-daughters/story?id=76437596 [https://perma.cc/UZ56-M9AY].

[40] Christina Morales, *Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders*, N.Y. TIMES (Mar. 14, 2021), https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html [https://perma.cc/4QJU-395C].

[41] *Id*.

[42] *Id*.

[43] *Deepfake Technology*, PROOFPOINT, https://www.proofpoint.com/us/threat-reference/deepfake [https://perma.cc/LM9S-MU3K] (last visited Apr. 3, 2025); *see generally* JOHANNES TAMMEKÄND ET AL., SENTINEL, DEEPFAKES 2020: THE TIPPING POINT (2020), https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf [https://perma.cc/PT2A-H5KX] (overviewing the recent rise of deepfakes across all industries).

[44] GlobeNewswire, *Deepfake Fraud Doubles Down: 49% of Businesses Now Hit by Audio and Video Scams, Regula's Survey Reveals*, FIN. POST, (Sep. 30, 2024), https://financialpost.com/globe-newswire/deepfake-fraud-doubles-down-49-of-businesses-now-hit-by-audio-and-video-scams-regulas-survey-reveals [https://perma.cc/KUR6-ZJ4M].

[45] Catherine Chipeta, *Deepfake Statistics (2025): 25 New Facts for CFOs*, EFTSURE (July 12, 2024), https://eftsure.com/statistics/deepfake-statistics/ [https://perma.cc/L9SS-WNR3].

deepfakes can also "make it easier for liars to avoid accountability for things that are in fact true."[46] In 2018, Robert Chesney and Danielle Keats Citron introduced the term "liar's dividend" to refer to the benefit received by those who question legitimate information in order to confuse what is true with what is not.[47] In at least two criminal cases brought by the U.S. Government in the aftermath of the January 6, 2021, attack on the Capitol, defendants asserted a "deepfake defense." Defendants in *U.S. v. Doolin*[48] and *U.S. v. Reffitt*[49] challenged the authenticity of video evidence proffered against them by the prosecution, claiming that the videos could have been manipulated using deepfake technology. The deepfake defense was also attempted in a highly publicized civil case involving an Apple engineer who died in a fatal car crash when his Tesla vehicle—operated in autopilot mode—drove into a highway barrier.[50] In their wrongful death suit against Tesla, the engineer's family pointed to Tesla CEO Elon Musk's assertion that "[a] Model S and Model X at this point can drive autonomously with greater safety than a person. Right now."[51] Despite the existence of online video recordings of the conference where Musk made the claim, his lawyers suggested that the statement could have been fabricated using deepfake technology.[52] In her ruling, Judge Everette Pennypacker refused to condone the argument that public figures—who are ostensibly more likely to be targets of deepfake technologies than others—could avoid accountability for their public statements by claiming, after the fact, that they were fabricated.[53] She ordered Musk to appear for a deposition that he would ordinarily have avoided under the "Apex Doctrine."[54]

---

[46] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1758 (2019).

[47] *Id.* at 1785.

[48] Defendant's Response to U.S. Motion in Limine Regarding Authentication of Certain Video Evidence, at 2, U.S. v. Doolin, No. 21-cr-00477 (filed D.D.C. Aug 5, 2022), ECF No. 135.

[49] U.S v. Reffitt, 602 F. Supp. 3d 85, 90 (D.D.C. 2022).

[50] Guardian STAFF AND AGENCIES, *ELON MUSK'S STATEMENTS COULD BE 'DEEPFAKES', TESLA DEFENCE LAWYERS TELL COURT*, GUARDIAN, (APR. 23, 2023, 9:30 PM), https://www.theguardian.com/technology/2023/apr/27/elon-musks-statements-could-be-deepfakes-tesla-defence-lawyers-tell-court [https://perma.cc/3ANM-S9TP].

[51] Order Granting, in Part, Plaintiffs' Motion to Compel Written Discovery Responses and the Deposition of Elon Musk, at 4, Sz Huang *ex rel.* Wei Lun Huang v. Tesla, Inc., No. 19CV346663 (Cal. Super. Ct. Apr. 28, 2023) [hereinafter Tesla Order]. To watch the video at issue, see Recode, *Elon Musk | Full Interview | Code Conference 2016*, YOUTUBE (June 2, 2016), https://www.youtube.com/watch?v=wsixsRI-Sz4&t=4765s.

[52] Tesla Order, *supra* note 51, at 11-12.

[53] *Id.* at 12. Some have described Musk's argument as invoking "memelord immunity." *Can a Celebrity Claim "Memelord Immunity" over Videoed Statements Due to Deep Fakes?*, NORTON DIGIT. FORENSICS (May 29, 2023), https://notiondigitalforensics.com.au/cyber-security-news/memelord-immunity-deep-fake-videos [https://perma.cc/8LFZ-ZUFX].

[54] *Id*. The "Apex Doctrine" is a legal principle that limits or prevents the deposition of high-ranking corporate or government officials in certain circumstances. *See id.* at 7. The doctrine is intended to protect these individuals from harassment and abuse of the discovery process, and to prevent litigants from using depositions as a means to extract settlements. *See* Andrea L. McDonald, *Existing in Tension: Courts Grapple with the Apex Doctrine*, LITIG. NEWS, Winter 2025, at 23.

Family court is another venue that is ripe for the infiltration of deepfakes. In a 2020 U.K. child custody dispute, a father was portrayed as threatening and violent when an alleged telephone recording of him was played in court.[55] A subsequent analysis of the recording's metadata by counsel for the father revealed that the mother had used "software and online tutorials" to manufacture the recording.[56] Thus, it is clear that actually and allegedly AI-generated audio, video, and image evidence is likely to find its way into court in many different types of cases for the foreseeable future.

III. WAYS THAT AI-GENERATED AND POTENTIALLY AI-GENERATED EVIDENCE MAY PRESENT IN COURT

There are least four different ways that AI-generated evidence (or potentially AI-generated evidence) is likely to appear in court. The first is where both parties agree that the evidence at issue is the product of an AI system. For example, a potential employee is screened using a human resources AI screening tool and does not receive a job offer. The applicant sues the potential employer claiming that the AI tool or the employer's use of the AI tool was discriminatory. The second is where a party's expert discloses the use of an AI tool in support of an analysis set forth in their expert report. The third is where a party seeks to introduce an AI-enhanced exhibit or demonstrative, as was the case in *State v. Puloka*.[57] In *Puloka*—a case of first impression—the court rejected the admission of video exhibits enhanced by AI for use in a jury trial.[58] The defendant was charged with three counts of murder stemming from a 2021 shooting that was captured (unaltered) on a bystander's smartphone.[59] The defense sought to admit AI-enhanced versions of the video aimed at enhancing the clarity of the original video's low resolution, motion blur, fuzzy images, and "blocky" edge patterns.[60] To improve clarity, the defense expert used an AI video-editing tool to "intelligently scale up the video to increase resolution," add sharpness, definition, and smoother edges to the objects in the video.[61] The State challenged the admissibility of the AI-enhanced videos asserting that the defense had failed to satisfy the standard set forth in *Frye* v.

---

[55] Patrick Ryan, *Deepfake Audio Evidence Used in UK Court to Discredit Dubai Dad*, NAT'L NEWS (Feb. 8, 2020), https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764 [https://perma.cc/2HWK-6RQX]; Gabriella Swerling, *Doctored Audio Evidence Used to Damn Father in Custody Battle*, TELEGRAPH (Jan. 31, 2020), https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/ [https://perma.cc/R6AN-P5Q4].

[56] Ryan, *supra* note 55.

[57] State v. Puloka, No. 21-1-04851-2 (Wash. Super. Ct. Mar. 29, 2024).

[58] *See id.*, slip op. at 4-7.

[59] Kevin J. Quilty, *Washington Court Rejects Novel Use of AI-Enhanced Video in Trial*, GREENBERGTRAURIG (May 23, 2024), https://www.gtlaw.com/en/insights/2024/5/washington-court-rejects-novel-use-of-ai-enhanced-video-in-trial [https://perma.cc/2ZKQ-3L74].

[60] *Puloka*, slip op. at 2.

[61] *Id.*

United States[62]—a standard requiring that evidence using a novel scientific theory or principle must have "achieved general acceptance in the relevant scientific community."[63] After hearing oral argument from both parties, the court declined to admit the AI-enhanced videos finding that the machine-learning algorithm used by the defense expert to enhance the videos had not been peer reviewed by the forensic video analysis community and was not generally accepted in that community and, therefore, could not meet the *Frye* standard.[64] The court noted that the defense expert himself had admitted that he did not know how the AI was trained, whether it was tested for reliability, and could not explain how it worked.[65]

What these cases have in common is that the parties all agree that certain evidence has been generated, analyzed, or altered using AI. We refer to the evidence in these cases as "acknowledged AI-generated evidence." Here, the admissibility dispute in front of the court involves the so-called "*accuracy*" of the AI-generated evidence, or the AI tool used to create it. We contrast this with the situation where the admissibility dispute in front of the court involves the *authenticity* of the evidence; one party claims it is genuine, while the other asserts it is AI-generated, in whole or in part. This is the fourth way that AI-generated evidence (or potentially AI-generated evidence) is likely to appear in court. We refer to the evidence in such cases as "unacknowledged AI-generated evidence." These four situations present very different issues for the court.

In the first three instances, the dispute is essentially the same as any dispute involving the admissibility of novel scientific or technical evidence, where the questions facing the court center around the scientific bona fides or propriety of the evidence. What may make AI seem like a harder case than those involving other technologies is that its operation may be "black-box" or proprietary, such that it cannot be readily explained to the parties or the court.[66] Granted, there are many inventions and technologies as to which the modes of operation may not be transparent, including, but not limited to, the biological mechanisms surrounding many medicines and medical treatments.[67] But, in these circumstances, there are often other ways for developers and manufacturers to demonstrate the effectiveness or safety of their tools and products, including by independent testing, validation, publication, and peer review, to name a few.[68]

---

[62] *Id.* at 4.

[63] Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

[64] *Puloka*, slip op. at 6.

[65] *Id.* at 2-3.

[66] Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 CORNELL L. REV. 561, 563 (2024).

[67] *See, e.g.*, Glastetter v. Novartis Pharm. Corp., 252 F.3d 986, 992 (8th Cir. 2001) (affirming exclusion of expert testimony connecting a lactation-suppressing drug to stroke, noting that evidence of causation was not adequately presented).

[68] People routinely take aspirin or board airplanes without understanding how they work, trusting that they have been sufficiently tested and approved by regulatory bodies like the FDA or FAA. And indeed, in the context of drugs and medical devices regulated by the FDA, guidance

The authors eschew the use of the term "accuracy' to describe this requirement of the evidence or its source because it is vague, imprecise,[69] and does not properly account for the multiple factors the court must consider in resolving whether or not to admit the evidence, which are "validity," "reliability," and "bias" or prejudice. "Validity" refers to whether the AI tool measures or predicts what it is intended to.[70] A scale is a valid measure of weight; a ruler is not. "Reliability" refers to whether the AI tool measures or predicts what it is intended to, consistently under substantially similar circumstances.[71] A scale is reliable if it records one's weight, on the same day, at the same time, as the same, within a small margin of (random) error. "Bias" refers to a systematic distortion of a measurement or prediction due to one or more factors that should not be considered.[72] It is also viewed as a prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.[73]

We posit that many—but not all—courts have at their disposal the basic evidentiary tools they need to deal with considerations of validity and reliability. These are the factors set forth in Rule 702,[74] which embody the well-known *Daubert factors*:

> (1) whether the expert's technique or theory can be or has been tested . . . ; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the [relevant] scientific community.[75]

---

exists for establishing substantial evidence of effectiveness, demonstrating the relative availability of methodologies for assessing the functionality of these technologies. *See generally* U.S. FOOD & DRUG ADMIN., DEMONSTRATING SUBSTANTIAL EVIDENCE OF EFFECTIVENESS WITH ONE ADEQUATE AND WELL-CONTROLLED CLINICAL INVESTIGATION AND CONFIRMATORY EVIDENCE: DRAFT GUIDANCE (2023).

[69] As the second author (Grimm) is fond of pointing out, "even a broken watch is 'accurate' twice a day!"

[70] *See* Fiona Middleton*, The 4 Types of Validity in Research: Definitions & Examples*, SCIBBR (June 22, 2023), https://www.scribbr.com/methodology/types-of-validity/ [https://perma.cc/XN7J-M28Z].

[71] *See What's the Difference between Reliability and Validity?*, SCRIBBR, https://www.scribbr.com/frequently-asked-questions/reliability-and-validity/ [https://perma.cc/2RVU-HBF7] (last visited Mar. 24, 2024).

[72] *See Bias: Meaning & Use*, OXFORD ENG. DICTIONARY (Mar. 2025), https://www.oed.com/dictionary/bias_n?tab=meaning_and_use [hereinafter *Bias*, OXFORD ENG. DICTIONARY]; *Bias*, WIKIPEDIA (Mar. 18, 2025), https://en.wikipedia.org/wiki/Bias [https://perma.cc/6RCV-4ZRD]; see also *Types of Bias in Research: Definition & Examples*, SCRIBBR, https://www.scribbr.com/category/research-bias/ [https://perma.cc/KER9-TB5R] (last visited Mar. 24, 2025).

