DRAFT

DEEPFAKES ON TRIAL 2.0: A REVISED PROPOSAL FOR A NEW FEDERAL RULE OF EVIDENCE TO MITIGATE DEEPFAKE DECEPTIONS IN COURT

Professor Rebecca A. Delfino[*]
LMU Loyola Law School, Los Angeles

Loyola Law School, Los Angeles Legal Studies Research Paper No. 2025-10,
DOI: 10.13140/RG.2.2.12632.61447
Available at SSRN: https://ssrn.com/abstract=5188767 or
http://dx.doi.org/10.2139/ssrn.5188767

INTRODUCTION

The increasing sophistication of generative artificial intelligence creates unique and unprecedented challenges for courts in assessing the authenticity of computer-generated or electronic evidence. Despite the importance of authenticating evidence before it is admitted, Federal Rule of Evidence 901 does not explicitly address AI-generated falsifications, and existing authentication standards may be insufficient to detect them. Thus, courts lack clear guidance on evaluating alleged deepfake evidence, increasing the risk that it will be admitted without adequate scrutiny.

In *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery*,[1] I argued that the unique dangers of deepfake evidence call for an amendment to FRE 901. Specifically, I proposed a new subsection, FRE 901(c), that would reallocate the determination of authenticity from the jury to the court, ensuring that AI-generated or manipulated evidence is properly authenticated before admission.[2] The proposal sought to mitigate the risk of juror misjudgment regarding deepfakes, which exploit cognitive biases such as the "seeing is believing" heuristic.[3] Moreover, reallocating the admissibility determinations to the court protects the integrity of legal proceedings

---

[*] Associate Dean for Clinical Programs and Experiential Learning, and Professor of Law at LMU Loyola Law School, Los Angeles.

[1] 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) (*Deepfakes on Trial*).

[2] *Id.* at p. 341. The section 901(c) in *Deepfakes on Trial*, provided: "901(c). Notwithstanding subdivision (a), to satisfy the requirement of authenticating or identifying an item of audiovisual evidence, the proponent must produce evidence that the item is what the proponent claims it is in accordance with subdivision (b). The court must decide any question about whether the evidence is admissible." (the "Original Proposal".) *Id.*

[3] *Id.* at pp. 337, 346-47.

Citation: Delfino, Rebecca, DEEPFAKES ON TRIAL 2.0: A REVISED PROPOSAL FOR A NEW FEDERAL RULE OF EVIDENCE TO MITIGATE DEEPFAKE DECEPTIONS IN COURT (February 15, 2025). Loyola Law School, Los Angeles Legal Studies Research Paper No. 2025-10, DOI: 10.13140/RG.2.2.12632.61447, Available at SSRN: https://ssrn.com/abstract=5188767 or http://dx.doi.org/10.2139/ssrn.5188767

while avoiding delay, confusion, and prejudice caused by deepfake evidence allegations.[4]

Since the publication of *Deepfakes on Trial*, other scholars have contributed to this discourse, proposing alternative frameworks for addressing AI-generated evidence.[5] Considering these developments, new experiential studies on deepfake detection, and advancements in deepfake detection technology, this paper sets forth a revised proposal for FRE 901(c), balancing the necessity of rigorous authentication with the evidentiary burdens placed on litigants.

I. THE REVISED PROPOSAL FOR FRE 901(c)

Generative artificial intelligence can create realistic but entirely fabricated content, which may be used to falsely implicate individuals in crimes, create fake confessions, alter historical records, or fabricate news events.[6] Courts are already encountering challenges in assessing the authenticity of AI-generated evidence.[7] Existing authentication methods under Rule 901(b), such as witness testimony and metadata verification, may be insufficient to detect AI manipulation.[8]

---

[4] *Id.* at pp. 341-42, 345-46, 348.

[5] See Daniel J. Capra, *Deepfakes Reach the Advisory Committee on Evidence Rules,* 92 Fordham L. Rev. 2491 (2024) (summarizing the scholarship addressing the challenges deepfakes pose to the current evidentiary framework and exploring potential amendments to the Federal Rules of Evidence to enhance the authentication process).

[6] *Id*. at pp. 299-302; Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,* 107 Calif. L. Rev. 1753, 1772–1785 (2019).

[7] See e.g., *In re Woori Bank,* 2021 WL 2645812, p. *1-2 (N.D. Cal. 2021) (plaintiff sought discovery from social media platform to support his defamation action based on claim that a "deepfake" image of the plaintiff engaging in an improper intimate act had been posted on a social media platform); *Hohsfield v. Staffieri,* 2021 WL 5086367, p. *1 (N.J. 2021) (plaintiff brought a 42 USC 1983 action against police officers, claiming that they created a deepfake photo of him engaging in a lewd act to frame him and justify his arrest.); *Schaffer v. Shinn,* 2021 WL 6101435, p.*7 (Ariz. 2021) (defendant attacked sufficiency of the evidence supporting sentencing enhancement arguing that the pornographic image was a deepfake); *People v. Smith*, 969 N.W.2d 548, 565-567 (Mich. 2021) (defendant challenged the admission of Facebook posts belong to others which purportedly included his image and gang moniker, suggesting that they were fake).

[8] Delfino, *supra* note 1 at pp. 333-35, 341; Agnieszka McPeak, *The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood,* 23 Vand. J. Ent. & Tech. L. 433, 456–460 (2021) (arguing that traditional authentication methods under Federal Rule of Evidence 901(b), including eyewitness testimony and metadata verification, are inadequate to reliably detect sophisticated AI-generated deepfake manipulations); Marie-Helen Maras & Alex Alexandrou, *Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos,* 23 Int'l J. Evidence & Proof 255, 260–64 (2019) (explaining how conventional authentication techniques such as witness testimony and metadata examination are likely insufficient to accurately detect and authenticate video and audio manipulated by advanced generative AI technology).

