Fordham Law Review

Volume 92 | Issue 6 Article 7

2024

Deepfakes Reach the Advisory Committee on Evidence Rules

Daniel J. Capra
Fordham University School of Law

DEEPFAKES REACH THE ADVISORY COMMITTEE ON EVIDENCE RULES

*Daniel J. Capra**

INTRODUCTION

A number of articles have been written in the last couple of years about the evidentiary challenges posed by "deepfakes"—inauthentic videos and audios generated by artificial intelligence (AI) in such a way as to appear to be genuine.[1] You are probably aware of some of the widely distributed examples, such as: (1) Pope Francis wearing a Balenciaga jacket;[2] (2) Jordan Peele's video showing President Barack Obama speaking and saying things

* Reed Professor of Law, Fordham Law School; Reporter to the Judicial Conference Advisory Committee on Evidence Rules. This Essay was prepared as a companion to the Philip D. Reed Lecture held on October 27, 2023, at the University of St. Thomas School of Law, under the sponsorship of the Judicial Conference Advisory Committee on Evidence Rules.

[1] *See, e.g.*, 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); Russell Spivak, *"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339 (2019).

[2] *See* Simon Ellery, *Fake Photos of Pope Francis in a Puffer Jacket Go Viral, Highlighting the Power and Peril of AI*, CBS NEWS (Mar. 28, 2023, 11:39 AM), https://www.cbsnews.com/news/pope-francis-puffer-jacket-fake-photos-deepfake-power-peril-of-ai/ [https://perma.cc/TH5D-JYTZ].

that President Obama never said;[3] (3) Nancy Pelosi speaking while appearing to be intoxicated;[4] and (4) Robert DeNiro's de-aging in *The Irishman*.[5]

The evidentiary risk posed by deepfakes is that a court might find a deepfake video to be authentic under the mild standards of Rule 901 of the Federal Rules of Evidence,[6] that a jury may then think that the video is authentic because of the difficulty of uncovering deepfakes, and that all this will lead to an inaccurate result at trial. The question for the Advisory Committee on Evidence Rules (the "Committee") is whether Rule 901 in its current form is sufficient to guard against the risk of admitting deepfakes (with the understanding that no rule can guarantee perfection) or whether the rules should be amended to provide additional and more stringent authenticity standards to apply to deepfakes.

At the fall 2023 Committee meeting, Professor Maura R. Grossman and Judge Paul Grimm (former district judge for the U.S. District Court for the District of Maryland and now the Director of the Bolch Judicial Institute at Duke Law School) made a helpful and incisive presentation on deepfakes and proposed an amendment to Rule 901.[7] This short Essay provides (1) a brief introduction to deepfakes; (2) a short description of how Rule 901 operates; (3) a description of the Committee's review of social media and digital communication—the previous technological developments that challenged the evidence rules on authentication; and (4) a description of the Grimm-Grossman proposal to add a new provision to Rule 901 that will provide a procedure for assessing deepfakes, as well as of two other suggestions for change made in recent law review articles.

[3] *See* Stuart A. Thompson & Sapna Maheshwari, *'A.I. Obama' and Fake Newscasters: How A.I. Audio Is Swarming TikTok*, N.Y. TIMES (Oct. 13, 2023), https://www.nytimes.com/2023/10/12/technology/tiktok-ai-generated-voices-disinformation.html [https://perma.cc/FMY8-7QZY].

[4] *See* Amanda del Castillo, *Altered Videos of Speaker Nancy Pelosi Slurring Words Goes Viral*, ABC7NEWS (May 24, 2019), https://abc7news.com/deep-fake-deepfake-ai-reddit-nancy-pelosi/5315149/ [https://perma.cc/4FWC-5YFV].

[5] *See* Jordan Crucchiola, *Robert De Niro Reacts to* The Irishman *De-aging: 'It Was Okay,'* VULTURE (Oct. 25, 2019), https://www.vulture.com/2019/10/robert-de-niro-the-irishman-cgi-de-aging.html [https://perma.cc/E8J9-QNAD].

[6] *See* FED. R. EVID. 901 ("[T]he proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is.").