[73] *See Bias*, OXFORD ENG. DICTIONARY, *supra* note 72.

[74] FED. R. EVID. 702.

[75] FED. R. EVID. 702 advisory committee's note to 2000 amendment (referencing the factors laid out in Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993)).

The error rate of an AI tool is particularly important because false-positive and false-negative errors can have very different consequences, as was amply demonstrated in the well-known example of the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"). COMPAS is a tool that is used in many states to predict the likelihood of recidivism of a person charged with or convicted of a crime.[76] It is typically used for decisions about pre-trial release, sentencing, and/or parole.[77] In a 2016 exposé, ProPublica reported that COMPAS improperly made false- positive and false-negative errors along racial lines.[78] "False-positive" errors occur when the AI system incorrectly predicts the presence of 'X' (in the case of COMPAS, incorrectly predicts a high risk of recidivism) when 'X' is not actually present. A "false-negative" error occurs when the AI system incorrectly predicts the absence of 'X' (in the case of COMPAS, incorrectly predicts a low risk of recidivism) when 'X' actually is present.[79] ProPublica discovered that when COMPAS made false-positive errors, it incorrectly concluded that a Black person was likely to be a recidivist twice as often, and when it made false-negative errors, it incorrectly concluded that a White person was not likely to be a recidivist twice as often.[80] While ProPublica's analysis has been challenged by subsequent scholarship,[81] this ongoing debate illustrates one of the biggest challenges in measuring bias: We do not currently have public or even scientific consensus on what it means for an algorithm to be "unbiased" or "fair."[82]

To address the challenge of acknowledged AI-generated evidence, the authors propose a minor revision to Rule 901(b)(9) of the Federal Rules of Evidence ("the Rules")—"Evidence About a Process or System"—as described in greater detail below in Part VII, replacing the term "accurate" with the terms "valid and reliable," and requiring the proponent of the evidence to "describe[] the training data and [AI] software or program that was used" and demonstrate that "they produced valid and reliable results in this instance."

---

[76] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[77] NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS CORE 27-28 (2015), https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf [https://perma.cc/HJL4-8KWZ].

[78] Angwin et al, *supra* note 76.

[79] Prathamesh Patalay, *COMPAS: Unfair Algorithm?*, MEDIUM (Nov. 21, 2023), https://medium.com/%40lamdaa/compas-unfair-algorithm-812702ed6a6a [https://perma.cc/LGC9-Z9PZ].

[80] Angwin et al, *supra* note 76.

[81] *See, e.g.*, Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to* "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks," 80 FED. PROBATION 38 (2016) (reanalyzing ProPublica's data and arguing that there is no evidence of racial bias in COMPAS).

[82] *See* Angwin et al., *supra* note 76; Patalay, *supra* note 79 (discussing the different approaches and metrics for evaluating fairness that led to varying analysis by ProPublica and subsequent researchers).

The admittedly harder problem for the courts is the one where the actual authenticity of the evidence is in dispute. Currently, there is a very low threshold for the admissibility of non-testimonial evidence: The proponent must simply show that the evidence is more likely than not what the proponent claims it to be, typically referred to as the "preponderance" standard.[83] Questions of authenticity are generally left in the hands of the jury under Rule 104(b),[84] and the court can only withhold relevant evidence from the jury under Rule 403 if the evidence's likely prejudice substantially outweighs its probative value.[85] These standards are problematic in the era of deepfakes.

Take for example, a recording of one of the authors' voices. To have that recording admitted in court under Rule 901(b)(1)—"Testimony of a Witness with Knowledge"—or Rule 901(b)(5)—"Opinion About a Voice"—the proponent would simply need to present a witness who could testify that the voice on the recording was the author either based on personal experience or "based on hearing the voice under circumstances that connect it with the alleged speaker."[86] Presumably any decent deepfake of the author's voice would pass both those tests. Since it is likely that the court will deem the audio evidence probative, and because Rule 403 tilts strongly in favor of its admission, virtually nothing will prevent deepfake audio recordings from getting to the jury.[87] The problem is, once this audio evidence is heard by the jury, the effect is often permanent and indelible.[88] Dissimilar to an instruction to strike or ignore something, a court cannot unring this bell.[89]

IV. RESEARCH ON THE IMPACT OF AUDIOVISUAL EVIDENCE ON THE TRIER OF FACT

Many commentators have argued that today's deepfakes are no different than the forged and counterfeit evidence encountered by courts in the past. The authors disagree. First, the sheer scope, scale, and speed of deepfake evidence today is very different from the fakes of bygone days. A comprehensive report issued by Sentinel in 2020 noted that online deepfake videos nearly doubled from approximately 7,964 in 2018 to approximately 14,678 in 2019—and then increased by a factor of nearly ten to approximately 145,227 in 2020—showing exponential growth.[90] The majority of these deepfakes were posted on popular social media platforms, and

---

[83] FED. R. EVID. 901(a); *Preponderance of the Evidence*, LEGAL INFO. INST., https://www.law.cornell.edu/wex/preponderance_of_the_evidence [https://perma.cc/49Y7-RMCH] (last visited Mar. 25, 2025).

[84] FED. R. EVID. 104(b); *see* discussion *infra* Section VI.A.

[85] FED. R. EVID. 403; *see also* discussion *infra* Section VI.B.

[86] FED. R. EVID. 901(b)(1), (5).

[87] *See* discussion *infra* Section IV.B.

[88] *See, generally*, Taurus Myhand, *Once The Jury Sees It, The Jury Can't Unsee It: The Challenge Trial Judges Face When Authenticating Video Evidence in the Age of Deepfakes*, 29 WIDENER L. REV. 171, 180-182 (2023).

[89] *See id.*

[90] TAMMEKÄND ET AL. *supra* note 43, at 3.

"between them . . . amassed close to 6B views."[91] An estimated eight million Twitter (now X) accounts a week spread disinformation, and false stories reached an audience six times faster than true stories.[92] The Executive Summary of the Sentinel report concluded that 2020 was the tipping point where deepfakes began "to be used at scale to disrupt economies, influence political events, undermine journalism and wider society . . . . As the exponential trend continues, the majority of the world's digital information will be eventually produced by AI."[93] In fact, it has been estimated that by 2026, as much as 90% of online content will be synthetically generated.[94]

Today, virtually anyone can make convincing fake images, audios, and videos for free, in under five minutes.[95] There is no need for any skill, talent, sophistication, resources, or time, as was typically required in the past; the applications today are ubiquitous, cheap, easy to use, and remarkably convincing.[96] Second, the technology is improving at lightning speed and AI-detection technology, as we explain below in Part V, has simply not kept up with the proliferation of synthetic content.[97] Deepfake-detection technology is not expected to be sufficiently reliable or readily available to individuals in the immediate future or, possibly, ever.[98] There will be a continual arms race, as there was and is for spam filtering. Third, and most importantly, *deepfake evidence can have a profound impact on the cognition of the trier of fact*, particularly when that trier of fact is a lay juror.[99]

Rebecca A. Delfino was one of the first to write about this phenomenon in her 2023 paper, *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping*

---

[91] *Id.* at 3.

[92] *Id.* at 6.

[93] *Id.* at 4.

[94] Kimberly T. Mai et al., *Warning: Humans Cannot Reliably Detect Speech Deepfakes*, PLOS ONE, Aug. 2, 2023, at 1, 2 (citing NINA SCHICK, DEEP FAKES AND THE INFOCALYPSE: WHAT YOU URGENTLY NEED TO KNOW (2020)).

[95] *See, e.g.*, *AI Talking Head Video Generator*, SYNTHESIA, https://www.synthesia.io/tools/talking-head-video-maker? [https://perma.cc/WW36-J5SU] (last visited Mar. 25, 2025) (touting a service that can create videos of talking heads for free and in five minutes); Shannon Bond, *It Takes a Few Dollars and 8 Minutes to Create a Deepfake. And That's Only the Start*, NPR (Mar. 23, 2023, 5:00 AM), https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-start [https://perma.cc/3X4C-P779]; Lutz Finger, *Overview of How To Create Deepfakes - It's Scarily Simple*, FORBES (Sept. 8, 2022, 8:00 AM), https://www.forbes.com/sites/lutzfinger/2022/09/08/overview-of-how-to-create-deepfakesits-scarily-simple [https://perma.cc/SR2H-QFKX].

[96] *See* sources cited *supra* note 95.

[97] *See* discussion *infra* Part V.; Sarah A. Fisher et al., *Moderating Synthetic Content: the Challenge of Generative AI*, 37 PHIL. & TECH. (2024)

[98] *See id; see also* Gil Press, *Detecting Deepfakes: Fighting AI With AI*, FORBES (Aug. 7, 2024) https://www.forbes.com/sites/gilpress/2024/08/06/detecting-deepfakes-fighting-ai-with-ai/ [https://perma.cc/L68Z-T7TC].

[99] *See* discussion *infra* Part IV.

Role to Protect Legal Proceedings from Technological Fakery.[100] She observed that, "[i]n general, humans tend to accept images and other forms of digital media at face value"[101] and that "humans value visual perception above other indicators of truth."[102] She pointed to studies that demonstrate that "jurors who hear oral testimony along with video testimony are *650%* (i.e., seven times) more likely to retain the information . . . . Indeed, studies have demonstrated that video evidence powerfully affects human memory and perception of reality." [103] "The dangerousness of deepfake videos lies in the incomparable impact these videos have on human perception . . . . Video evidence is more cognitively and emotionally arousing to the trier of fact, giving the impression that they are observing activity or events more directly."[104]

In a 2010 study conducted at the University of Warwick, researchers examined the psychological impact of video on the reconstruction of personal observations.[105] The researchers had 60 college students engage in a computerized gambling task in a common room.[106] Following the completion of the task, half the students were shown a digitally altered video depicting a co-subject cheating, when none of the students had in fact cheated.[107] Nearly half of the subjects who viewed the video were willing to testify that they had *personally witnessed* the co-subject cheating after viewing the fake video, while only one in ten was willing to testify to the same after the researcher simply told them about the cheating incident, rather than showing them the deepfake evidence.[108]

Even more startling are the results of a 2009 study[109] reported by Tarus Myhand in *Once the Jury Sees It, the Jury Can't Unsee It: The Challenge Trial Judges Face When Authenticating Video Evidence in the Age of Deepfakes*.[110] Myhand argues that "[v]ideo evidence enjoys a ring of truth."[111] He highlights the study showing that nearly all of the subjects who viewed fake video evidence *falsely confessed to an act that they did not commit*.[112] In a controlled experiment, the subjects completed a computerized gambling task in which they were told to return the

---

[100] Rebecca A. Delfino, *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery*, 74 HASTINGS L.J. 293 (2023).

[101] *Id.* at 310-11.

[102] *Id.* at 311.

[103] *Id* (emphasis added).

[104] Myhand, *supra* note 88, at 175.

[105] Kimberley A. Wade et al., *Can Fabricated Evidence Induce False Eyewitness Testimony?*, 24 APPLIED COGNITIVE PSYCH. 899, 900 (2010).

[106] *Id.* at 901-02.

[107] *Id.* at 903-04.

[108] *Id.* at 904-05.

[109] Robert A. Nash & Kimberley A. Wade, *Innocent but Proven Guilty: Eliciting Internalized False Confessions Using Doctored-Video Evidence*, 23 APPLIED COGNITIVE PSYCH. 624 (2009).

[110] Myhand, *supra* note 88, at 175.

[111] *Id.*

[112] *Id.*

money to the bank if they answered a question incorrectly.[113] Later, researchers falsely accused the subjects of cheating by stealing the money, presenting some of them with digitally fabricated video evidence of them taking money that did not belong to them, and falsely telling the rest of the subjects that such video evidence existed.[114] When presented with or informed of the doctored evidence, *all of the subjects confessed*, and most internalized their belief in their guilt—i.e., they actually believed that they had taken money they should not have.[115]

Accordingly, in this new world of deepfakes, jurors—let alone witnesses—can no longer blithely trust even their own perceptions and memories. This phenomenon is not limited to videos; in one experiment, 40% of subjects exposed to doctored photographs purportedly from their childhood reconstructed false memories based on those photos.[116] Moreover, there is also a phenomenon known as the "continued influence effect," which is the tendency for misinformation to influence a person's reasoning even after it has been corrected.[117] Studies show that people exposed to misinformation are unable to easily discard that information even when they subsequently receive contradictory or corrective information.[118] The more times a person is exposed to the divergent post-event misinformation, the more likely it is that a person's memory will be tainted.[119] Thus, after viewing a deepfake post-event, a person may no longer be able to distinguish what they themselves originally observed from what was suggested to them after the fact.[120] Thus, judicial instructions provided to jurors to "disregard" audiovisual evidence they have seen are unlikely to be effective.

Because of these psychological impacts and, as we shall see in the following section, the fact that humans cannot effectively and reliably distinguish AI-generated from human-generated content—and cannot yet count on the reliability

---

[113] Nash & Wade, *supra* note 109, at 625.

[114] *Id.*

[115] *Id.* at 629-30.