The challenge presented by deepfakes requires a heightened authentication standard because traditional evidence verification techniques were not designed to address highly sophisticated AI-generated falsifications.[9] Without a new rule to address fraudulent AI-generated evidence, fake evidence could be admitted based on authentication methods that are ineffective in addressing the challenges presented by the technology, increasing the risk that jurors will be exposed to convincing but entirely false evidence.[10] The lack of explicit procedural safeguards also risks inconsistent application of authentication requirements to AI-generated content, leading to evidentiary confusion and unfair trial outcomes.[11] The Revised Proposal fills this gap by establishing a clear and structured approach to determining the admissibility of AI-generated evidence. Thus, Rule 901 should be amended to add a new subdivision (c), which would provide:

> 901(c). Notwithstanding subdivision (a), if a party challenging the authenticity of computer-generated or other electronic evidence presents evidence sufficient to support a factual finding that the challenged evidence has been manipulated or fabricated, in whole or in part, by generative artificial intelligence, the proponent of the evidence must authenticate the evidence under subdivision (b) and provide additional proof establishing its reliability. The court must decide the admissibility of the challenged evidence under Rule 104(a). (the "Revised Proposal")

II. COMPARATIVE ANALYSIS: THE REVISED PROPOSAL AND EMERGING ALTERNATIVES

The Revised Proposal differs from previous frameworks, including the Original Proposal in *Deepfakes on Trial* and alternative proposals by Professor Paul Grimm and Professor Maura R. Grossman.[12] ("Grimm & Grossman Proposal") and

[9] Delfino, *supra* note 1 at pp. 332-35; McPeak, *supra* note 8 at pp. 456-61; Riana Pfefferkorn, *"Deepfakes" in the Courtroom,* 29 B.U. Pub. Int. L.J. 245, 249–257 (2020).
[10] Delfino, *supra* note 1 at pp. 332-35, 340-42; McPeak, *supra* note 8 at pp. 456-62.
[11] Delfino, *supra* note 1 at pp. 336-42; McPeak, *supra* note 8 at pp. 459-62; Pfefferkorn, *supra* note 9 at pp. 255-57.
[12] Daniel J. Capra, Reporter to the Judicial Conference Advisory Committee on Evidence Rules, *Memorandum to Advisory Committee on Evidence Rules, Re: Artificial Intelligence, Machine Learning, and Possible Amendments to the Federal Rules of Evidence,* October 1, 2024, pp. 22-23. (Describing the proposal offered by Professor Paul W. Grimm and Professor Maura R. Grossman to amend the Federal Rules of Evidence to explicitly address the authentication and admissibility challenges posed by deepfake and AI-generated evidence. Their proposal states: "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, the evidence is admissible only if the proponent demonstrates that its probative value outweighs its prejudicial effect on the party challenging the evidence."); *see also Symposium on Scholars' Suggestions for Amendments and Issues Raised by Artificial Intelligence,* 92 FORDHAM L. REV. 2375, 2430-32 (2024).

the Committee Reporter's Amendment to the Grimm & Grossman Proposal ("Reporter's Amendment").[13] The following sections explore the key distinctions between the alternative proposals.

A. PRECISION IN SCOPE: FOCUSING ON TERMINOLOGY TO TARGET AI THAT CREATES FABRICATED EVIDENCE

The Original Proposal in *Deepfakes on Trial* applied to "digital audiovisual evidence."[14] Although commentary to the rule could supply interpretative guidance on the meaning of these terms in the context of deepfake allegations, in recognition vast array of evidence that could be characterized as "digital audiovisual evidence" (though unrelated to fake evidence) and the impact on the courts of making such determinations in each case involving digital audiovisual evidence, the Revised Proposal employs more exact and focused terminology, designed to limit the application of the rule only to those cases involving deepfake technology.

Thus, the Revised Proposal uses the term "generative artificial intelligence" to ensure clarity and precision in discussions concerning AI-generated evidence. Unlike the broader and more ambiguous phrase "artificial intelligence by an automated system," used in the Grimm & Grossman Proposal[15] and the Reporter's Amendment,[16] "Generative Artificial Intelligence" accurately identifies the specific type of AI technology responsible for creating fabricated content.[17] This specificity is essential in avoiding ambiguity and unnecessary overbreadth in legal and regulatory discussions.

Generative artificial intelligence refers to AI models that create new content, including synthetic videos, images, and audio, that can fabricate events that never occurred.[18] Technologies such as deepfake generators, text-to-image models like DALL·E and Midjourney, AI voice cloning, and synthetic video editing tools fall within this category. The primary concern in authentication disputes is the ability of generative AI to create evidence that appears real but is entirely fabricated. Other AI-driven enhancements, such as AI-powered photo enhancement, voice amplification, and predictive text tools, do not pose the same risk. Unlike generative AI, these tools enhance or clarify existing content rather

---

[13] Capra, *supra* note 12 at p. 32. The Reporter's Amendment to the Grimm & Grossman Proposal provides: "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." *Id.*

[14] Delfino, *supra* note 1, at p. 341.

[15] Capra, *supra* note 12 at p. 23.

[16] *Id.* at p. 32.

[17] Fed. Trade Comm'n, *Protecting Consumers from Fraud and Deception in the Age of AI* 2 (2023), https://www.ftc.gov/system/files/ftc_gov/pdf/P231200%20AI%20Guidance.pdf.