[7] *See* ADVISORY COMM. ON EVIDENCE RULES, MINUTES OF THE MEETING OF OCTOBER 27, 2023, at 97 (2023), https://www.uscourts.gov/sites/default/files/2023-10_evidence_rules_agenda_book_final_10-5.pdf [https://perma.cc/MN5L-2HTD]. Professor Grossman and Judge Grimm have written several important articles about deepfakes and about AI more broadly. *See* Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. 9, 84 (2021); Maura R. Grossman, Paul W. Grimm & Daniel G. Brown, *Is Disclosure and Certification of the Use of Generative AI Really Necessary?*, 107 JUDICATURE, no. 2, 2023, at 68; Maura R. Grossman, Paul W. Grimm, Daniel G. Brown & Molly (Yiming) Xu, *The GPTJudge: Justice in a Generative AI World*, 23 DUKE L. & TECH. REV. 1, 9 (2023).

I. THE PROBLEM OF DEEPFAKES

A deepfake is an inauthentic audiovisual presentation prepared by software programs using AI.[8] Of course, photos and videos have always been subject to forgery, but developments in AI make deepfakes much more difficult to detect.[9] Software for creating deepfakes is already freely available online and fairly easy for anyone to use.[10] As the software's usability and the videos' apparent genuineness keep improving over time, it will become harder for computer systems, much less lay jurors, to tell real from fake.[11]

Generally speaking, there is an arms race between deepfake technology and the technology that can be employed to detect deepfakes. Deepfakes involve machine-learning algorithms that are simultaneously pitted against one another.[12] One of these programs is a generative model that creates new data samples; the other, known as a discriminator model, evaluates this data

[8] *See* Meredith Somers, *Deepfakes, Explained*, MIT SLOAN SCH. MGMT. (July 21, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained [https://perma.cc/T7J7-Z2QT].

[9] *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1759–60 (2019). Some of the famous deepfakes are pretty easy to root out with minimal inquiry. The Nancy Pelosi video was debunked simply by playing it more slowly. The Pope picture, upon scrutiny, shows up as a fake because his medal is not sitting on his chest and his fingers are not accurate. *See supra* notes 1–2. But it is very likely that future developments will make deepfakes harder to detect.

[10] *See* Arjun Sha, *12 Best Deepfake Apps and Websites You Can Try for Fun*, BEEBOM, https://beebom.com/best-deepfake-apps-websites [https://perma.cc/4973-YWDF] (Apr. 3, 2024).

[11] MIT has provided a checklist that can be used to help detect a deepfake, though MIT makes no promises:

> When it comes to AI-manipulated media, there's no single tell-tale sign of how to spot a fake. Nonetheless, there are several DeepFake artifacts that you can be on the lookout for.
>
> 1. Pay attention to the face. High-end DeepFake manipulations are almost always facial transformations.
>
> 2. Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? DeepFakes may be incongruent on some dimensions.
>
> 3. Pay attention to the eyes and eyebrows. Do shadows appear in places that you would expect? DeepFakes may fail to fully represent the natural physics of a scene.
>
> 4. Pay attention to the glasses. Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, DeepFakes may fail to fully represent the natural physics of lighting.
>
> 5. Pay attention to the facial hair or lack thereof. Does this facial hair look real? DeepFakes might add or remove a mustache, sideburns, or beard. But, DeepFakes may fail to make facial hair transformations fully natural.
>
> 6. Pay attention to facial moles. Does the mole look real?
>
> 7. Pay attention to blinking. Does the person blink enough or too much?
>
> 8. Pay attention to the lip movements. Some deepfakes are based on lip syncing. Do the lip movements look natural?

Detect DeepFakes: How to Counteract Misinformation Created by AI, MIT MEDIA LAB, https://www.media.mit.edu/projects/detect-fakes/overview/ [https://perma.cc/P6GX-B297] (last visited Apr. 3, 2024).