[116] Miriam S. Johnson et al., *Doctored Photographs Create False Memories of Spectacular Childhood Events. A Replication of Wade et al. (2002) with a Scandinavian Twist*, 31 MEMORY 1011, 1014 (2023) (replicating results from an earlier study in 2002 showing the same effect in 50% of subjects).

[117] Ulrich K. H. Ecker et al., *The Psychological Drivers of Misinformation Belief and Its Resistance to Correction*, 1 NAT. REVS. PSYCH. 13, 15 (2022).

[118] *See id.* (discussing "the typical CIE [continued influence effect] laboratory paradigm"*)*; *see also* Nathan Walter & Riva Tukachinsky, *A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?* 47 COMMC'N. RSCH. 155, 168 (2020) (finding that continued influence effect was observed across multiple studies).

[119] Ralph Norman Haber & Lyn Haber, *Experiencing, Remembering and Reporting Events*, 6 PSYCH. PUB. POL'Y & L. 1057, 1069 (2000). Post-event misinformation refers to new false information obtained after the initial acquisition of information. *Id.* at 1068.

[120] *See, e.g.*, Gillian Murphy et al., *Face/Off: Changing the Face of Movies with Deepfakes*, PLOS ONE, July 6, 2023, at 1, 7-8 (Study participants "readily formed false memories" after being falsely told that well-known films had been remade with well-known modern actors, like The Matrix starring Will Smith, and then being shown fabricated movie remakes in which deepfake technology replaced the original actors.).

of AI-detection technology[121]—cases involving unacknowledged AI-generated evidence will invariably need to involve experts who may not be available to the parties or the court in many state-court and low-value or criminal federal-court cases, and who will undoubtedly add delay and cost to complex litigation in state and federal court. As a consequence of the intractability of this problem, the fix is also more complicated than in the case of acknowledged AI-generated evidence, and in this case, the authors argue for the need for a new Rule 901(c). As explained in detail in Part VII below, the authors propose removing the authenticity determination from the jury's hands under Rule 104(b)) when there is evidence that a reasonable juror could equally find to be authentic or AI-generated. The new rule would bolster the gatekeeping role of the judge by adding a familiar balancing test—weighing the probative value versus the prejudice posed by the unacknowledged AI-generated evidence—before the potentially AI-generated evidence is admitted and shown to the jury.

V. THE CURRENT STATE OF THE ART IN DEEPFAKE DETECTION: BOTH HUMAN AND ALGORITHMIC

There are various information sources and guidelines that have been promulgated to help distinguish human-generated content from AI-generated content. One example is "Detect Fakes," an online tool developed by MIT Media Labs and maintained by Northwestern University.[122] "Detect Fakes" aims to collect data about commonly identifiable distinctions between authentic and AI-generated images.[123] In June 2024, Northwestern University published a comprehensive guide detailing signs to look for to identify fake images, such as anatomical or sociocultural implausibilities, inconsistent rendering of texture, light, shadows, and text, and/or violations of physics, including perspective.[124] The problem with these guides is that deepfake technology is improving so rapidly that any guidance from six months ago may already be outdated.[125]

Regrettably, studies show that humans are not very good at making judgments about authenticity, regardless of whether the media is visual or auditory. "In the context of images, multiple studies show that humans do not perform much better than chance," and "are overconfident in their deepfake-detection abilities."[126] For example, in a 2024 study of 244 participants from all over the U.S., more than half

---

[121] *See* discussion *infra* Part V.

[122] DETECT FAKES, https://detectfakes.kellogg.northwestern.edu/ [https://perma.cc/CKV4-Q4VE] (last visited Mar. 16, 2025).

[123] *About*, DETECT FAKES, https://detectfakes.kellogg.northwestern.edu/about [https://perma.cc/2NJU-WUG7] (last visited Mar. 16, 2025).

[124] Negar Kamali et al., *How to Distinguish AI-Generated Images from Authentic Photographs*, ARXIV 7 (Jun. 12, 2024), https://arxiv.org/pdf/2406.08651 [https://perma.cc/G2G5-FD36].

[125] *See* Matthew Hutson, *Detection Stays One Step Ahead of Deepfakes—for Now*, IEEE SPECTRUM (Mar. 6, 2023), https://spectrum.ieee.org/deepfake [https://perma.cc/56QR-QKPG].

[126] Mai et al., *supra* note 94, at 4.

could not distinguish between AI-generated and genuine content.[127] Another 2024 study, with 3,002 subjects from the U.S., Germany, and China tested on text, image, and audio detection, found that current state-of-the-art deepfake media was basically indistinguishable from "real" media, leading most participants to simply guess which was which.[128] AI-generated media was rated as more likely to be human-generated for all media types in all three countries.[129] "The average detection accuracy of participants [was] below 50% for images and never exceed[ed] 60% for the other media types."[130]

A 2024 study involving high-quality videos from Facebook (sixty genuine and sixty deepfake) found that humans were confused by high-quality deepfakes 75.5% of the time.[131] Curiously, algorithms performed quite differently than humans; they struggled to detect videos that appeared obviously fake to humans but were sometimes able to accurately detect fake videos that were difficult for people to discern.[132]

There are fewer studies that examine how well humans can detect speech deepfakes. A comprehensive 2023 study found that humans did not fare much better with this type of media, regardless of how many times they could listen to the audio clip before rendering a decision.[133] Subjects listened either to a single audioclip that they were asked to classify as bona fide or synthetic, or to two audiotapes, one of which was real and the other synthetic.[134] Listeners overall had somewhat limited detection capabilities; they made correct classifications 70.35% of the time in the first scenario and 85.59% in the second, which was a less realistic setting because listeners are generally unlikely to have multiple utterances containing the exact same speech to compare simultaneously.[135] Familiarizing listeners with examples of deepfakes in advance of the detection experiment boosted detection ability, but only by a small degree (3.84%).[136]

---

[127] UTAH VALLEY UNIV. CTR. FOR NAT'L SEC. STUD., RESEARCH SUMMARY: IMPACT OF AI GENERATED MEDIA 5 (2024), https://www.uvu.edu/news/2024/10/media/research-summary.pdf [https://perma.cc/K2BK-GCL6]; s*ee also* Hugo Rikard-Bell & Lindsay Aerts, *Deepfakes Fool More Than Half of Americans, UVU Study Shows*, KSLNEWSRADIO (Oct. 29, 2024, 4:54 PM), https://kslnewsradio.com/elections-politics-government/elections/uvu-deepfake-study/2149385/ [https://perma.cc/H8CL-HGPJ].

[128] Joel Frank et al., *A Representative Study on Human Detection of Artificially Generated Media Across Countries*, 2024 IEEE SYPM. ON SEC. & PRIV. 55, 55 (2024).

[129] *See id.*

[130] *Id.*

[131] Pavel Korshunov & Sebastien Marcel, *Subjective and Objective Evaluation of Deepfake Videos*, 2021 IEEE INT. CONF. ON ACOUSTICS, SPEECH & SIGNAL PROCESSING 2510, 2513.

[132] *Id.*

[133] Mai et al., *supra* note XX, at 10.

[134] *Id.* at 5-6.

[135] *Id.* at 8-9.

[136] *Id.* at 10.

The same 2023 study examined audio deepfake detection by humans in both English and Mandarin.[137] No significant difference was found between the detectability of deepfakes in each language; however, speakers of each language used different methods to detect fake audio.[138] English speakers tended to mention irregular breathing as a sign of AI-generated audio, while Mandarin speakers primarily mentioned fluency and word pacing.[139] This suggests that deepfake "tells" can vary with the suggested cultural background of the generated person and that recordings may need to be evaluated by someone well-versed in a particular culture or language in order to be distinguished.

Many believe that it is only a matter of time until deepfake-detection technology catches up with and solves the deepfake problem for us.[140] What these people may not realize is that the challenge in developing such tools may be inherent to generative AI technology more generally—and GAN technology more specifically—where, as the discriminator gets better, so too does the generator, such that the two algorithms may be locked in a perpetual arms race.[141] It may be impossible to develop a discriminator that will prevail once and for all.

A related technical problem is that while Developer A may be able to develop a discriminator that works well in detecting the use of generative AI tool 'A' (developed by Developer A), the same discriminator often will not work well in distinguishing the use of generative AI tool 'B' (developed by Developer B). For example, OpenAI's image-detection tool was shown to correctly flag 98% of the images generated by its own DALL•E image generator, even when they were manipulated in an effort to defeat detection, but it only correctly flagged 5-10% of images generated by competitors' image generators.[142] The same may be true as to ElevenLabs' AI audio detector, which reported a high (>90%) detection rate for its

[137] *Id.* at 1.

[138] *Id.* at 10, 14-15

[139] *Id.* at 15.

[140] *See, e.g.*, Ping Liu et al., *Automated Deepfake Detection,* ARXIV 1-3 (Aug. 12, 2021), https://arxiv.org/pdf/2106.10705 [https://perma.cc/K4LG-EGBN]; Hutson, *supra* note 125; *Intel and Intel Labs Develop New AI Methods to Restore Trust in Media*, INTEL, https://www.intel.com/content/www/us/en/research/blogs/trusted-media.html [https://perma.cc/9MSC-GQDT] (last visited Mar. 16, 2025); Umur Aybars Ciftci et al., *FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (forthcoming) (manuscript at 1-2) (available via IEEE early access).

[141] Linda Laurier et al., *The Cat and Mouse Game: The Ongoing Arms Race Between Diffusion Models and Detection Methods*, ARXIV 2-3 (Oct. 24, 2024), https://arxiv.org/pdf/2410.18866v1 [https://perma.cc/86ZU-ZLSU] .

[142] *Understanding the Source of What We See and Hear Online*, OPENAI (Aug. 4, 2024), https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/ [https://perma.cc/JZ2Z-867W].

own synthetic audio but failed to report how well it detected synthetic audio generated by other audio generators.[143]

A related issue is that the reliability and error rates of these detection technologies are often not disclosed to the public. For example, after the release of various ChatGPT text discriminators, it was discovered by third parties that many had unacceptably high false-positive error rates, especially by labelling the text of non-native English speakers as AI-generated.[144] Some of these tools were even pulled from use,[145] and many universities will not rely on the results of such tools when investigating alleged cases of cheating or plagiarism.[146]

Regardless of whether the effort is undertaken by a forensic expert, by using an algorithmic solution, or by a combination of both, there are two basic types of deepfake detection: (1) inference-based, and (2) provenance-based.[147] Both involve direct inspection of the media for signs of irregularity or inconsistency with what one would expect in the "real world."[148] The first method—inference-based—looks for signals in the actual media *content* that do not sync, line up, or match properly, such as fabric textures, reflections on the cornea, lines on the floor not extending back in the proper perspective, mispronunciations of words, irregular breathing patterns, peculiar background noises, etc.[149] The second method—provenance-

---

[143] *AI Speech Classifier,* ELEVEN LABS, https://elevenlabs.io/blog/ai-speech-classifier [https://perma.cc/CX33-JCC2] (last visited Mar. 16, 2025); *See also* Rowan Philp, *How to Identify and Investigate AI Audio Deepfakes, a Major 2024 Election Threat*, GLOB. INVESTIGATIVE JOURNALISM NETWORK (Feb. 26, 2024), https://gijn.org/resource/tipsheet-investigating-ai-audio-deepfakes/ [https://perma.cc/B35L-QRDA].

[144] Andrew Myers*, AI-Detectors Biased against Non-Native English Writers*, STANFORD INST. FOR HUMAN-CENTERED AI (May 15, 2023), https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers [https://perma.cc/LL8C-BXA4].

[145] Emily Forlini, *OpenAI Quietly Shuts Down AI Text-Detection Tool over Inaccuracies*, PCMAG (July 25, 2023), https://www.pcmag.com/news/openai-quietly-shuts-down-ai-text-detection-tool-over-inaccuracies [https://perma.cc/8K6L-U4CK].

[146] *See, e.g.*, *Generative AI: Encouraging Academic Integrity*, UNIV. OF PITTSBURGH: UNIV. CTR. FOR TEACHING & LEARNING, https://teaching.pitt.edu/resources/encouraging-academic-integrity (last updated Jan. 24, 2025) [https://perma.cc/YK4F-NESF]; *AI & Academic Integrity*, CORNELL UNIV.: CTR. FOR TEACHING INNOVATION, https://teaching.cornell.edu/generative-artificial-intelligence/ai-academic-integrity [https://perma.cc/WC5Q-8PFR] (last visited Mar. 16, 2025).

[147] *Deepfake Detection: Provenance, Inference, and Synergies Between Techniques*, REALITY DEFENDER (Mar. 24, 2024), https://www.realitydefender.com/blog/provenance-and-inference [https://perma.cc/Y2AX-66C9] [hereinafter *Deepfake Detection*, REALITY DEFENDER].