[18] Chesney & Citron, *supra* note 6 at pp. 1758-59.

than fabricate new content. Their outputs are tethered to authentic data sources and are generally verifiable through traditional authentication methods, which means they pose a significantly lower threat to evidentiary reliability than synthetic content. As the National Institute of Standards and Technology (NIST) has recognized, "enhancing tools like noise reduction or predictive text typically maintain fidelity to source material and are not inherently deceptive in nature," and thus present different risk profiles than generative AI systems that produce entirely synthetic outputs.[19] However, the term "artificial intelligence by an automated system" used in other proposals under consideration could mistakenly encompass these legitimate AI tools, subjecting them to undue scrutiny. This proposal avoids unnecessary complications in authenticating digital evidence by specifically targeting generative AI.

In comparison, the alternative phrase, "artificial intelligence by an automated system," is broad and imprecise. "Artificial intelligence" broadly includes all machine-learning systems, even those that do not generate synthetic content.[20] The additional phrase "by an automated system" further expands the scope to include any AI-driven process, such as predictive analytics, automated transcription, machine vision analysis, and digital forensics tools.[21] This lack of specificity increases the risk of misapplication, leading to situations where AI-enhanced evidence, rather than AI-created evidence, is subjected to unnecessary scrutiny. For instance, a security camera video enhanced using AI-based sharpening filters could be wrongly challenged as synthetic evidence despite being legitimate. Similarly, AI-powered speech-to-text transcription of court proceedings could be mistakenly classified under this vague definition, imposing unnecessary authentication burdens on standard transcription evidence.

Furthermore, judicial and legislative trends favor "generative artificial intelligence" as a distinct category. Courts and regulators are already differentiating between generative AI and other AI applications. The European Union AI Act[22] and discussions surrounding U.S. AI policy clearly distinguish between AI used for automation—fraud detection and content moderation—and AI used for content generation, including deepfakes and synthetic voices. The Federal Trade Commission (FTC) has also issued guidance addressing generative AI fraud, demonstrating that "generative AI" is already well-established in legal and regulatory discussions.[23] Aligning with these emerging legal and technological

[19] Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 22 (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[20] Org. for Econ. Co-operation & Dev. [OECD], *A Framework for Classifying AI Systems: An Overview* 6–7 (2022), https://oecd.ai/en/classification.
[21] *Ibid.*
[22] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, available at EUR-Lex.[https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng]
[23] Fed. Trade Comm'n, Generative AI and the Risk of Consumer Harm, FTC (Jan. 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2025/01/ai-risk-consumer-harm.

standards ensures consistency and clarity in judicial interpretation. Conversely, using the vague phrase "artificial intelligence by an automated system" risks confusion, as courts would be tasked with determining what falls under this broad term. Using "generative AI" aligns with emerging legal frameworks that distinguish generative AI from other forms of AI, ensuring that courts apply the rule consistently and avoid evidentiary confusion.

B. A STRUCTURED BURDEN-SHIFTING MODEL: BALANCING ACCESS AND ACCURACY

The revised proposal establishes a clear burden-shifting framework and appropriate burdens on the challenger and the proponent of the evidence. Like the Grimm & Grossman Proposal, the Revised Proposal requires the party challenging the authenticity to present evidence sufficient to support a factual finding that the challenged evidence has been altered or fabricated *before* requiring the proponent of the evidence to come forward to demonstrate the evidence is genuine.

However, the Revised Proposal differs from the alternative proposals. It articulates a clear proponent's burden, thus creating a structured approach to evaluating AI-manipulated evidence by establishing a burden-shifting framework. Also, unlike the balancing test in the Grimm & Grossman Proposal that presumes authenticity and requires courts to weigh probative value against prejudicial effect,[24] the Revised Proposal introduces a clear and structured burden-shifting framework to evaluate alleged deepfake evidence. The mechanism is grounded in authenticity rules to ensure that generative AI-manipulated evidence meets authenticity and reliability standards before admission.

1. *The Challenger's Burden: "Presents Evidence Sufficient to Support a Factual Finding"*

Under the Revised Proposal, the party challenging the authenticity of AI-generated evidence must provide sufficient evidence to support a factual finding that AI manipulation may have occurred. This standard aligns with Rule 104(b) and deters frivolous challenges to legitimate digital evidence. A challenger must provide expert testimony, forensic evidence, or AI-detection analysis that suggests the evidence could be AI-generated. Requiring the challenger to make a threshold showing that the evidence is a deepfake is a crucial regulatory check against deepfake claims that might be raised in every case involving digital audio-visual evidence. This threshold requirement ensures that authentication challenges are legitimate while preventing unnecessary litigation.

[24] *Id*. at p. 22-23, 31.

2. The Proponent's Burden: "Authenticate the Evidence Under 901(b) and Provide Additional Proof Establishing Its Reliability."

Although FRE 901 has historically been concerned only with authenticity, deepfake evidence presents unique authentication challenges that traditional standards do not address. Because AI-generated evidence can be so convincingly realistic, traditional authentication alone does not ensure the evidence is genuine. Thus, under the Revised Proposal, once a credible challenge is made, the proponent of the evidence must meet a heightened authentication standard by first authenticating the evidence under traditional Rule 901(b) methods, such as metadata or witness verification. Second, the proponent must provide additional proof of reliability.