[12] Chris V. Nicholson, *A Beginner's Guide to Generative AI*, PATHMIND, https://pathmind.com/wiki/generative-adversarial-network-gan [https://perma.cc/JEY9-K283] (last visited Apr. 3, 2024).

against a training dataset for authenticity.[13] The discriminator model estimates the probability that the sample came from the generative model (a machine creation) or sample data (a real-world original).[14] These two models operate in a cyclical fashion and learn from each other.[15] The generative model program is learning to create false data, and the discriminator model is learning to identify whether the data is artificial.[16] The generative model constantly improves its ability to create datasets that have a lower probability of failing the detection algorithm as the discriminator model learns to keep up, a process that continuously improves the apparent genuineness of the creation.[17] Therefore, any time new software is developed to detect fakes, deepfake creators can use that to their advantage in their discriminator models. A *New York Times* report reviewed some of the currently available programs that try to detect deepfakes.[18] The programs varied in their accuracy.[19] None were accurate 100 percent of the time.[20]

It should be noted that various digital tools have been introduced for authenticating video recordings that a party has prepared. These tools allow the proffering party to vouch for video recordings' authenticity through an electronic seal of approval.[21] Although the use of such methods increases the costs of litigation, they do appear to answer any "deepfake" claim from

[13] *See id.*

[14] *See id.*

[15] *See id.*

[16] *See id.*

[17] *See id.*

[18] *See* Stuart A. Thompson & Tiffany Hsu, *How Easy Is It to Fool A.I.-Detection Tools?*, N.Y. TIMES (June 28, 2023), https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html [https://perma.cc/W7XG-DVSC].

[19] *See id.*

[20] *See id.*; *see also* Tiffany Hsu & Steven Lee Myers, *Another Side of the A.I. Boom: Detecting What A.I. Makes*, N.Y. TIMES (May 18, 2023), https://www.nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html [https://perma.cc/RMZ6-2PMG] ("Detection tools inherently lag behind the generative technology they are trying to detect. By the time a defense system is able to recognize the work of a new chatbot or image generator, like Google Bard or Midjourney, developers are already coming up with a new iteration that can evade that defense. The situation has been described as an arms race or a virus-antivirus relationship where one begets the other, over and over.").

[21] WITNESS, TICKS OR IT DIDN'T HAPPEN: CONFRONTING KEY DILEMMAS IN AUTHENTICITY INFRASTRUCTURE FOR MULTIMEDIA 6 (2019), https://lab.witness.org/ticks-or-it-didnt-happen/ [https://perma.cc/87U9-ZLZK] ("The idea is that if you cannot detect deepfakes, you can, instead, authenticate images, videos and audio recordings at their moment of capture."); Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 B.U. PUB. INT. L.J. 245, 268 (2020) ("So-called 'verified media capture technology' can help 'to ensure that the evidence [users] are recording . . . is trusted and admissible to courts of law.' For example, an app called eyeWitness to Atrocities 'allows photos and videos to be captured with information that can firstly verify when and where the footage was taken, and secondly can confirm that the footage was not altered,' all while the company's 'transmission protocols and secure server system . . . create[] a chain of custody that allows this information to be presented in court.' That information, paired with the app-maker's willingness to provide a certification to the court or send a witness to testify if needed, could satisfy a court that the video is admissible, even if the videographer is unavailable." (alterations in original) (quoting WITNESS, *supra*, at 22–30)).

the opponent.[22] The limitation on the software is that the electronic stamp of genuineness occurs during the process in which the video is being generated; it does not work with videos taken off of the internet, for example.[23]

Besides the challenge of determining whether a video is fake, some commentators are concerned about a "*CSI* effect."[24] Jurors, aware of phenomena such as deepfakes and "fake news," may start expecting the proponent of a video to use sophisticated technology to prove to their satisfaction that the video is not fake.[25] Another concern is that over time, skepticism over video evidence may undermine the use of perfectly authentic videos—though how that concern is to be addressed in an evidence rule is a mystery.[26]

II. BASIC RULES ON AUTHENTICITY

Under Rule 901(a), the standards for authenticity are low. The proponent must only "produce evidence sufficient to support a finding that the item is what the proponent claims it is."[27] Under the rule, the question of authenticity is one of conditional relevance—an item of evidence is not relevant unless it is what the proponent purports it to be.[28] For example, a sexually harassing statement in an email, purportedly sent from the plaintiff's supervisor, is probative only if it is the supervisor who sent it. As a question of conditional relevance, the admissibility standard under Rule 901 is the same as that provided by Rule 104(b): Has the proponent offered a foundation from which the jury could reasonably find that the evidence is what the proponent says it is?[29] This is a mild standard—favorable to admitting the evidence. Authenticity should generally be a jury question because, if a juror finds the item to be inauthentic, it just drops from the case, so no real damage is done.[30] Rule 901 basically operates to prevent the jury from wasting its time evaluating an item of evidence that clearly is not what the proponent claims it to be.[31]

[22] *See* WITNESS, *supra* note 21, at 16.