[148] *Id.*

[149] *See, e.g.*, Joel R. McConvey, *Everything you need to know about deepfake detection – for now*, BIOMETRIC UPDATE (Oct. 23, 2024), https://www.biometricupdate.com/202410/everything-you-need-to-know-about-deepfake-detection-for-now [https://perma.cc/ZBM2-PWFA] (explaining how the audio deepfake detection tool "SAFE and Sound" identifies artifacts in higher frequencies, including phase mismatches, irregular annunciation, and breathing cadence inconsistencies); Umur Aybars Ciftci et al., *How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals*, ARXIV 1 (Aug. 26, 2020), https://arxiv.org/pdf/2008.11363 [https://arxiv.org/pdf/2008.11363] (discussing using biological signals, such as

based—involves the review of metadata, such as timestamps and GPS coordinates, to look for signals suggesting the use of AI in generating or manipulating the media that are not content-based per se.[150] Neither approach is certain or foolproof; they merely allow an expert to opine with some degree of confidence that the content of the media appears to be inconsistent with authentic media or that the metadata or other digital artifacts suggest that the media was created or altered using AI.[151]

To increase transparency around provenance, many companies are beginning to implement "watermarks"—either discernable or hidden signals embedded on the face of the media or in its metadata—which can be used in identifying the source or origin of media.[152] The problem with watermarks is that they can readily be removed from synthetic media or added to genuine media.[153] As of yet, there are no reliable, permanent watermarks, and the tools that exist are not readily accessible to the creators of all types of media.[154] For example, provenance indicators cannot

---

irregular breathing patterns, to identify deepfake); Edmund L. Andrews, *Using AI to Detect Seemingly Perfect Deep-Fake Videos*, STANFORD INST. FOR HUMAN-CENTERED AI (Oct. 13, 2020), https://hai.stanford.edu/news/using-ai-detect-seemingly-perfect-deep-fake-videos [https://perma.cc/X3DA-J9PB] (discussing deepfake detection techniques that look for inconsistencies between mouth formations and phonetic sounds); Dan Milmo et al., *Weird Hands, Dodgy Numbers: Seven Signs You're Watching a Deepfake*, GUARDIAN (July 1, 2024), https://www.theguardian.com/technology/article/2024/jul/01/seven-signs-deepfake-artificial-intelligence-videos-photographs [https://perma.cc/V8GH-DLCH] (discussing deepfake detection that relies on mispronunciations of words and distorted floor patterns); Shu Hu et al., *Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights*, ARXIV 1 (Aug. 26, 2020), https://arxiv.org/pdf/2008.11363 [https://perma.cc/V9MT-ZXQA] (introducing deepfake detection based on extraction and comparison of corneal specular highlights).

[150] *See Deepfake Detection*, REALITY DEFENDER, *supra* note 147.

[151] *See e.g.*, Huo Jingnan, *Using AI to Detect AI-generated Deepfakes Can Work for Audio — but Not Always*, NPR (Apr. 5, 2024, 5:23 AM), https://www.npr.org/2024/04/05/1241446778/deepfake-audio-detection [https://perma.cc/UL7Q-WG3Q] (noting that while AI tools can detect deepfake audio, they are not foolproof and can struggle with nuanced manipulation and adversarial techniques); *Understanding the Source of What We See and Hear Online*, *supra* note 142.

[152] *See, e.g.*, *Watermarks Are Just One of Many Tools Needed for Effective Use of AI in News*, CTR. FOR NEWS, TECH. & INNOVATION, https://innovating.news/article/watermarks-are-just-one-of-many-tools-needed-for-effective-use-of-ai-in-news/ [https://perma.cc/SM22-CZSJ] (last updated Dec. 3, 2024) (discussing companies including Google, Meta and OpenAI implementing watermarks in AI generated content); Lev Craig, *AI Watermarking*, TECHTARGET, https://www.techtarget.com/searchenterpriseai/definition/AI-watermarking [https://perma.cc/KQQ3-V9QN] (last visited Mar. 15, 2025) (explaining mechanisms and limitations of watermarks).

[153] *See* Xuandong Zhao et al., *Invisible Image Watermarks Are Provably Removable Using Generative AI*, ARXIV 1 (Oct. 31, 2023), https://arxiv.org/pdf/2306.01953 [https://perma.cc/RVB8-GWMH]; Guanlin Li et al., *Warfare: Breaking the Watermark Protection of AI-Generated Content*, ARXIV 1 (Mar. 8, 2023), https://arxiv.org/pdf/2310.07726v3 [https://perma.cc/G62F-7RVC].

[154] *See e.g.*, Justyna Lisinska, *Watermarking in Images Will Not Solve AI-Generated Content Abuse*, CTR. FOR DATA INNOVATION (Aug. 15, 2024), https://datainnovation.org/2024/08/watermarking-in-images-will-not-solve-ai-generated-content-abuse [https://perma.cc/J2XZ-HUBF] (noting a lack of reliable watermarking solutions); Bob Gleichauf & Dan Geer, *Digital Watermarks Are Not Ready for Large Language Models*, LAWFARE (Feb. 29, 2024, 8:00 AM), https://www.lawfaremedia.org/article/digital-watermarks-are-not-ready-for-large-language-models [https://perma.cc/AX3A-A72E] (discussing the challenges of concealing watermarks in textual content).

be properly incorporated into media generated in the past, and with most contemporary watermarking tools, the decision whether or not to add a watermark at the time of creation may be at the user's or platform's discretion.[155] The bottom line is that lay people are not skillful at the detection task; there is no readily accessible, easy-to-use, and reliable technical solution that is available across different content generators, and the courts simply cannot wait until uniform, industry-wide provenance standards, like C2PA,[156] are implemented on every device.

With this understanding of the challenges generative AI applications create for distinguishing fact from fiction in the litigation system under our belt, we now turn our attention to the current rules of evidence that judges and lawyers will be called on to use to meet these challenges and explore whether they are sufficient. Where we believe the rules are insufficient, we suggest how they might be strengthened by modifications or new rules tailored to this specific challenge.

VI. THE APPLICABLE RULES OF EVIDENCE

When called upon to determine whether scientific evidence is admissible in civil and criminal cases, judges must apply the rules of evidence.[157] While Rule 702 deals specifically with the admissibility of scientific, technical, and specialized evidence, it must be applied in concert with a series of other rules that address: (1) the role of the judge and jury in the decision as to whether the evidence is admissible

---

[155] *See* Alex Hern, *'Time is Running Out': Can a Future of Undetectable Deepfakes Be Avoided?*, GUARDIAN (Apr. 8, 2024), https://www.theguardian.com/technology/2024/apr/08/time-is-running-out-can-a-future-of-undetectable-deepfakes-be-avoided [https://perma.cc/WZJ7-P358] (noting that smaller companies might not devote resources to watermarking and that users of open source platforms could create "forks" that do not implement watermarks); *Label AI Content on Facebook*, FACEBOOK, https://www.facebook.com/help/7434563519957988/ (last visited Apr. 5, 2025) (laying out AI labeling guidelines that do not extend to all content that is generated or altered by AI, suggesting that users still have discretion in deciding to watermark their posts).

[156] COAL. FOR CONTENT PROVENANCE & AUTHENTICITY, https://c2pa.org/ [https://perma.cc/H3FY-NENV] (last visited Mar. 15, 2025).

[157] For convenience, we will refer in this paper to the Federal Rules of Evidence, which are applicable in federal courts in all 50 states. While each state has enacted its own rules of evidence, in the main, they either follow the Federal Rules of Evidence, or represent a modified version thereof, or address the same evidentiary concepts. As will be noted in the discussion about the evidentiary rules governing scientific, technical, and specialized evidence, most of the states have adopted the same approach as that set forth in FED. R. EVID. 702, as it was amended in 2000 to address the Supreme Court's decisions in Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993), Gen. Elec. Co. v. Joiner, 118 S. Ct. 512 (1997), and Kumho Tire Co. v Carmichael, 119 S. Ct. 1167 (1998) (hereinafter the "*Daubert* Standard" or the "*Daubert* Test"). Still, about a quarter of the states have eschewed the *Daubert* Test, preferring instead to adhere to the much older standard for admitting scientific evidence, the "general acceptance" test first articulated in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). *See, e.g.*, Rochkind v. Stevenson, 471 Md. 1, 15 (Md., 2020) (noting that "[a] supermajority [39 out of 50, or 78%] of jurisdictions have departed from [the] Frye [test] in favor of the flexible Daubert approach"). As we explain in more detail *infra*, the *Daubert* Test established a multi-factor test for admitting scientific, technical, and specialized evidence. While it includes "general acceptance" as one of the factors, it adds others that were intended to make the analysis more flexible.

(i.e., Rule 104(a) and 104(b)); (2) whether the evidence is relevant to the issues that must be resolved in the case (i.e., Rule 401); (3) whether, if relevant, the evidence should nonetheless be excluded because it is unfairly prejudicial (i.e., Rule 403); and (4) whether scientific and technical evidence is "authentic" (meaning that it is what the party offering it says it is) (i.e., Rule 901-903). In this section, we will describe how these various rules impact the ultimate decision as to whether scientific evidence should be admitted into evidence, particularly in the context of the challenges presented by deepfake evidence.

A. Relevance and the Roles of the Judge and Jury in Admitting Evidence

It is no exaggeration to say that the Federal Rules of Evidence are jury-centric, meaning that the rules implicitly contemplate that juries comprised of lay members of the public will resolve factual disputes and determine how much, if any, of the evidence that is admitted is worthy of belief.[158] The rules allocate roles between the trial judge and the jury. The trial judge acts as a "gatekeeper," charged with deciding whether the jury may consider the litigant's proffered evidence.[159] Rule 104(a) describes this function: "The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except for privilege."[160] While the judge decides whether the jury will hear the evidence, the jury decides how much weight to give it and can entirely disregard evidence that it finds insufficient.[161]

But when the relevance of evidence is disputed, meaning that the facts that one party relies on to establish the relevance of the evidence are challenged by an opposing party, then the roles of the judge and jury are more complicated. Rule 104(b) governs this situation, and states "[w]hen the relevance of evidence depends on whether a fact exists, proof must be introduced sufficient to support a finding that the fact does exist. The court may admit the proposed evidence on the condition that the proof be introduced later."[162]

Read in the abstract, this rule is enigmatic. To better understand it, we must start with the definition of "relevant" evidence. Rule 401 states "[e]vidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be

---

[158] Given that the right to a jury trial in both criminal and civil cases is enshrined in the U.S. Constitution, this is unsurprising. *See, e.g.*, U.S. CONST. amend. VI, cl. 1 ("In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the state and district wherein the crime shall have been committed . . . ."); U.S. CONST. amend. VII ("In Suits at common law, where the value in controversy shall exceed twenty dollars, the right of a trial by jury shall be preserved, and no fact tried by a jury, shall be otherwise reexamined in any Court of the United States, than according to the rules of the common law.").

[159] *Daubert*, 509 U.S. at 598.

[160] FED. R. EVID. 104(a).

[161] FED. R. EVID. 104(e) ("This rule [104] does not limit a party's right to introduce before the jury evidence that is relevant to the *weight or credibility* of other evidence." (emphasis added)).

[162] FED. R. EVID. 104(b).

without the evidence; and (b) the fact is of consequence in determining the action."[163] Simply put, evidence is "relevant" if it has any tendency—however slight—to prove or disprove a fact or facts that must be established given the nature of the claims or charges alleged in the civil or criminal case, and the legal defenses offered to challenge those claims or charges.[164] The party that initiates the civil suit or criminal case determines the nature of the legal claims, which it must then offer evidence to prove. Similarly, the party that is sued or charged decides the legal defenses that it intends to assert to oppose the claims or charges brought. Thus, the evidence each side offers is "relevant" if it has any tendency to support or undermine the proof offered by the party bearing the burden of proving their claims, charges, or defenses.[165] The words "any tendency" emphasize that the threshold for showing relevance is very low.[166] Evidence does not have to be convincing to be relevant; it only has to have a logical tendency to prove or disprove a consequential fact.[167] The Advisory Committee Note to Rule 401 helpfully describes relevance:

> Relevancy is not an inherent characteristic of any item of evidence but exists only as a relation between an item of evidence and a matter properly provable in the case. Does the item of evidence tend to prove the matter sought to be proved? Whether the relationship exists depends upon principles evolved by experience or science, applied logically to the situation at hand . . . .
>
> The standard of probability under the rule is "more . . . probable than it would be without the evidence." Any more stringent requirement is unworkable and unrealistic.[168]

With this in mind, Rule 104(b) becomes clearer, especially if it is examined through a hypothetical fact pattern. Consider a case in which one party—Party A—sues another—Party B—for defamation because Party B published false information that injured the reputation of Party A to third parties. Party A wants to introduce an email that purports to be from Party B to A's employer, accusing A of embezzling money from the employer. The email is only relevant to prove defamation by Party B if B was responsible for composing and sending it. But suppose that Party B disputes that they wrote and sent the email. Party B claims that it is a fake email sent by someone else via a software application that allowed the user to fabricate the text of the email, Party B's actual email address, and the email address of Party A's employer, producing an email purportedly sent by Party B from their email account, when it actually was not.

---

[163] FED. R. EVID. 401.

[164] *See* Old Chief v. United States, 519 U.S. 172, 179 (1997).

[165] *See* FED. R. EVID. 401.