The unique risks of AI-generated evidence justify a heightened standard. Deepfakes and AI-generated content are fundamentally different from traditional manipulated evidence. Unlike traditional manipulated evidence, deepfakes do not merely distort reality; they fabricate it entirely, making traditional authentication standards insufficiently rigorous to reliably detect falsification.[25] Moreover, deepfakes can mimic real individuals with near-perfect accuracy—posing unique risks of deception. They can also be mass-produced quickly and spread widely, raising concerns about their impact on judicial truth-seeking.[26] Because AI-generated evidence can so convincingly mimic reality, requiring additional proof of reliability ensures that courts apply a heightened evidentiary standard to AI-generated content. FRE 901(b) alone cannot address the unique risks AI-generated deepfakes pose.

Courts need an extra reliability safeguard because deepfakes differ from traditional forms of manipulated evidence in ways that complicate authentication and increase the potential for deception for three reasons. First, the traditional FRE 901(b) standard is too lenient for AI-generated evidence. Many traditional authentication methods under FRE 901(b) do not work well for AI-generated deepfakes. For example, witness testimony (901(b)(1)) may be unreliable; AI can generate false but hyper-realistic content, making it hard even for eyewitnesses to detect manipulation. In addition, metadata (901(b)(4)) can be easily falsified. AI-generated content can be inserted into real files, and metadata can be modified to make it appear legitimate.[27] Finally, expert comparison (901(b)(3)) may be difficult because AI-generated videos, images, and audio can be nearly indistinguishable from real content.[28] Requiring an extra layer of scrutiny ensures that authentication is not just a formal check-box process but instead that courts actually evaluate whether the evidence is trustworthy.

[25] Pfefferkorn, *supra* note 9 at pp. 248-50.
[26] Chesney & Citron, *supra* note 6, at pp. 1768-77.
[27] McPeak, *supra* note 8 at pp. 460-61.
[28] Delfino, *supra* note 1 at pp. 333-35.

Second, FRE 901(b) concerns authentication (showing that evidence is what it purports to be), but it does not necessarily establish reliability. For example, a perfectly forged AI-generated video may technically be authenticated under 901(b)(4) (appearance, contents, substance), even if it is entirely fake. Under traditional authentication rules, if a witness testifies, "Yes, this looks like what I saw," the evidence could pass authentication—even if it is unreliable. Courts need a reliability check beyond authentication to ensure that AI-generated evidence is technically authenticated, truthful, and accurate. This safeguard is analogous to the *Daubert* standard for expert testimony under Rule 702, which requires that expert evidence be relevant and reliable.[29]

Although imposing a reliability/trustworthiness requirement creates an additional burden on the proponent of deepfake evidence, that burden is outweighed by the risks associated with admitting unreliable evidence. The proponent is best positioned to establish reliability/trustworthiness through forensic analysis, expert testimony, or digital verification methods. Furthermore, even though the requirement may introduce additional steps in judicial proceedings, courts already handle complex evidentiary issues involving forensic science and chain-of-custody disputes. Implementing a reliability standard for deepfake evidence is a natural evolution of evidentiary safeguards rather than an insurmountable challenge. While requiring proof of reliability may impose a higher evidentiary burden, it is justified to prevent legal manipulation and ensure the integrity of audiovisual evidence in court proceedings.

Finally, requiring more proof of reliability does not burden proponents of legitimate AI-enhanced evidence unnecessarily. The Revised Proposal does not impose a heightened burden on all electronic evidence; it applies only when a credible challenge has been made that generative AI manipulated the evidence. The proponent may authenticate the evidence under Rule 901(b) if no such challenge exists. This safeguard ensures that courts do not impose an undue burden on parties relying on AI-enhanced but authentic evidence while preventing AI-generated fabrications from being admitted into evidence.

The Revised Proposal would effectively address the problem of deepfakes. Imposing a stricter evidentiary standard acknowledges the reality that deepfakes can be highly realistic fabrications, necessitating more safeguards in the judicial process. This dual requirement acknowledges that even though authentication confirms the source of the evidence, it doesn't necessarily attest to its content's veracity, especially given the sophisticated nature of deepfakes. By mandating both authentication and additional proof of trustworthiness, the revised proposal provides a more robust framework to address the unique challenges deepfakes present in legal proceedings, which will become increasingly important, particularly as deepfake technology becomes more sophisticated. Ensuring both

[29] *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 589 (1993).

authentication and reliability is essential to maintaining the integrity of judicial proceedings and preventing the admission of deceptive audiovisual content.

While the Revised Proposal focuses on ensuring the accuracy and admissibility of AI-generated evidence, it also intersects with broader concerns about fairness and equity in litigation.[30] In a related article, Deepfakes and Access to Justice, I explore how the cost of proving or disproving the authenticity of deepfakes can serve as a barrier to justice for economically disadvantaged litigants.[31] The high cost of digital forensic experts and the absence of meaningful cost-shifting mechanisms may leave parties unable to challenge or defend against deepfake evidence, effectively denying them access to the courts.[32] By requiring the proponent of disputed AI-generated evidence to establish both authenticity and reliability, the Revised Proposal not only safeguards evidentiary integrity but also helps distribute litigation burdens more equitably—particularly in cases where one party lacks the resources to meet the technological demands of modern evidence.[33] Thus, evidentiary reform and access-to-justice reform are complementary responses to the deepfake threat, each necessary to ensure a fair and functional legal process.[34]

C. RECLAIMING JUDICIAL GATEKEEPING IN THE AGE OF DEEPFAKES

The revised proposal retains the requirement for judicial determination under FRE 104(a) of authenticity. Before the 1930s, courts in the United States applied the traditional English view that the judge had plenary authority to decide all questions of fact conditioning the admissibility of evidence.[35] However, under the current evidentiary framework in FRE Rule 104, judges make a preliminary determination regarding authenticity, but jurors ultimately decide whether the evidence is genuine.[36] The decision to assign the ultimate decision on authenticity to juries reflects the view that based on their innate human perceptive skills and lived experience, jurors are equally capable and effective at making authenticity assessments as judges. The FRE 104 framework assumes that jurors can reasonably evaluate the credibility of audiovisual evidence.[37] However, deepfake technology upends this process by making distinguishing between real and manipulated content difficult.