[23] *See, e.g.*, Lily Hay Newman, *A New Tool Protects Videos from Deepfakes and Tampering*, WIRED (Feb. 11, 2019, 5:15 PM), https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/ [https://perma.cc/NPQ4-PXND] ("Called Amber Authenticate, the tool is meant to run in the background on a device as it captures video. At regular, user-determined intervals, the platform generates 'hashes'—cryptographically scrambled representations of the data—that then get indelibly recorded on a public blockchain. If you run that same snippet of video footage through the algorithm again, the hashes will be different if anything has changed in the file's audio or video data—tipping you off to possible manipulation.").

[24] *See* Delfino, *supra* note 1, at 338 n.310.

[25] *See id.*

[26] *See id.* at 297.

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

[28] *See id.* 901(b)(1).

[29] *Id.* 104(b).

[30] *See Rule 901—Authenticating or Identifying Evidence: Summary and Explanation*, FED. R. EVID., https://www.rulesofevidence.org/fre/article-ix/rule-901/ [https://perma.cc/ZV88-KNPF] (last visited Apr. 3, 2024).

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

The authenticity rules operate as follows: First, Rule 901(a) sets the general standard for authenticity—enough admissible evidence for a juror to believe that the proffered item is what the proponent says it is.[32] Second, Rule 901(b) provides examples of sufficient authentication; if the standard set forth in any of the illustrations is met, then the authenticity objection is overruled and any further question of authenticity is for the jury.[33] Third, the illustrations are not intended to be independent of each other, so a proponent can establish authenticity through a single factor or combination of factors in any particular case. Finally, Rule 902 provides certain situations in which the proffered item will be considered self-authenticating—no reference to any Rule 901(b) illustration need be made or satisfied if the item is self-authenticating.[34]

In order for the trier of fact to make a rational decision as to authenticity, the foundation evidence must itself be admissible.[35] If the opponent still contests authenticity at trial, the proponent will need to present admissible evidence of the authenticity of the challenged item.[36] When an item is offered, a jury must be provided sufficient admissible evidence for it to find that the evidence is what the proponent claims, or the requirement of authentication is not satisfied.[37] A judgment by the court as to whether a reasonable jury will find evidence to be authentic can only be made by examining the evidence that the jury will be permitted to hear.[38]

Applying the current authentication rules to deepfakes raises at least two concerns. On the one hand, because deepfakes are hard to detect, many deepfakes will probably satisfy the low standards of authenticity; on the other hand, the prevalence of deepfakes will lead to blanket claims of forgery, requiring courts to have an authenticity hearing for virtually every proffered video.

III. PRIOR COMMITTEE DECISION ON SPECIAL AUTHENTICATION RULES FOR ELECTRONIC EVIDENCE

The rise of deepfakes is not the only technological advancement that has challenged the existing rules on authentication. In 2014, the Committee undertook a project to consider whether rules should be added to Article IX of the Federal Rules of Evidence to address digital communications and

[32] *See id.* 901(a).

[33] *See id.* 901(b).

[34] *See id.*; *id.* 902.

[35] *See id.* 104(b).

[36] *See, e.g.*, United States v. Bonds, 608 F.3d 495, 508 (9th Cir. 2010) (holding that records could not be authenticated when the only basis for authentication was a hearsay statement that was not admissible under any exception); Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 537 (D. Md. 2007) ("Because, under Rule 104(b), the jury, not the court, makes that factual findings that determine admissibility, the facts introduced must be admissible under the rules of evidence.").

[37] *See Lorraine*, 241 F.R.D. at 540.