[166] *Daubert*, 509 U.S. at 587 ("The Rule's basic standard of relevance . . . is a liberal one"); FED. R. EVID. 401 Advisory Committee's Note to 1972 proposed rules.

[167] *See* FED. R. EVID. 104(e) (acknowledging that relevant and admissible, evidence may have no weight in the eyes of the jury, and that admitted relevant evidence may still be challenged as having little credibility.

[168] FED. R. EVID. 401 Advisory Committee's Note to proposed rules (citations omitted).

In this situation, the email is relevant to prove Party B's defamation of Party A only if Party B was, in fact, responsible for it being drafted and sent to A's employer. If it was, it has a tendency to prove defamation by Party B; if it was not, it has no tendency to prove Party B's defamation of Party A and is irrelevant, therefore, inadmissible. Party A says Party B wrote and sent it; B says they did not. Who decides this dispute: the judge, under Rule 104(a), or the jury, under Rule 104(b)?

The answer is that the jury normally must decide by considering the evidence offered by Parties A and B and determining which version it finds more credible. If it agrees with Party A's evidence, the jury may consider the email in deciding if Party B defamed Party A. If it agrees with Party B's evidence, the email is irrelevant, and inadmissible. The role of the judge under Rule 104(a) is simply to assess whether a reasonable jury could find, more likely than not, that Party B wrote and sent the email. If a reasonable jury could not find under the proffered facts that Party B wrote and sent it, it is not relevant and the judge may not allow the jury to consider it. But if the judge finds that a reasonable jury could find that Party B wrote and sent the email, then the judge must allow the jury to hear the evidence and give it the weight they think it deserves. This concept is often described as "conditional relevance."[169] Before the email can be admitted against Party B there must be *sufficient* evidence that Party B actually wrote and sent it.[170] As we explain below, the role of Rule 104(b) becomes especially important when one party claims that evidence offered against them is a deepfake.

B. Unfair Prejudice

As noted, the threshold for establishing that evidence is relevant is low. If evidence is relevant under Rule 401, then Rule 402 creates a presumption that it is admissible, unless the Constitution, a statute, a rule of procedure, or other source of law makes it inadmissible.[171] If establishing relevance is an easy task, and relevant evidence is presumed to be admissible, then there is a danger that marginally relevant evidence (i.e., that having some, but not much, tendency to prove a consequential fact) may carry with it the danger of unfair prejudice to the party against whom it is introduced, especially when the audience receiving it is composed of lay members of the public who may be misled by the evidence. Rule 403 exists to counterbalance this danger. It states that "[t]he court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence."[172] As the rule makes quite clear, it is for the *judge* to make the determination of

[169] *See* FED. R. EVID. 104 Advisory Committee's Note to proposed rules.

[170] *Id.* ("[I]f a letter purporting to be from Y is relied upon to establish an admission by him, it has no probative value unless Y wrote it or authorized it. Relevance in this sense has been labeled 'conditional relevancy.'").

[171] FED. R. EVID. 402.

[172] FED. R. EVID. 403.

whether the danger of unfair prejudice substantially outweighs the probative value of the relevant evidence. If the judge determines that any of the examples of unfair prejudice mentioned in Rule 403 exists, and that they substantially outweigh the probative value of the evidence, the judge must exclude the evidence, and the jury never sees it. The 1972 Advisory Committee Note to Rule 403 provides examples of the circumstances where such unfair prejudice may exist:

> The case law recognizes that certain circumstances call for the exclusion of evidence which is of unquestioned relevance. These circumstances entail risks which range all the way from inducing decision on a purely emotional basis, at one extreme, to nothing more harmful than merely wasting time, at the other extreme. Situations in this area call for balancing the probative value of and need for the evidence against the harm likely to result from its admission . . . .
>
> . . . "Unfair prejudice" within . . . [the] context [of Rule 403] means an undue tendency to suggest decision on an improper basis, commonly, though not necessarily, an emotional one.[173]

It is important to keep in mind that as the probative value of the challenged evidence increases, so too must the demonstration that the prejudice of admitting it *substantially* outweighs its probative value for it to be excluded under Rule 403.[174] In essence, the balancing under Rule 403 tilts in favor of admissibility, not exclusion.[175] Since all evidence offered by one party against the other in a civil or criminal case is to some degree "prejudicial" to that party, more than just routine prejudice must be shown.[176] There must be "unfair" prejudice. And the weight of the prejudice also must be substantially greater than the probative weight of the evidence that is being challenged.[177] As we will show, the structure of Rule 403 becomes very important when questions about the admissibility of alleged deepfake evidence are considered.[178]

---

[173] FED. R. EVID. 403 advisory committee's note to 1972 proposed rules (citations omitted).

[174] *See* Sprint/United Mgmt. Co. v. Mendelsohn, 552 U.S. 379, 387 (2008) (holding that "prejudice . . . under Rule[] 403 [is] determined in the context of the facts and arguments in a particular case").

[175] *E.g.*, GN Netcom, Inc. v. Plantronics, Inc., 930 F.3d 76, 85 (3d Cir. 2019) ("[T]here is a strong presumption that relevant evidence should be admitted, and thus for exclusion under Rule 403 to be justified, the probative value of evidence must be 'substantially outweighed' by the problems in admitting it.").

[176] *See, e.g.*, United States v. DiRosa, 761 F.3d 144, 153 (1st Cir. 2014) ("We stress that it is only unfair prejudice which must be avoided . . . because [b]y design, all evidence is meant to be prejudicial." (internal citations omitted)).

[177] FED. R. EVID. 403.

[178] FED. R. EVID. 403 is not the only balancing test contained in the Federal Rules of Evidence. Several other rules address circumstances in which the judge must decide whether evidence that has "passed" the relevance threshold should nonetheless be excluded from consideration by the jury. Like FED. R. EVID. 403, some of these balancing tests favor admissibility, but others are intended to exclude the evidence, unless its probative value greatly outweighs its potential prejudice.

C. Authenticity

Non-testimonial evidence (such as written text (whether electronic or "hard copy"), photographs/audios/videos, and tangible things) must be "authentic" in order to be relevant. Three rules establish when such evidence meets the authenticity test, the most important of which is Rule 901, which we discuss here. As argued in this Article, the authenticity test is the most important one to consider when the evidence is AI-generated, as well as when one party offers image, audio, or audiovisual evidence that their opponent challenges as a deepfake.

Rule 901(a) establishes the test for authenticity: "To satisfy the requirement of authenticating or identifying an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims

---

For example, FED. R. EVID. 609(a)(1)(A), which deals with whether a witness who testifies in a criminal case may be impeached by their prior felony conviction, adopts the FED. R. EVID. 403 balancing test for all witnesses who testify in civil or criminal cases other than the defendant. FED. R. EVID. 609(a)(1)(B), on the other hand, states that if the defendant testifies in a criminal case, they may only be impeached by a prior felony conviction if "the probative value of the evidence outweighs its prejudicial effect to that defendant"—making it easier for the defendant to exclude the evidence of a prior felony conviction if the judge determines that it is even slightly more prejudicial than probative.

FED. R. EVID. 609(b)(1) excludes the use of a conviction if more than 10 years have passed since the witness's conviction or release from confinement, unless the judge determines that the probative value of the evidence "supported by specific facts and circumstances, substantially outweighs its prejudicial effect," which leans strongly against admissibility of the dated conviction. This same test was adopted in FED. R. EVID. 412(b)(2), which states that "[i]n a civil case, the court may admit evidence proffered to prove a victim's sexual behavior or sexual predisposition if its probative value substantially outweighs the danger of harm to any victim and of unfair prejudice to any party." Similarly, FED. R. EVID. 703 discusses when an expert witness (who will testify about scientific, technical, or specialized evidence) may consider facts that ordinarily would be inadmissible. The facts may be disclosed to the jury when other experts in the field reasonably would find these facts reliable in reaching an opinion about the evidence in a case. It states that "if the facts or data [considered by the expert in reaching their opinion] would otherwise be inadmissible, the proponent of the opinion may disclose them to the jury only if their probative value in helping the jury evaluate the opinion substantially outweighs their prejudicial effect."

Thus, the Federal Rules of Evidence have adopted a variety of tests that the trial judge must apply when deciding whether to admit relevant evidence that also may be prejudicial. These tests fall on a continuum, ranging from strongly favoring admissibility (*e.g.*, FED. R. EVID. 403 and 609(a)(1)(A)), to strongly disfavoring admissibility (*e.g.*, FED. R. EVID. 609(b)(1) and 703), to slightly favoring admissibility, provided the evidence is at least more probative than prejudicial, but excluding it if it is even slightly more prejudicial than probative in the middle (*e.g.*, FED. R. EVID. 609(a)(1)(B)).

These rules reflect the judgment of the Advisory Committee on the Evidence Rules that there may be special attributes or concerns about certain types of evidence that justify making it either easier or more difficult to admit. This is a policy decision. It is important, because, as we argue below, there are strong policy reasons that justify using a test similar to the one applied in FED. R. EVID. 609(a)(1)(B) when evidence is challenged as a deepfake (making it easier to exclude if the judge determines it is more prejudicial than probative—even if only slightly so), as opposed to the more stringent test in FED. R. EVID. 403, which makes it much more difficult for the judge to exclude evidence even when it may be quite prejudicial.

it is."[179] As the Advisory Committee Note to Rule 901(a) states, "Authentication and identification represent a special aspect of relevancy."[180]

The means by which the proponent can meet their authentication obligation are set forth in a list of non-exclusive examples in Rule 901(b) and 902. For example, as we mentioned above, if the evidence offered is a voicemail recording of the defendant's voice, the plaintiff can call a witness with personal knowledge of the sound of the defendant's voice to testify that they listened to the voicemail, and it is the defendant.[181] Or, if the plaintiff is familiar with the defendant's manner of speaking, they may testify as to their opinion that it is the defendant's voice on the voicemail.[182] Or the plaintiff can introduce examples of the defendant's voice as to which there is no dispute that they were generated using AI, and the jury can listen to them, compare them to the voice on the voicemail, and decide for themselves whether or not it is the voice of the defendant.[183]

The concept is very straightforward, but an example helps make the key points. Suppose the plaintiff contends that they have a voicemail message from the defendant threatening them. To authenticate it, they must introduce sufficient evidence for the judge (as gatekeeper) and the jury (as fact finder) to conclude that the voice is that of the defendant.[184] In this regard, there is a definite link between the requirement of authenticity in Rule 901(a), and the "conditional relevance" concept that underlies Rule 104(b).[185] The proponent of the evidence must meet the authenticity requirement by a preponderance of the evidence—that it is more likely authentic than not.[186]

What this means is that when the facts the proponent of the evidence wants to rely on to establish authentication are disputed by the opposing party (i.e., plaintiff says the voice on the voicemail is the defendant's; defendant denies it is their voice), then both the judge and jury are involved in determining whether the evidence should be admitted. Initially, the judge, as gatekeeper, must determine whether a reasonable jury could conclude by a preponderance of evidence that the voice on the voicemail message is that of the defendant. If the judge concludes this showing has not been made by the proponent, then the judge excludes the evidence from

---

[179] FED. R. EVID. 901(a).

[180] FED. R. EVID. 901(a) Advisory Committee's Note to 1972 amendment (citation omitted).

[181] FED. R. EVID. 901(b)(1) (a witness with personal knowledge of the defendant's voice).

[182] FED. R. EVID. 901(b)(5) (opinion as to voice).

[183] *See* FED. R. EVID. 901(b)(3) (factfinder comparison of known and disputed examples); FED. R. EVID. 901(b)(4) (distinctive characteristics of defendant's voice match those of the voicemail).

[184] *See* FED. R. EVID. 901(a).

[185] FED. R. EVID. 901(a) Advisory Committee's Note to 1972 amendment ("[The] requirement of showing authenticity or identify falls in the category of relevancy dependent upon fulfillment of a condition of fact and is governed by the procedure set forth in Rule 104(b)").

[186] Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 542 (D. Md. 2007); Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROPERTY 9, 88 n.348 (2021) ("The party introducing the evidence bears the burden of proving that the offered evidence meets the requirements of 104(a) by a preponderance of the evidence."). This is a relatively low threshold (i.e., 51%).

consideration by the jury. But if the judge determines that a reasonable jury could find that the voice is more likely than not the voice of the defendant, then the judge submits the disputed facts to the jury to consider. If they agree with the proponent that it is more likely the defendant's voice than not, then they may consider the evidence, and give it the weight they think it deserves. But if they conclude that the defendant's facts disputing that the voice is theirs are more likely true than not, then they are instructed by the judge to disregard the voicemail and give it no weight in their deliberations.

This concept is critical in considering how evidence challenged as being a deepfake should be evaluated to determine if it is admissible. Sticking with our voicemail example, the plaintiff contends that the challenged evidence is a voicemail message that the defendant left on their voicemail. But the defendant denies that it is their voice, insisting instead that it is a deepfake created by the plaintiff (or someone associated with them). The defendant claims that someone used a generative AI application that allows the creation of a cloned (i.e., fake) recording of the voice of the defendant by inputting an authentic sample of the defendant's voice (perhaps taken from a YouTube post the defendant made) into an online application that allows the plaintiff to type the text of the threatening message, then produces a recording using the sample recording of the defendant's voice, but instead speaking the content that the plaintiff typed. The finished product is a cloned audio recording in the defendant's voice making a threat they never made. How do the rules of evidence handle the admissibility of such a voicemail?