[30] See Rebecca A. Delfino, *Pay-to-Play: Access to Justice in the Era of AI and Deepfakes*, 55 Seton Hall L. Rev. 789, 796–97 (2025).
[31] *Id.* at pp. 790–91.
[32] *Id.* at pp. 792–93.
[33] *Id.* at pp. 801–02.
[34] *Id.* at pp. 795–96.
[35] Delfino, *supra* note 1 at pp. 323-24, 342-43.
[36] Fed. R. Evid. 104(a) & (b).
[37] Delfino, *supra* note 1 at pp. 323-24, 341-43.

In *Deepfakes on Trial*, I argued that deepfake evidence offered unprecedented challenges to the legal proceedings, which the current FRE 104 legal framework governing authenticity determinations in the federal rules of evidence could not effectively address.[38] The human tendency to rely on visual perception as a primary source of truth, combined with the sophisticated realism of deepfakes, makes jurors particularly susceptible to deception.[39] Additionally, the widespread awareness of deepfake technology fosters skepticism, leading some jurors to doubt authentic digital evidence. The mere existence of deepfakes enables bad actors to exploit uncertainty and cast doubt on legitimate evidence, further eroding confidence in the legal system process.[40]

Thus, the Original Proposal relocated the determination of authenticity from the jury under FRE Rule 104(b) to the judge under FRE Rule 104(a) because the traditional allocation of fact-finding responsibilities exacerbates the risks associated with deepfake evidence.[41] If jurors incorrectly authenticate deepfakes as real, fabricated evidence may be admitted and relied upon in court. Conversely, genuine evidence may be dismissed as fake due to increasing skepticism over digital manipulation.[42] The Original Proposal reallocated this responsibility to the judge to ensure a consistent and informed authenticity assessment before evidence reaches the jury.[43]

1. Deepfake Detection and the Comparative Competence of Judges

The approach reflected in the Original Proposal was grounded in computer science research on deepfake detection.[44] Although few empirical studies on the human ability to detect deepfakes had been published before I wrote *Deepfakes on Trial*, one behavioral experiment study published in late 2021, conducted by researchers from the Center for Humans and Machines at the Max Planck Institute for Human Development and the University of Amsterdam School of Economics, (the "Planck Institute Study") found that laypersons struggled to differentiate deepfakes from real footage, even after being trained in detection techniques.[45] The study found that participants consistently failed to detect deepfakes, and their overconfidence in their ability further exacerbated the problem.[46]

[38] *Id.* at pp. 297-98, 332-33.
[39] *Id.* at. p. 337.
[40] *Id.* at pp. 311-13, 338.
[41] *Id.* a pp. 341-42.
[42] *Id.* at pp. 309-10.
[43] *Id.* at p. 342.
[44] *Id.* at p. 337.
[45] Nils C. Köbis, Barbora Doležalová & Ivan Soraperra, *Fooled Twice: People Cannot Detect Deepfakes, but Think They Can,* iScience, Oct. 29, 2021, at 1, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8602050/pdf/main.pdf.
[46] *Id*. at p. 5.

In *Deepfakes on Trial*, I argued that, compared to lay juries, available research suggested that judges were better suited to assess the authenticity of digital audiovisual evidence because of their training and ability to engage in disciplined evaluation.[47] I cited studies showing that although judges are not entirely immune to cognitive biases, they are generally more resistant to certain heuristic errors that affect laypersons. Their professional experience in assessing legal evidence enables them to filter out misleading arguments and focus on technical indicators of authenticity.[48] Moreover, I pointed out that judges can develop expertise in forensic technology and deepfake detection outside the context of a specific case, allowing them to apply more informed scrutiny when evaluating evidence.[49] Given these advantages, reallocating authenticity determinations to judges under Rule 104(a) would enhance the accuracy of evidentiary assessments and help safeguard the integrity of judicial proceedings.[50]

The Revised Proposal retains the requirement in the Original Proposal to reallocate the final decision on questions related to authenticity to the court. New experimental computer science research confirms the argument in *Deepfakes on Trial* that the task of detecting deepfakes is a task better suited to judges than juries.

2. Empirical Support for Judicial Gatekeeping in Deepfake Cases

In mid-2024, Alena Birrer and Natascha Just, research scholars from the University of Zurich, published a review of recent experiments and research on deepfake detection Alena Birrer and Natascha Just, *What We Know and Don't Know About Deepfakes: An Investigation into the State of the Research and*