[38] *See Bonds*, 608 F.3d at 508; *Lorraine*, 241 F.R.D. at 540.

social media postings.[39] The proposal considered was to have special rules on authenticating emails, texts, social media postings, and so forth.[40] After significant discussion, the Committee decided not to proceed with the project.[41] According to the meeting minutes of the fall 2014 meeting, the reasons for rejection were as follows.[42]

First, the current rules are flexible enough to handle questions about the authenticity of digital communications. For digital evidence, the most useful authentication rules within Rule 901(b) are: 901(b)(1), which allows a witness with personal knowledge to testify that the evidence is what it purports to be;[43] 901(b)(3), which allows comparison of the evidence "with an authenticated specimen by an expert witness or the trier of fact";[44] 901(b)(4), which allows evidence of the "appearance, contents, substance, internal patterns, or other distinctive characteristics of the item, taken together with all the circumstances";[45] 901(b)(5), which allows "an opinion identifying a person's voice—whether heard firsthand or through electronic transmission or recording—based on" having heard that voice in the past;[46] and 901(b)(9), which allows "evidence describing a process or system and showing that it produces an accurate result."[47] These rules give the court all the tools it needs to determine the authenticity of digital evidence.[48]

Second, any rules directed specifically toward digital communications would likely overlap with the provisions already in Rule 901(b).[49] Certainly, distinctive characteristics would be important for authenticating digital evidence; further, authentication of email, for example, would use analogous principles of authenticating telephone conversations. This overlap between new and old rules would likely cause confusion.

Third, listing factors relevant to authentication would run the risk of misleading courts and litigators into thinking that all of the listed factors can or should be weighed equally, when in fact a case-by-case approach is required.[50]

Finally, given the deliberateness of rulemaking—which takes three years minimum—there was a risk that any rule on digital communications could be dead on arrival.[51] I called it the MySpace problem.

[39] *See* ADVISORY COMM. ON EVIDENCE RULES, MINUTES OF THE MEETING OF OCTOBER 24, 2014, at 8–12 (2014), https://www.uscourts.gov/sites/default/files/minutes_oct_2014_evidence_committee_0.pdf [https://perma.cc/FW5B-FYFS].

[40] *See id.*

[41] *See id.* at 9.

[42] *See id.* at 8–12.

[43] *See* FED. R. EVID. 901(b)(1).

[44] *See id.* 901(b)(3).

[45] *See id.* 901(b)(4).

[46] *See id.* 901(b)(5).

[47] *See id.* 901(b)(9).

[48] *See* ADVISORY COMM. ON EVIDENCE RULES, *supra* note 39, at 9.

[49] *See id.*

[50] *See id.*

[51] *See id.* It should be noted that the Committee did propose two new rules to deal with authenticating digital evidence—Rules 902(13) and 902(14), which became effective in 2017. *See id.* at 12. But these rules do not add or change any grounds of authentication for digital

In hindsight, it is fair to state that the Committee's decision to forego amendments setting forth specific grounds for authenticating digital evidence was the prudent course. Courts have sensibly, and without extraordinary difficulty, applied the grounds of Rule 901 to determine the authenticity of digital evidence.[52] Courts have specifically rejected blanket claims like "my account was hacked" because such an argument can always be made.[53] Courts properly require some showing from the opponent before inquiring into charges of hacking and falsification of digital information.[54] Thus, courts have consistently held that the mere allegation of fabrication "does not and cannot be the basis for excluding [electronically stored information] as unidentified or unauthenticated as a matter of course, any more than it can be the rationale for excluding paper documents."[55]

It is true that litigators have to know what they are doing when they try to authenticate digital evidence, and it is also true that authenticating digital evidence can be costly, but no rule of evidence would change that.[56]

evidence. *See id.* at 11–12. Rather, they allow the existing grounds to be established by a certificate of a person with knowledge, thus dispensing with the requirement of in-court testimony. *See id.*