The proponent of the evidence (here, the plaintiff) goes first, as they must meet the threshold requirement of authenticating the voicemail. We have already described several ways the plaintiff might easily do so. The judge first considers the proffered evidence to see if a reasonable jury could conclude that the voicemail more likely than not was left by the defendant. If the judge concludes "no," the jury does not hear the evidence—it is excluded for want of authenticity. But if the judge concludes the jury *could* find by a preponderance that it is the defendant's actual voice, or that the jury reasonably could conclude either that it might or might not be the defendant's actual voice, then the judge does not make the final call on admissibility but submits the disputed facts to the jury to consider and decide one way or the other.

But what should the judge do if they determine that the dispute as to authenticity of the voicemail could be decided either way by the jury, but that the content of the email is so graphic, shocking, and extreme that if the jury hears it, they might rule based on their emotional reaction to the nature of the voicemail, even though there is evidence that the message may be fake? Are they required to let the jury hear it in order to resolve the disputed facts as to its authenticity (even though the judge is concerned that hearing it may cause the defendant unfair prejudice in the eyes of the jury), or may the judge first evaluate the potential prejudicial impact of hearing the voicemail and, if convinced that its probative value is substantially outweighed by the danger of unfair prejudice, exclude it under Rule 403? While the rules themselves do not provide a clear answer to this question, there is analogous case law that provides very helpful guidance as to what the judge's options are.

In *Huddleston v. United States*,[187] the Supreme Court considered the evidence rules that govern the admissibility of evidence of "other crimes, wrongs or acts" offered under Rule 404(b).[188] This evidence is often offered in criminal cases and is usually objected to by the defendant as irrelevant or, if relevant, unfairly prejudicial.[189] The issue before the Court was whether the trial judge was required to make an initial finding that the government had proved the existence of the prior act by a preponderance of the evidence, before allowing the jury to hear it.[190] The Court concluded the trial judge did not personally have to make this initial determination, but only to determine whether the government had offered sufficient evidence for the jury to find by a preponderance of the evidence that the televisions were stolen.[191] Having done so, the Court considered the defendant's argument that even if that showing had been made, the evidence nevertheless should be excluded as excessively prejudicial under Rule 403.[192] The Supreme Court held:

> [W]e share the petitioner's concern that unduly prejudicial evidence might be introduced under Rule 404(b). We think however, that the protection against such unfair prejudice emanates not from the requirement of a preliminary finding by the trial court, but rather from . . . other sources: . . . from the requirement of Rule 404(b) that the evidence be offered for a proper purpose; second, from the relevancy requirement of Rule 402—as enforced through Rule 104(b) [the "conditional relevance rule"]; [and] . . . from the assessment the trial court must make under Rule 403 to determine whether the probative value of the similar acts evidence is substantially outweighed by its potential for unfair prejudice . . . .[193]

Similarly, in *Johnson v. Elk Lake School District*,[194] the Third Circuit followed the approach used in *Huddleston* when deciding (in a civil case alleging sexual assault), whether evidence that the defendant had committed a prior offense that constituted a "sexual assault," pursuant to Rule 415, should be admitted.[195] The Court again determined that the trial judge needed only to make the initial determination that a reasonable jury could find that the prior conduct was a sexual assault, not make that finding themself, before submitting it to the jury to consider

---

[187] Huddleston v. United States, 108 S. Ct. 1496 (1988).

[188] FED. R. EVID. 404(b) is part of the character evidence rules. It prohibits introduction of evidence of "other crimes, wrongs, or acts" to prove propensity, stating, "Evidence of any other crime, wrong, or act is not admissible to prove a person's character in order to show that on a particular occasion the person acted in accordance with the character." FED. R. EVID. 404(b)(1).

[189] *See, e.g.*, Steven Goode, *It's Time to Put Character Back into the Character Evidence Rule*, 104 MARQ. L. REV. 709, 713-154 (2021) (discussing a case where evidence of defendant's association with prior disappearances was admitted despite Rule 404(b) objections).

[190] The "prior act" at issue in *Huddleston* was whether the televisions sold by the defendant had been stolen. 108 S. Ct. at 1499.

[191] *Id.* at 1501.

[192] *Id.* at 1500.

[193] *Id.* at 1502 (emphasis added) (internal citations omitted).

[194] Johnson v. Elk Lake School Dist., 283 F. 3d 138 (3d. Cir. 2002).

[195] *Id.* at 143-44.

under Rule 104(b).[196] But it qualified this ruling, adding, "We also conclude, however, that even when the evidence of a past sexual offense is relevant, the trial court retains discretion to exclude it under Federal Rule of Evidence 403 if the evidence's probative value 'is substantially outweighed by the danger of unfair prejudice . . . .'"[197]

Returning to our prior example of the disputed authenticity of the voicemail, *Huddleston* and *Johnson* offer important guidance. The proponent has met their initial authentication obligation to show by a preponderance of the evidence that the defendant's voice is that heard on the voicemail (e.g., by the plaintiff's opinion testimony, under Rule 901(b)(5), that the voice is the defendant's). This is enough for the judge to find that a reasonable jury could agree with the plaintiff that the voicemail was left by the defendant. The trial judge does not have decide whether they agree with this, only that the jury could. Ordinarily, the judge would then admit the plaintiff's evidence supporting authenticity as well as the defendant's evidence challenging authenticity for the jury to decide which version they accepted—under Rule 104(b). But, before doing so, *Huddleston* and *Johnson* permit the judge to make an assessment under Rule 403 on whether allowing the jury to hear the disputed evidence would create an unacceptable likelihood of unfair prejudice to the defendant (if, for example, the defendant has introduced evidence from which a reasonable jury could find that the voicemail was the product of a sophisticated generative AI application, and that the jury would not be able to put this out of their mind even if they suspected that the recording was a deepfake). Thus, while the current rules of evidence regarding relevance and prejudice do not provide a perfect solution for the "deepfake dilemma," they can—if properly applied—help to mitigate that dilemma. As we will soon argue, however, it would be very helpful to have new evidence rules that address this unique problem directly.

D. Scientific, Technical, and Specialized Evidence

The final evidence rule that needs to be considered when evaluating the admissibility of acknowledged AI-generated evidence and possible deepfake evidence is Rule 702, which governs admissibility of scientific, technical, and specialized evidence, and expert opinion testimony.[198] In this regard, Rule 702 does not directly address AI-generated evidence but provides a very helpful framework to "borrow from" in the context of acknowledged AI-generated and potential deepfake evidence.[199]

---

[196] *Id.* at 154-55.

[197] *Id.* at 144.

[198] FED. R. EVID. 702.

[199] See, e.g., FED. R. EVID. 102, which states that the rules of evidence "should be construed so as to administer every proceeding fairly, eliminate unjustifiable expense and delay, and promote the development of evidence law, to the end of ascertaining the truth and securing a just determination." This rule—which applies to the consideration of all the other rules of evidence—permits the "borrowing" of analysis helpful in one evidentiary context when considering a new evidentiary problem not otherwise explicitly addressed by the evidence rules. *See, e.g.*, G. Alexander Nunn, *The Living Rules of Evidence* 170 U. PENN. L. REV. 937, 982 (2022).

Rule 702 was revised effective December 1, 2023. It now states:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise, *if the proponent demonstrates to the court that it is more likely than not that*:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on *sufficient facts or data*;
>
> (c) the testimony is the product of *reliable principles and methods*; and
>
> (d) the expert's opinion reflects a *reliable application of the principles and methods to the facts of the case*.[200]

The rule was amended to strengthen the requirement that the trial judge fulfill their gatekeeping role under Rule 104(a) to ensure that evidence regarding scientific, technical, and specialized matters should not be admitted for consideration by a jury unless the proponent has met its initial burden of showing the four factors in the rule, by a preponderance of the evidence.[201] While it does not directly address admissibility of acknowledged AI-generated evidence or evidence asserted to have been created by an AI application (such as a deepfake), Rule 702 is nevertheless invaluable when analyzing this evidence for the obvious reason that AI-generated evidence is, by definition, scientific, technical, and specialized.[202] When the judge uses the four factors set forth in Rule 702 to evaluate the admissibility of scientific, technical, or specialized evidence, the judge usually considers the *Daubert* factors.[203]

The authors of this Article have previously argued that the authentication of AI-generated evidence should align with established evidentiary standards, as its admissibility depends on demonstrating validity, reliability, and adherence to scientific principles, making the application of Rule 702 and the *Daubert* factors particularly useful:

> The usefulness of borrowing [from Rule 702 and the *Daubert* factors] in assessing whether acknowledged AI-generated evidence should be admitted is readily apparent. To authenticate AI technology, its proponent must show that it produces accurate, that is to say valid, results. And it must perform reliably, meaning that it

---

[200] FED. R. EVID. 702 (emphasis added).

[201] FED. R. EVID. 702 advisory committee's note to 2023 amendment.

[202] *See* Grimm et al., *supra* note 186, at 95 (arguing that the best way to assess whether AI evidence is sufficiently accurate or reliable "is to employ Rule 102, which requires the rules of evidence to be 'construed so as to administer every proceeding fairly . . . and promote the development of evidence law' to 'borrow' from Rule 702 and the cases that have interpreted it, when determining the standard for admitting scientific, technical, or other specialized information that is beyond the understanding of lay jurors and generalist judges.")

[203] FED. R. EVID. 702 Advisory Committee's Note to 2000 amendment.

> consistently produces accurate results when applied in similar circumstances. When the accuracy and reliability of technical evidence has been verified through independent testing and evaluation of the AI system that produced it, [when] the methodology used to develop the evidence has been published and subject to peer review by others in the same field of science or technology, when the error rate associated with the AI system['s] use is not unacceptably high, when the standard testing [and operational] methods and protocols have been followed, and when the methodology used is generally accepted within the field of similar scientists or technologists, then [the AI-generated evidence] has been authenticated. It does what its proponents say it does. And introducing it produces none of the adverse consequences that Rule 403 is designed to guard against.[204]

With Rule 702 and the *Daubert* factors in mind, one authentication rule in particular—Rule 901(b)(9)—seems especially helpful in evaluating the admissibility of acknowledged AI-generated evidence as well as evidence asserted to be an AI-generated deepfake. Rule 901(b)(9) provides that a party can authenticate evidence if it demonstrates that it was created by a "process or system and showing that it produces an accurate result."[205] Rule 901(b)(9) uses the word "accurate" to determine whether the evidence is authentic,[206] but accuracy is essential to, but not necessarily sufficient, to show that AI-generated evidence should be admitted. After all, as we previously observed, a broken watch "accurately" tells the time twice a day, but it is not a reliable way of doing so throughout the day. As the preceding quote explains, the best way to assess the "accuracy" of AI evidence is to consider the related concepts of "validity" and "reliability," which is what is done in the field of science.[207]

In short, the best way under the current rules of evidence to address admissibility of acknowledged AI-generated evidence and evidence asserted to be the product of an AI application, such as a deepfake, is to focus on the authenticity of the evidence, borrowing heavily from the analysis of Rule 702[208] and the

---

[204] Grimm et al., *supra* note 186, at 96.

[205] FED. R. EVID. 901(b)(9).

[206] *Id.*

[207] Grimm et al., *supra* note 186, at 48 ("*Validity* is the quality of being correct or true, in other words, whether and how accurately an AI system measures (i.e., classifies or predicts) what it is intended to measure. *Reliabilit*y refers to the *consistency* of the output of an AI system; that is, whether the same (or a highly correlated) result is obtained under the same set of circumstances. Both need to be measured and both need to exist for an AI system to be trustworthy." (italics in original) (internal citations omitted)); *see also* discussion *supra* Part III and sources cited notes 70-71 (defining validity and reliability).

[208] One proposal that the Advisory Committee on Federal Evidence Rules is considering is a new rule, FED. R. EVID. 707, that would explicitly recognize the value of considering the FED. R. EVID. 702 factors when assessing machine-generated evidence, such as that generated by AI. Daniel J. Capra, Memorandum to Advisory Committee on Evidence Rules (Oct. 1, 2024), *in* ADMIN. OFF.

Daubert factors, bearing in mind that, with respect to deepfakes in particular, Rule 403's assessment of the potential for unfair prejudice to occur if the jury is exposed to certain types of evidence also is an essential factor to consider.

While we concede that the bench and bar can likely "make due" with the current rules of evidence if needs be, there would be considerable benefit to enacting two new evidence rules: one that supplements Rule 901(b)(9) to address acknowledged AI-generated evidence (establishing an already accepted way to authenticate scientific, technical, and specialized evidence, and encouraging counsel and the courts to employ it) and another to deal with the unique problem that unacknowledged AI-generated or deepfake evidence presents.