[47] Delfino, *supra* note 1 at p 347.
[48] *Ibid*. (Citing the following studies and articles: Valerie Hans, *Judges, Juries, and Scientific Evidence,* 16 J.L. & Pol'y 19, 36–37 (2007) (finding that only 15% of judges accepted a fallacious argument about mitochondrial DNA evidence, compared to 49% of mock jurors); Elizabeth Thornburg, (*Un)Conscious Judging,* 76 Wash. & Lee L. Rev. 1567, 1620–23 (2019) (arguing that with sufficient training, judges can overcome heuristic responses that might otherwise affect their decision-making); Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind,* 86 Cornell L. Rev. 777, 784, 816–17 (2001) (studying of 167 federal magistrate judges, finding that judges were just as susceptible as laypeople to biases like anchoring, hindsight bias, and egocentric bias but were less affected by framing and the representativeness heuristic); Andrew J. Wistrich, Chris Guthrie & Jeffrey J. Rachlinski, *Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding,* 153 U. Pa. L. Rev. 1251, 1318–22 (2005) (testing judges' ability to disregard legally inadmissible evidence and found that their decision-making was generally not influenced by such information in certain contexts); Jeffrey J. Rachlinski, Chris Guthrie & Andrew J. Wistrich, *Probable Cause, Probability, and Hindsight,* 8 J. Empirical Legal Stud. 72, 72 (2011) (finding that 900 state and federal judges' probable cause judgments were generally unaffected by knowledge of case outcomes).
[49] Delfino, *supra* note 1 at p. 348.
[50] *Id*. at p. 342.

Regulatory Landscape.[51] Birrer and Just described 22 experimental computer science studies that explored the effectiveness of both humans and artificial intelligence in identifying deepfake images and videos.[52] These studies relied on extensive datasets to assess detection accuracy and influencing factors. The findings revealed that human participants could correctly identify deepfakes with an average accuracy of 63.3%.[53] However, their success rate varied depending on several factors, including image resolution, familiarity with the person depicted, and demographic similarities between the observer and the deepfake subject.[54]

The studies reviewed also confirmed the Planck Institute Study's finding that many participants overestimated their ability to distinguish real images from deepfakes, a cognitive bias often linked to the *Dunning-Kruger* effect,[55] where people with limited expertise tend to overrate their skills. Another striking pattern that emerged from the recent studies was that the age, race, and gender of the deepfake subject influenced detection accuracy in several ways. Participants were generally more successful in identifying the forgery when the deepfake subject was younger or more well-known, such as a celebrity.[56] This was particularly evident in cases where high-resolution images were available, making manipulation artifacts more visible. Furthermore, demographic similarities between the observer and the deepfake subject also played a role.[57] The research found that participants were better at detecting deepfakes of individuals whose age, gender, or race aligned with their own.[58] This finding suggests that implicit biases shape an individual's ability to discern digital forgeries.

Birrer and Just's report also evaluates various interventions to improve deepfake detection. Some strategies, such as raising awareness through financial incentives or informing participants about common deepfake artifacts, did not impact detection accuracy.[59] Similarly, providing immediate feedback on detection performance produced limited benefits across three studies.[60] Meanwhile, AI-

[51] Alena Birrer and Natascha Just, *What We Know and Don't Know About Deepfakes: An Investigation into the State of the Research and Regulatory Landscape,* New Media & Society (2024), https://doi.org/10.1177/14614448241253138.)
[52] *Id.* at pp. 6-7.
[53] *Ibid.*
[54] *Ibid.*
[55] *Ibid.* The *Dunning-Kruger* effect is a cognitive bias in which individuals with low ability in a particular domain overestimate their competence, while highly skilled individuals tend to underestimate their abilities. Kruger, Justin & David Dunning, *Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*, 77 *J. Personality & Soc. Psychol.* 1121 (1999). This phenomenon occurs because those with limited knowledge lack the metacognitive skills to recognize their own incompetence.
[56] Birrer & Just, *supra* note 51 p. at 7.
[57] *Ibid*
[58] *Id.* at p. 6.
[59] *Id.* at p. 7.
[60] *Ibid.*

assisted detection significantly improved performance, yet it introduced a new challenge—users often placed excessive trust in AI-generated assessments, sometimes altering their judgments based on incorrect AI suggestions.[61]

Among the tested interventions, the most effective was offering participants a detailed walkthrough of examples, helping observers recognize specific deepfake artifacts.[62] This structured and intensive training proved beneficial in enhancing detection skills.[63] Gamification and literacy-based training also showed promise.[64] The type of training needed to increase deepfake detection rates is likely more time-consuming and resource-intensive than what an average trial would allow. However, judges are well suited to receive such training on deepfake detection in connection with judicial training and continuing education requirements. As argued in *Deepfakes on Trial*, the investment in judicial training on deepfake detection would yield benefits in multiple cases.[65]

3. Legal Precedent and Policy Support for Judicial Determination of Authenticity

As I pointed out in *Deepfakes on Trial*, the expansion of a judge's role in determining foundational facts, while unusual, is not without precedent.[66] Although Rule 104(a) and (b) generally allow for jury fact-determination, certain circumstances necessitate the judge's intervention, particularly when the evidence is highly prejudicial or complex.

This approach is clear in policy-based exceptions. One key exception is prejudice-based exclusion, where the judge determines admissibility under Rule 104(a) to prevent undue influence on the jury.[67] For instance, applying attorney-client privilege depends on whether a communication is private. If an eavesdropper overhears a defendant's confession to an attorney, the prosecution may argue that privilege does not apply due to the presence of a third party, while the defense denies this.[68] Since jurors exposed to such testimony might struggle to disregard it even if they find it privileged, courts allocate this determination to the judge to

[61] *Ibid.*
[62] *Ibid.*
[63] *Ibid.*
[64] *Ibid.*
[65] Delfino, *supra* note 1 at p. 348.
[66] *Id.* at p. 342.
[67] Fed. R. Evid. 104(a) advisory committee's note (1972) ("The judge rules on the admissibility of evidence and is not bound by the rules of evidence except those with respect to privileges.").
[68] 8 John Henry Wigmore, Evidence in Trials at Common Law § 2311 (McNaughton rev. 1961) (explaining judicial determination of preliminary facts related to privilege); *see United States v. Gann*, 732 F.2d 714, 723 (9th Cir. 1984) (holding that presence of a third party can negate attorney-client privilege, and such factual determinations are made by the judge under Rule 104(a)).