[52] *See, e.g.*, United States v. Fluker, 698 F.3d 988, 999 (7th Cir. 2012) (outlining the variety of ways in which an email could be authenticated and stating that testimony from a witness who purports to have seen the declarant create the email in question was sufficient for authenticity under Rule 901(b)(1)); United States v. Lundy, 676 F.3d 444, 452–54 (5th Cir. 2012) (finding that testimony by one party to an electronic chat conversation that the chats were as he recorded them was enough to meet the low threshold for authentication); United States v. Needham, 852 F.3d 830, 836 (8th Cir. 2017) ("Exhibits depicting online content may be authenticated by a person's testimony that he is familiar with the online content and that the exhibits are in the same format as the online content. Such testimony is sufficient to provide a rational basis for the claim that the exhibits properly represent the online content. . . . [The witness] testified that he personally viewed the [webpages] and that the screenshots accurately represented the online content of both sites. Thus, the district court did not abuse its discretion by admitting the screenshots."); United States v. Recio, 884 F.3d 230, 237 (4th Cir. 2018) ("[T]he government sufficiently tied [the] 'Facebook User' to [the defendant] by showing that: (1) the user name associated with the account was 'Larry Recio,' (2) one of the four email addresses associated with the account was 'larryrecio20@yahoo.com,' (3) more than 100 photos of Recio were posted to the account, and (4) one of the photos posted to the user's timeline was accompanied by the text 'Happy Birthday Larry Recio.'"). In *United States v. Barnes*, the court found that the government had laid a proper foundation to authenticate Facebook and text messages as having been sent by the defendant. 803 F.3d 209, 217 (5th Cir. 2015). The defendant was a quadriplegic, but the witness who received the messages testified that she had seen the defendant use Facebook, that she recognized his Facebook account, and that the Facebook messages matched the defendant's "manner of communicating." *Id.* "Although she was not certain that [the defendant] authored the messages, conclusive proof of authenticity is not required for admission of disputed evidence." *Id.*

[53] *See supra* note 52 and accompanying text; *see also* Paul W. Grimm, Daniel J. Capra & Gregory P. Joseph, *Authenticating Digital Evidence*, 69 BAYLOR L. REV. 1, 7–9 (2017).

[54] *See supra* note 52 and accompanying text.

[55] United States v. Safavian, 435 F. Supp. 2d 36, 41 (D.D.C. 2006).

[56] *See* Jeffrey Bellin & Andrew Guthrie Ferguson, *Trial by Google: Judicial Notice in the Information Age*, 108 NW. U. L. REV. 1137, 1157 (2014) ("Although much is made of [the authentication] hurdle in the Information Age, it is . . . an easy one to surmount. Success generally depends not on legal or factual arguments, but rather the amount of time and resources a litigant devotes to the problem.").

Moreover, some costs of proving authenticity can be saved by the affidavit procedures established for authentication of digital evidence in Rules 902(13) and 902(14).[57]

The fact that the Committee decided not to promulgate special rules on digital communication is a relevant data point, but it is not necessarily dispositive of amending the rules to treat deepfakes.[58] Although a special rule setting forth the grounds for possible authentication of audiovisual evidence runs a similar risk of overlap, perhaps a rule of procedure (such as the requirement of a special showing made to the court or a notice requirement) or a higher standard of proof could be useful. It is for the Committee to determine whether it is interested in exploring such a procedural alternative.

IV. CALLS FOR CHANGE

There are several calls for change to the authenticity rules to deal with the rise of deepfakes. This part discusses two suggestions made in law review articles, as well as a third suggestion from Professor Grossman and Judge Grimm.

A. Allocating Responsibility to the Court

Professor Rebecca Delfino argues that the danger of deepfakes demands that the judge decide authenticity, not the jury.[59] She contends that "[c]ountering juror skepticism and doubt over the authenticity of audiovisual images in the era of fake news and deepfakes calls for reallocating the factfinding authority to determine the authenticity of audiovisual evidence."[60] She argues that jurors cannot be trusted to fairly analyze whether a video is a deepfake because deepfakes appear to be genuine and "seeing is believing."[61] Professor Delfino suggests that Rule 901 should be amended to add a new subdivision (c), which would provide:

> 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.[62]

[57] *See* WITNESS, *supra* note 21, at 30 (noting that Rules 902(13) and 902(14) "streamlin[e] authentication for those with limited legal resources").