VII. PROPOSED CHANGES TO FEDERAL RULES OF EVIDENCE TO ADDRESS ACKNOWLEDGED AI-GENERATED EVIDENCE AND EVIDENCE CHALLENGED AS DEEPFAKE

The rules of evidence largely are "technology neutral," meaning that they are intended to apply broadly to all types of evidence. There is wisdom in this, as technology—especially AI technology—changes constantly and extremely quickly, much faster than the ability of the Advisory Committee on the Rules of Evidence to keep up with the pace of change with bespoke new rules addressing specific types of technological evidence. This is because the adoption of new rules of practice and procedure in the U.S. Courts is governed by the Rules Enabling Act.[209] In brief, the Rules Enabling Act ensures that the process of adopting new rules of practice and procedure is transparent to the public at large as well as attorneys and clients, and that there are layers of review that apply to any rule change—starting with the Federal Evidence Rules Committee, the Standing Committee on Rules of Practice and Procedure, the Judicial Conference of the United States, the Supreme Court, and finally, Congress.[210] From start to finish, the process of enacting a new rule of evidence can take as long as four or more years.[211] But regardless of the time it may take to adopt new rules, there are times when the process needs to be undertaken when particular types of evidence create new challenges for judges and lawyers, and the existing rules are not adequate for the task.

We believe that there is a need to adopt two new rules—one to address acknowledged AI-generated evidence, and another to deal with the unique problems presented by unacknowledged AI-generated evidence or potential

---

OF THE U.S. CTS, ADVISORY COMMITTEE ON EVIDENCE RULES AGENDA BOOK 220, 252 (Nov. 8, 2024). As currently proposed, new FED. R. EVID. 707 would say, "Where the output of a process or system would be subject to Rule 702 if testified to by a human witness, the court must find that the output satisfies the requirements of Rule 702(a)-(d). This rule does not apply to the output of basic scientific instruments or routinely relied upon commercial software." *Id.* at 253.

[209] 28 U.S.C. § 2071-2077.

[210] *See id.* §§ 2073-2074.

[211] *See* Nate Raymond, *US Judicial Panel to Develop Rules to Address AI-produced Evidence*, REUTERS (Nov. 8, 2024, 4:38 PM), https://www.reuters.com/legal/transactional/us-judicial-panel-develop-rules-address-ai-produced-evidence-2024-11-08/ (citing Judge Jesse Furman, Chair of the Advisory Committee on Evidence Rules).

deepfakes. The authors have drafted these proposed rules, discussed below, and submitted them to the Advisory Committee on Evidence Rules, which has considered them, but has decided not to move forward with them as drafted at the present time.[212]

Why do we think there is a need for these particular proposed rules at this time? First, as we show below, there is a "catch 22" in the current rules that makes addressing challenges to evidence that the non-proffering party claims is a deepfake especially problematic. And second, while the existing authentication rules—especially Rule 901(b)(9)—are useful for authenticating acknowledged AI-generated evidence, it would be extremely beneficial to have a rule that employs more precise language and focuses on an acceptable way to obtain the admission of such evidence, that provides a "recipe" lawyers can follow when preparing for trials and hearings, and that judges can refer to in ruling on evidentiary challenges to acknowledged AI-generated evidence. We briefly discuss each of these reasons, followed by our proposed new rules and the justification for them.

A. The Deepfake Dilemma

As we have already discussed, the rules of evidence allocate different roles to the trial judge and jury with respect to the admissibility of evidence, and the weight given to admitted evidence in resolving disputes at trial. Under Rule 104(a), the trial judge is the "gatekeeper," making *preliminary* assessments on the admissibility of evidence, qualification of witnesses, and the existence of evidentiary privileges.[213] Under Rule 104(b), the jury is charged with resolving disputed facts when the relevance of evidence one party offers is challenged by the opposing party,[214] and for determining how much weight to give to any evidence that is admitted, or under Rule 104(e), how credible any witness is.[215] As we also have pointed out, the most frequent challenge raised to the admissibility of non-testimonial evidence—such as AI-generated and other digital evidence—is authentication under Rule 901(a) (i.e., whether the evidence "is what the offering party claims it to be").[216]

The example we used above in Part VI helps illustrate this point. Party A offers into evidence a voicemail recording which it claims Party B left on their cellphone. Party B objects, and claims that the voicemail is actually a deepfake. As the proponent, Party A must come forward with sufficient facts to convince the trial judge that the jury reasonably could find that the voice in the voicemail message actually is A's voice. In response, Party B has some options, ranging from objecting to the voicemail as a deepfake in conclusory or hypothetical terms ("Your Honor, we all know how easy it is to make a fake voicemail, how do we know that this one isn't a deepfake?") to offering evidence to show the voicemail is a fake (for example,

---

[212] *See id.* (agreeing to develop a potential rule).

[213] FED. R. EVID. 104(a).

[214] FED. R. EVID. 104(b).

[215] FED. R. EVID. 104(a), (e)

[216] *See* FED. R. EVID. 901(a); discussion *supra* Section VI.C;.

a forensic expert who has examined the voicemail and found indicia that it was fabricated by a generative AI application). When Party A has offered sufficient evidence that a jury *could* find the voicemail is authentic, and Party B has offered sufficient evidence from which a reasonable jury *could* find that it is fake, Rule 104(b) requires the judge to let the jury hear both versions and to make its own assessment as to whether the voicemail is real or fake.[217] But this means the jury will be exposed to the voicemail, and there is a real danger that even if the jury is persuaded that the voicemail is—or likely is—fake, they will have great difficulty putting it out of their minds when they deliberate.[218]

We have noted that Rule 403 allows the court to balance the potential probative value of the evidence against its possible unfair prejudice, but the introductory language to Rule 403 qualifies its use as follows: "The court may exclude *relevant* evidence if its probative value is substantially outweighed by a danger of . . . unfair prejudice."[219] Evidence is not relevant if it is not authentic.[220] Relevance is now conditional—and the question goes to the jury—if the evidence offered to authenticate the evidence is opposed by bona-fide evidence that it is not authentic; the judge must then allow the jury to resolve the disputed facts.[221] If the jury agrees it more likely is authentic, they can consider it. If they agree it is more likely not authentic, they are told by the judge to disregard it. But the "catch 22" with respect to our deepfake hypothetical is that the jury must listen to the challenged evidence in order to make its determination of authenticity. And we have pointed to several studies that show that with respect to audio, visual, or audiovisual evidence challenged as a deepfake, the very fact of hearing or seeing it may irreparably prejudice the jurors even if they determine it likely is fake.[222]

We have argued that the *Huddleston* and *Johnson* cases could support an argument that under these circumstances, the judge should first consider prejudice under Rule 403 in determining whether to let the jury hear the disputed evidence of authenticity.[223] But those two cases dealt with different types of evidence than unacknowledged AI-generated evidence, and a court could find that they are inapplicable to the deepfake situation. But even if our analysis is accepted, recall

---

[217] *See* FED. R. EVID. 104(b); discussion *supra* Section VI.C.

[218] *See* discussion *supra* Section VI.C.

[219] FED. R. EVID. 403 (emphasis added).

[220] *See* FED. R. EVID. 901.

[221] See *supra* Section VI.A. (discussing conditional relevance).

[222] *See, e.g.*, discussion *supra* Section IV; Myhand, *supra* note 88, at 174-75 ("The dangerousness of deepfake videos lie in the incomparable impact these videos have on human perception. Videos are not merely illustrative of a witnesses' testimony, but often serve as independent sources of substantive information for the trier of fact. Since people tend to believe what they see, images and other forms of digital media are often accepted at face value . . . . Video evidence is more cognitively and emotionally arousing to the trier of fact, giving the impression that they are observing activity or events more directly." (internal quotation omitted)). Myhand's concern regarding deepfake videos also extends to fake audio evidence. If "seeing is believing," so too is hearing.

[223] *See* discussion *supra* Section VI.C.

that Rule 403 tilts strongly in favor of admissibility,[224] allowing the judge to exclude the evidence only if its probative value is substantially outweighed by the danger of *unfair* prejudice. Since potential deepfakes will often involve highly relevant evidence (such as in our defamation case with the disputed voicemail recording), then the showing of prejudice needed to exclude them must correspondingly be much greater. This may prove to be too great a showing for the opposing party to make, rendering Rule 403 virtually ineffective as it applies to deepfake challenges. For that reason, we have proposed a new rule with a familiar balancing test, to deal with the unique challenge posed by deepfakes.

The second rule that we propose is more straightforward. We have noted that Rule 901(b)(9) allows a party to authenticate evidence by showing that it was produced by a system or process that produces an accurate result.[225] This rule is technology neutral, but could be helpfully amended to provide greater specificity by editing a few words and by including a new subsection that deals specifically with how acknowledged AI-generated evidence could be shown to be the product of a system or process that produces a "valid and reliable" result: that is, by describing the training data and software or program that was used, and showing that it produced valid and reliable results in the particular case in which it is offered. This rule change amounts to a modest amendment of a longstanding rule. Its benefits are (i) that it uses more precise, scientific language, and (ii) it describes an acceptable method to authenticate acknowledged AI-generated evidence for those who wish to use it. By doing so, the proponent will have a rule to rely on in arguing that they have met their authentication obligation, and lawyers and judges will have a clear roadmap for what is sufficient to authenticate AI-generated evidence, which should encourage its use and avoid uncertainty.

Here are the two rules that the authors have proposed:

1. *Proposed New Rule 901(c)*

> **901(c) *Potentially Fabricated or Altered Electronic Evidence*. If a party challenging the authenticity of computer-generated or other electronic evidence demonstrates to the court that a jury reasonably could find that the evidence has been altered or fabricated, in whole or in part, using artificial intelligence, the evidence is admissible only if the proponent demonstrates that its probative value outweighs its prejudicial effect on the party challenging the evidence.**

This rule has the following advantages: (i) It limits the scope to evidence challenged as having been fabricated by AI, thereby narrowing its scope and preventing it from being used to challenge *any other* kind of digital evidence, preventing the danger of "flooding" the courts with objections to run-of-the-mill digital evidence; (ii) It substitutes the words "valid and reliable" for the current rule's use of the term "accurate," to better and more accurately capture what needs

[224] *See* discussion *supra* Section VI.B.

[225] *See* discussion *supra* Section VI.D.

to be addressed when assessing the authenticity of AI evidence; (iii) It preserves the complete autonomy of the party that is offering the evidence to choose any means they wish to authenticate the evidence; (iv) It places the burden on the party objecting to the evidence as AI-generated or deepfake to come forward with facts—not conclusory argument or hypothetical possibilities—sufficient to persuade a reasonable jury by a preponderance of the evidence that the challenged evidence is a AI-generated or fake; (v) It does not require the trial judge to decide whether the objecting party's evidence in fact is sufficient to show that the evidence is AI-generated or a deepfake, only to decide whether a reasonable jury could make this determination. This is in keeping with the accepted approach taken in *Huddleston* and *Johnson*; (vi) It adopts a familiar but different balancing test than the one in Rule 403, which tilts too far towards admissibility in this instance, instead using the balancing test already found in Rule 609(a)(1)(B), which requires that the proponent show that the probative value of the evidence outweighs its prejudicial impact, which is a more appropriate balancing test given the particular challenges associated with AI-generated or deepfake evidence; and (vii) It explicitly allows the judge to employ the balancing test before the jury is exposed to the potentially highly prejudicial evidence.

2. Proposed New Rule 901(b)(9)

The proposed new language is shown in bold font, the existing language is in regular font, and deleted language is shown with strikethrough:

> [901] (b) Examples. The following are examples only—not a complete list—of evidence that satisfies the requirement [of Rule 901(a)]:
>
> (9) Evidence about a Process or System. For an item generated by a process or system:
>
> (A) evidence describing it and showing that it produces ~~an accurate~~ **a valid and reliable** result; and
>
> **(B) if the proponent acknowledges that the item was generated using artificial intelligence, additional evidence that:**
>
> **(i) describes the training data and software or program that was used; and**
>
> **(ii) shows that they produced valid and reliable results in this instance.**

This proposed rule has the following advantages: (i) It is a minor adjustment to a familiar and a well-established rule of authentication; (ii) It replaces the less specific and less appropriate word "accurate" with the more specific and appropriate phrase "valid and reliable," consistent with language used in the scientific community and with the standards in Rule 702 for evaluating the reliability of scientific, technical, and specialized evidence, which is especially appropriate when evaluating acknowledged AI-generated evidence; (iii) It does not mandate a particular method of authenticating AI-generated evidence, but rather

provides an accepted example, which is consistent with the existing structure of Rule 901(b), and preserves the autonomy of the party offering AI-generated evidence to choose any method desired to authenticate the evidence; and (iv) It provides a method of authentication that, if used, is recognized as sufficient, thereby providing an incentive for this rule to be applied, and gives certainty to lawyers who use it that they will succeed in authenticating their AI-generated evidence. It also lessens the likelihood of a challenge to AI-generated evidence when the rule is properly satisfied and gives a definitive standard to a judge who must decide a challenge to authentication of AI-generated evidence.