uphold the integrity of privilege protections.[69] Similarly, the voluntariness of a confession is judged by the court rather than the jury.[70] Another example of prejudicial evidence involves identifying the perpetrator of uncharged misconduct under Rule 404(b).[71] Suppose a personal injury case defendant is accused of assaulting the plaintiff. If evidence emerges that the defendant attempted to bribe a witness to provide false testimony, the judge must first determine whether the defendant committed the bribery before the evidence is admitted.[72] Historically, courts have been cautious about allowing such prejudicial evidence before juries, fearing jurors might assume guilt based on prior misconduct, even when the evidence is weak. This concern has led some courts to assign this fact-finding responsibility to the judge.[73] Beyond prejudice, courts also intervene when evidence, such as scientific testimony, is too complex for jurors to evaluate properly.[74] Establishing the methodological validity of a scientific theory often involves lengthy foundational testimony, sometimes stretching over weeks or months. Jurors exposed to such evidence may struggle to disregard it, even if deemed inadmissible.[75] Therefore, judges rule on the admissibility of scientific evidence under Rule 104(a) to prevent jurors from improperly weighing or misinterpreting the information.

---

[69] *See* Fed. R. Evid. 104(a) advisory committee's note (1972) ("Preliminary questions of fact governing admissibility and the application of privileges are determined by the judge. This includes questions of the existence of a privilege."); Andrew J. Wistrich, Chris Guthrie & Jeffrey J. Rachlinski, *Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding*, 153 U. Pa. L. Rev. 1251, 1255–56 (2005) (noting that even judges—let alone jurors—struggle to disregard inadmissible evidence once exposed to it, justifying judicial gatekeeping).

[70] *See United States v. James*, 576 F.2d 1121, 1127–32 (5th Cir. 1978) (acknowledging that jurors exposed to coerced confessions may nonetheless believe them to be true, making it unrealistic to trust them to set aside such evidence in deliberations.)

[71] Fed. R. Evid. 404(b).

[72] See e.g., *United States v. Sampson*, 486 F.3d 13, 42 (1st Cir. 2007) ("The trial judge, not the jury, must determine whether the proponent has introduced sufficient evidence to support a finding that the defendant committed the prior bad act.")

[73] *See United States v. Smith*, 451 F.3d 209, 217 (4th Cir. 2006) (Courts must take special care that 404(b) evidence is not used simply to paint a defendant as a bad person, especially where the act is prejudicial and only weakly probative.)

[74] *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 592–95 (1993) (holding trial judges must function as gatekeepers to ensure that all scientific testimony or evidence admitted is not only relevant but also reliable. This is a Rule 104(a) determination—judges assess the validity of scientific methodology before it reaches the jury.); Weinstein's Federal Evidence § 702.02[5][a] (2023 ed.) ("Because jurors are generally ill-equipped to evaluate complex scientific or technical evidence, the trial judge must determine admissibility under Rule 104(a) to ensure that such evidence is reliable and will assist rather than confuse the jury.")

[75] *United States v. Brown*, 415 F.3d 1257, 1266 (11th Cir. 2005) ("Once heard, expert testimony—even if ultimately ruled inadmissible—is difficult for jurors to disregard, especially in technical matters where jurors may defer to perceived expertise.")

These judicial fact-finding allocations are not aimed at enhancing truth-seeking but serve broader policy goals, such as protecting privilege, preventing undue prejudice, and ensuring fair deliberations.[76] Deepfakes are similarly complex and potentially prejudicial, justifying the same epistemic curation--where courts control the information presented to juries to shape decision-making in line with policy concerns to preserve the integrity and fairness of trials rather than purely evidentiary accuracy.

Beyond ensuring evidentiary reliability, the Revised Proposal advances several critical policy objectives that underscore its importance as an amendment to the Federal Rules of Evidence. Foremost, it aims to preserve public confidence in the judicial system. The judiciary's legitimacy depends on fair outcomes and public trust in courts to effectively manage technological threats. The admission of convincingly fabricated evidence—or the exclusion of genuine evidence under the mistaken belief that it is fake—risks undermining that confidence. Implementing a clear procedural rule addressing deepfakes signals to the public that courts are equipped to respond to digital deception and are committed to preserving trial integrity.[77]

In addition, the Revised Proposal protects the integrity of judicial fact-finding. Deepfakes introduce entirely fabricated audiovisual content that closely mimics reality, posing unique challenges to the trial process. Unlike traditional evidence disputes, which typically concern degrees of reliability, deepfakes threaten to falsify the underlying evidentiary narrative. The proposal prevents a profound distortion of the fact-finding process by ensuring that unauthenticated or unreliable synthetic content does not reach the jury.[78] It also promotes judicial economy and trial efficiency by requiring judges to resolve foundational authenticity questions early. This proactive approach reduces procedural uncertainty and minimizes mid-trial disruptions that can result in mistrials or protracted appeals.[79] At the same time, establishing a uniform standard ensures consistency and predictability in evidentiary rulings across federal courts, enhancing procedural fairness and reducing forum-dependent outcomes.[80]

Furthermore, the Revised Proposal aims to avoid the chilling effects of using legitimate digital evidence. As deepfakes become more prevalent, there is a