As we noted, the Advisory Committee on Evidence Rules met on November 8, 2024, to consider a number of potential changes to the evidence rules, including whether to go forward with proposed new rules addressing acknowledged AI-generated evidence and potential deepfakes.[226] We provided the Committee with our two proposed rules, and a memo explaining why we thought that they should be adopted by the Committee.[227] At the meeting, the Committee determined that it would develop a rule addressing the introduction of AI evidence, and to begin closer consideration of a possible approach that could be used in the future to assist judges in dealing with evidence claimed to be deepfake.[228] A memo from Fordham Law School Professor Daniel J. Capra, the longtime and highly respected reporter for the Committee, dated October 1, 2024, described what such a rule might look like, if approved by the Committee:

> If a party challenging the authenticity of computer-generated or other electronic evidence demonstrates to the court that a jury reasonably could find that the evidence has been altered or fabricated, in whole or in part, by artificial intelligence [by an automated system], the evidence is admissible *only if the proponent demonstrates to the court that it is more likely than not authentic*.[229]

Professor Capra explained this draft rule as follows:

> This burden-shifting alternative on the question of authenticity—once the opponent has made a prima facie case, the proponent has to establish authenticity more likely than not—may be questioned because it imports a Rule 104(a) standard for an authenticity question, while all other authenticity questions are decided under Rule 104(b). But this differentiation may be justified by the problems inherent in detecting deepfakes. And heightening the standard makes sense after the opponent has provided a prima facie case of fakery. After that triggering requirement is met, the proponent should have to show *something* more than the Rule 104(b) standard of authenticity. The logical conclusion is the

[226] ADMIN. OFF. OF THE U.S. CTS, *supra* note 208, at 40.

[227] *See* Capra, *supra* note 208, at 240-45.

[228] Raymond, *supra* note 211.

[229] Capra, *supra* note 208, at 250 (emphasis added).

> proponent must show authenticity by a preponderance of evidence. Note that the Rule 104(a) standard only applies if the opponent makes the initial showing of fakery. If that showing is not made, then the proponent authenticates under the Rule 104(b) standard.[230]

It is important to keep in mind that the Committee did not decide that a new rule would be adopted or what the proposed new "deepfake rule" would say, only that it would be helpful to move forward with drafting such a rule in the event that the Committee decides that one should be adopted. Accordingly, we can expect that in connection with their May 2, 2025 meeting.[231] the Committee will further consider this, and, of course, any proposed rule would be subject to public notice and comment, and have to be approved, as required by the Rules Enabling Act, a process that could take several years. Nevertheless, it seems that the Committee may have crossed the Rubicon with respect to its position on whether it is advisable to have a bespoke rule addressing potential deepfake evidence. That is a very significant and important step forward.

VIII. PRACTICE POINTERS FOR COURTS UNDER THE EXISTING RULES OF EVIDENCE

We close this Article with some suggestions about what lawyers and courts might do to deal with acknowledged and unacknowledged AI-generated evidence *now*, since any rules change is likely years away, and there is no developed case law at present to lend a hand.

A. Early Anticipation and Planning.

---

[230] *Id.* at 250 (emphasis in original). This version of the new rule begins with the language proposed by the authors of this article, but modifies the showing that must be made, for the reasons described by Professor Capra in his memo. The authors agree that this potential rule change would be a step forward in addressing the "deepfake dilemma" we have described, but it may not go far enough. While a critique of Professor Capra's proposal is beyond the scope of this paper, we will note that when the party opposing evidence as AI-generated has met its burden of coming forward with evidence from which the jury could reasonably find that the evidence has been fabricated or altered using AI, Professor Capra's proposed rule only requires the proponent to demonstrate to the court that the evidence is more likely than not authentic. It leaves to the judge how best to make that determination. Our proposal, instead, provides the judge with a familiar balancing test (drawn from FED. R. EVID. 609(a)(1)(B)) that they can apply to assess the potential impact of the evidence. As we have explained, *supra* Part IV, research has shown that non-expert humans—including judges—are not adept at distinguishing authentic from synthetic content, and that experts may not always be available to testify. For that reason, we believe that the balancing test that we propose provides the judge with a more workable standard. If the judge finds that the probative value of the evidence is greater than its prejudice to the objecting party, it can be admitted. If not, it is excluded. This balancing test would allow the judge to consider the totality of the circumstances of the particular case, including whether the challenged evidence is corroborated by other evidence that is admissible, thereby lessening the likelihood of prejudice to the objecting party. Notwithstanding our preference for the test stated in our proposed rule, Professor Capra's proposed rule is a significant improvement to the current evidence rules, and we commend the Committee for being willing to consider it.

[231] Judicial Conference of the U.S., *Advisory Committee on Evidence Rules; Meeting of the Judicial Conference*, 90 Fed. Reg. 19228 (May 6, 2025).

The use of AI and generative AI applications is becoming increasingly pervasive in all aspects of business, government, and personal matters. It is inevitable that litigation in the near term—and progressively more so in the future—will involve evidence generated by these applications. Judges need to anticipate that the cases assigned to them will invariably involve AI-generated evidence, and may also involve deepfake challenges, and must be prepared to handle both of these situations. Scheduling orders should require early disclosure from the parties about whether they intend to introduce AI-generated evidence, and set a deadline for doing so. There should also be a deadline for raising a challenge to AI-generated evidence that the opposing party has given notice it intends to introduce. The court should address the discovery the parties will reasonably need to be able to mount a challenge to AI-generated evidence that may be introduced against them, or to decide whether they need to retain an expert to evaluate or dispute potential AI-generated evidence. If so, there should be a deadline for disclosure of the experts' reports. Finally, the court should set a deadline for an evidentiary hearing and/or argument on the admissibility of acknowledged AI-generated or potentially deepfake evidence sufficiently far in advance of trial to be able to carefully evaluate the evidence and challenges and to make a pretrial ruling. These issues are simply too complex and time consuming to attempt to address on the eve of or during trial.

B. Discovery about Acknowledged or Unacknowledged AI-Generated Evidence

As we have shown in this Article, acknowledged and unacknowledged AI-generated evidence involve scientific and technical information. When the parties agree that the evidence is AI-generated, the key evidentiary issue of *admissibility* necessarily focuses on how the AI software was developed, trained, and tested, and whether it (or the results it supplies) is valid, reliable, and unbiased. This cannot be determined without discovery about the AI application involved (potentially including its source code), the data on which it was trained, validation of the AI system for its intended purpose and information about who performed that testing, information regarding false-positive and false-negative error rates or other limitations on the use of the AI system, and information concerning any potential biases that could affect the validity and reliability of the output of the AI system.

When the source of the evidence is in dispute, i.e., the party against whom the evidence is being proffered claims it is a deepfake, the key evidentiary issue of *authenticity* is implicated. That may require access to the original hardware or media on which the evidence was created or maintained or the native version of the evidence, including its metadata. This may be the only way of determining the provenance of the evidence. Both situations may involve discovery that is more extensive and intrusive than is the norm in most civil or criminal matters.

C. Use of Protective Orders to Address Issues Associated with Claims of Proprietary Information or Trade Secrets and Claims of Confidentiality or Privacy

Many if not most of the parties that will use AI applications will not have developed the software themselves; they will use applications they have licensed from others. The developers will undoubtedly consider the information needed to evaluate the validity and reliability of the systems as proprietary information, or as trade secrets, and can be expected to object to allowing access to the information that is needed to evaluate the AI application.[232] While trade-secret claims are legitimate and must be taken seriously, they seldom warrant a court order precluding the party against whom the evidence will be offered from having access to it to be able to examine and mount a challenge to the evidence.[233] The better practice is for the court to allow reasonable discovery subject to a protective order that can be tailored to the facts of the particular case. Simply put, if a party intends to use AI-generated evidence, it cannot be allowed to do so while simultaneously objecting to any discovery by the opposing party and thereby preventing them from evaluating and challenging the evidence. The same analysis applies to assertions of confidentiality or privacy when a party is proffering evidence that it claims is genuine but its opponent claims is a deepfake. The forensic examination of a device such as a phone may be intrusive, but the choice should be either to allow discovery to be able to use the evidence or otherwise be precluded from using the evidence at a hearing or trial. Anything else is unfair and may very well raise due process concerns.

D. Expert Witnesses

Given the fact that the evaluation of AI evidence is, by definition, scientific, technical, or specialized, and ferreting out deepfake evidence is beyond the capabilities of lay witnesses and jurors, it is almost unavoidable that expert witnesses will be involved in cases where acknowledged and unacknowledged AI-generated evidence is presented. In both instances, the party offering the evidence will need experts to authenticate the evidence in order for it to be admitted, and the opposing party will need them to evaluate and potentially challenge the evidence, either as to authenticity (in the case of unacknowledged AI-generated evidence) or validity, reliability, and bias (in the case of acknowledged AI-generated evidence). Both the Federal Rule of Civil Procedure and Criminal Procedure have rules dealing with expert witness disclosures.[234] Expert disclosures should be detailed and not conclusory and must address the evidentiary issues that judges have to consider when ruling on evidentiary challenges, such as the Rule 702 reliability factors and the *Daubert* factors that we have previously discussed. Further, the expert needs to be sufficiently qualified to be able to testify about the AI application at issue. For example, while a law enforcement officer may be sufficiently well

---

[232] *See* John G. Sprankling, *Trade Secrets in the Artificial Intelligence Era*, 76 S.C. L. Rev. 181, 218 (2024).

[233] *See also* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1395 (2018) (arguing that "when trade secret evidence is relevant to a case, protective orders, sealing, and limited courtroom closures provide sufficient safeguards").

[234] *See* FED. R. CRIM. P. 16(a)(1)(G), (b)(1)(C); FED. R. CRIM. P. 16.1; FED. R. CIV. P. 26(a)(2).

trained on how to *use* a particular AI application (e.g., a facial recognition technology), that does not mean they have the knowledge, training, or experience needed to explain how the application was developed, trained, and tested.

The more troublesome situation will be the one where the parties cannot afford to hire experts. In those cases, the court should consider engaging its own expert under Federal Rule of Civil Procedure 706,[235] but there may well not be funds available to pay for that. In such cases, courts might seek out forensic practitioners who are willing to volunteer a limited amount of their time pro bono or local districts might arrange for a pool of funds to provide for experts when necessary and appropriate.

E. Motions Practice

Finally, when the parties disclose that they will seek to use acknowledged or unacknowledged AI-generated evidence at trial and the court has addressed any necessary discovery, a pre-trial hearing to rule on evidentiary issues and challenges associated with the admissibility of the AI-generated or potentially deepfake evidence should be held. The court should require written motions that set forth in detail the basis for any requests that the court make a pre-trial determination to either admit or exclude the evidence. When scheduling the filing of the motions and any follow-on hearing to rule on them, the court should be very clear in letting the parties and their experts know what information the court needs in order to be able to rule, such as information concerning validity, reliability, error rates, bias, etc., in the case of acknowledged AI-generated evidence, and in the case of unacknowledged AI-generated evidence, information about the most likely source of evidence, what the content or metadata suggests about provenance or manipulation, and the probative value of the evidence versus the prejudice that could occur were the evidence to be admitted. This information will help counsel to focus on what the judge needs to know and provide the court with what it needs to make a proper pre-trial ruling as far in advance of trial as possible.

IX. CONCLUSION

In this Article, we have argued that, compared to run-of-the-mill computer-generated evidence, such as Excel Spreadsheets, AI-generated evidence is far more challenging. First, it is typically the product of complex and opaque systems that cannot be explained even by their developers. Second, evidence created using GenAI, in particular, has a unique ability to persuade and shape the attitudes and perceptions of those who view it. While there is certainly a plethora of promising uses for GenAI, it can readily be used to mislead and misinform, presenting novel and unique evidentiary challenges, especially due to its easy accessibility by those who would use it to create seemingly authentic content to manipulate or defraud others at scale.

[235] FED. R. CIV. P. 706 (providing for court-appointed expert witnesses).

The existing Federal Rules of Evidence are, for the most part, technology neutral. Because of the procedural requirements of the Rules Enabling Act, they are not easily or quickly revised. At a minimum, the process of revising the evidence rules takes three years, and more frequently as many as five or six. Any amendment to the rules should be crafted to remain applicable regardless of how GenAI and its applications may evolve.

As we have discussed, if applied flexibly, the current rules of evidence can be used today to deal with both acknowledged and unacknowledged AI-generated evidence. This is a good thing, because the earliest that the proposed new evidentiary rules that we discuss—or those being considered by the Advisory Committee on Evidence Rules—could make it through the rule-making process is 2027, and more realistically, 2029, assuming the Committee decides to move forward with proposed rules changes addressing acknowledged and unacknowledged AI-generated evidence. We have identified a real limitation in the current rules—what we describe as the evidentiary "catch 22"—in dealing with the authenticity of and potential unfair prejudice associated with—unacknowledged AI-generated evidence that is asserted to be a deepfake. We have cited to case law that may offer a viable means to address this potential unfair prejudice, but because it is borrowed from analogous—and not identical—evidentiary challenges, it is therefore of uncertain applicability. Our hope is that, in 2025, the Advisory Committee on Evidence Rules will move forward with proposed rules changes to address the challenges we have identified with acknowledged and unacknowledged AI-generated evidence. But regardless of what the Committee does, judges and lawyers will have to come to terms with these challenges now. We have therefore closed this piece with some practical advice as to how judges might mitigate the risks presented by this powerful new form of evidence.