[76] David P. Leonard, *The New Wigmore: A Treatise on Evidence: Selected Rules of Limited Admissibility* § 1.2 (2022 ed.) ("In some contexts, foundational fact-finding is assigned to the judge not because of doubts about the jury's competence, but due to overriding policy goals such as protecting confidential communications or shielding jurors from prejudicial or confusing material.")
[77] Rebecca Wexler, *Real Evidence, Fake Evidence,* 98 Tex. L. Rev. 1165, 1190 (2020).
[78] Andrea Roth, *Machine Testimony,* 126 Yale L.J. 1972, 1975–77 (2017).
[79] Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, *Artificial Intelligence as Evidence,* 25 Yale J.L. & Tech. 1, 60–62 (2023).
[80] Maura R. Grossman & Paul W. Grimm, *Deepfake Evidence: How We Can Distinguish What's Real from What's Not,* 72 Syracuse L. Rev. 243, 266–67 (2022).

growing concern that even authentic digital evidence may be challenged or discredited as fake. This chilling effect could discourage litigants from introducing important digital exhibits, especially those relying on video, audio, or photo documentation. By setting out a straightforward process for distinguishing real from fake, the rule protects the admissibility of legitimate digital evidence.[81] Additionally, by adopting precise terminology and a targeted scope, the proposal aligns with emerging regulatory and international standards, ensuring consistency with broader legal and policy efforts to distinguish between generative and non-generative AI systems.[82] Together, these policy considerations reinforce the Revised Proposal's significance—not only as a matter of evidentiary reform but also as a strategic response to the risks deepfakes pose to judicial legitimacy, fairness, and efficiency.

Moreover, the recommendation to assign judges the authority to decide questions of authenticity for deepfake evidence has raised concerns that jurors may struggle to understand why they are entrusted with determining the authenticity of traditional evidence—such as handwriting, phone calls, or physical objects—but not generative artificial intelligence, potentially undermining perceptions of procedural fairness unless carefully explained.[83] However, courts already engage in threshold admissibility determinations under FRE 104(a) without encroaching on the jury's role. Under FRE 901(a), the proponent of evidence must offer a foundation sufficient for a reasonable juror to find the evidence authentic.[84] But when there are serious concerns about fabrication—especially with generative AI—the question is not just whether the evidence is "more likely than not" authentic but whether it is so unreliable that no reasonable jury should consider it. Unlike traditional authentication issues (e.g., "Is this a real signature?" or "Did this witness write this letter?"), deepfakes can be so sophisticated that even expert witnesses may struggle to determine authenticity.[85] Applying Rule 104(a), the court is better suited to make an initial legal determination about whether the evidence meets minimum reliability standards.

Although reallocating authenticity determinations introduces additional judicial responsibilities, existing judicial training infrastructures readily accommodate these requirements, enhancing trial fairness. Courts routinely assess reliability in contexts such as: Expert testimony, where judges decide whether

[81] Delfino, *supra* note 1 at pp. 354-56.
[82] Eur. Comm'n, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence* (Artificial Intelligence Act), COM (2021) 206 final (Apr. 21, 2021).
[83] "Another concern is about how the jury will react when it is instructed to presume authenticity. . . . It could become especially confusing when the jury is told that authenticity is a question primarily for jurors when it comes to telephone calls, diaries, and physical evidence, but when it comes to videos --- hands off." See Daniel J. Capra, *Memorandum To: Advisory Committee on Evidence Rules, Re: "Deepfakes" and Possible Amendments to Article 9 of the FRE*, p. 8 October 1, 2023
[84] Fed. R. Evid. 901(a).
[85] Delfino, *supra* note 1 at pp. 333-35.

expert opinions are sufficiently reliable before they ever reach a jury;[86] hearsay exceptions, where judges determine whether an out-of-court statement meets the criteria for admissibility before a jury can consider its weight;[87] and the best evidence rule, where a judge determines whether the evidence meets the rule's requirements when the parties dispute whether original writing exists.[88] Just as jurors do not question why they are not required to decide whether an expert's testimony is admissible or whether a hearsay statement qualifies for an exception, they would not be expected to question why an obviously unreliable piece of AI-generated evidence was excluded before they could evaluate its weight. Furthermore, potential jury confusion can be managed with jury instructions. Judges can provide limiting instructions to clarify that their role is only to ensure that evidence meets basic reliability standards, not to determine the ultimate truth of the evidence.[89]

The reallocation of the admissibility determination to the court maintains the integrity of the trial process. If the jury were left to decide whether highly questionable evidence is authentic, there is a significant risk that jurors would be misled by sophisticated deepfakes, undermining the fairness of the trial. Courts must exercise their gatekeeping function to protect the integrity of the fact-finding process—just as they do with coerced confessions, unreliable expert testimony, or improperly obtained evidence.

CONCLUSION

The Revised Proposal for FRE 901(c) offers a necessary and balanced solution to the challenges posed by AI-generated evidence. It ensures that digital evidence is authenticated and reliable before admission, prevents fraudulent AI-generated content from misleading jurors, and establishes a clear procedural framework for courts. Finally, by reallocating authenticity determinations to the court under FRE 104(a), this amendment aligns with existing judicial safeguards against prejudicial and unreliable evidence. The Revised Proposal embodies a necessary modernization of the Federal Rules of Evidence in response to the evolving threat of generative AI falsifications.

[86] Fed. R. Evid. 702.
[87] Fed. R. Evid. 803 & 804.
[88] Fed. R. Evid. 1008.
[89] For example: "The court has determined that certain evidence did not meet the necessary reliability requirements to be considered. You should not speculate about evidence that was not admitted but evaluate all admitted evidence impartially."