[58] For one thing, stare decisis does not apply in this context. Examples of amendments that were previously rejected include the 2023 amendment to Rule 106 and the 2024 amendment that will add a new Rule 107. *See generally* ADVISORY COMM. ON EVIDENCE RULES, *supra* note 7, at 319, 321 (noting approval of Rules 106 and 107). Also, perhaps the dangers of fakery are greater with respect to deepfakes than were presented by digital evidence in 2014.

[59] *See* Delfino, *supra* note 1, at 341–48.

[60] *Id.* at 341.

[61] *Id.* at 307.

[62] *Id.* at 341.

She explains that the new Rule 901(c) "would relocate the authenticity of digital audiovisual evidence from Rule 104(b) to the category of relevancy in Rule 104(a)" and would "expand the gatekeeping function of the court by assigning the responsibility of deciding authenticity issues solely to the judge."[63]

The proposed rule would operate as follows: After the pretrial hearing to determine the authenticity of the evidence, if the court finds that the item is more likely than not authentic, the court admits the evidence.[64] The court would instruct the jury that it *must accept as authentic* the evidence that the court has determined is genuine.[65] The court would also instruct the jury not to doubt the authenticity simply because of the existence of deepfakes.[66] This new rule would take the jury out of the business of determining authenticity, "thereby avoiding the problems invited by juror distrust and doubt."[67] Finally, "the court would address the threat of counsel exploiting juror doubts over the authenticity of evidence using the deepfake defense by ordering counsel not to make such arguments."[68]

It should be noted that the Delfino proposal applies to *all* audiovisual evidence—including the video evidence that courts have been dealing with for about 100 years. Query whether the threat of deepfakes warrants such a dramatic change with respect to all video evidence. Assuming that any amendment is necessary, perhaps the goal is to set out procedures and higher standards—but only after the opponent specifically brings a credible deepfake argument.

Another concern is about how the jury will react when it is instructed to presume authenticity. Given the presence of deepfakes in society, it may well be that jurors will do their own assessment—regardless of the instruction—and that juror assessment will be done without the foundation for authenticity laid by the proponent in the admissibility hearing.[69] 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.

One can argue that the Delfino proposal could productively be cut in half. That is, apply the Rule 104(a) standard to the authenticity of visual evidence, but then allow the jury to make its own assessment. In other words, treat the authenticity of visual evidence the same way we treat expert testimony.[70] Delfino would object, though, due to her belief that jurors will not be able to assess the genuineness of the evidence given that deepfakes are getting harder

[63] *Id.*

[64] *See id.*

[65] *See id.* at 341–42.

[66] *See id.* at 342.

[67] *Id.*

[68] *Id.*

[69] *See id.* at 336–37.

[70] *See* FED. R. EVID. 702 advisory committee's note to 2023 amendment (imposing a preponderance standard for admissibility but leaving the jury to evaluate expert testimony if the preponderance standard is met).

and harder to detect.[71] But this half-proposal would at least address arguments that deepfakes will be too easily admitted under the mild standard for showing authenticity to the court. And it would not take away the jury's traditional role of evaluating admissible evidence.

Finally, Delfino's idea is that the court is to use the Rule 104(a) standard—a preponderance of the evidence.[72] Assuming that this is appropriate, it should be added to the text of the rule—that is a lesson that was learned by the Committee in the amendment to Rule 702.[73] Accordingly, the last sentence of the proposal should read something like the following: "The court must decide whether it is more likely than not that the item is authentic."

Such an explication is especially important because Delfino's proposal does not explicitly say that admissibility is governed by Rule 104(a). It states that "the proponent must produce evidence that the item is what the proponent claims it is in accordance with subdivision (b)."[74] But the illustrations of subdivision (b) are, as discussed above, decided on the less rigorous, prima facie proof standard of Rule 104(b).[75]

B. A Corroboration Requirement

John LaMonaca argues for a more stringent standard of authenticity with respect to deepfakes.[76] He contends that the traditional means of authentication—by a person with knowledge under Rule 901(b)(1)—will no longer work with deepfakes because a witness cannot reliably testify that the video accurately represents reality.[77] He states that "[b]ecause witnesses will no longer be able to meet the legacy standard of Rule 901(b)(1)'s knowledgeable witness by attesting that a video is a fair and accurate portrayal, courts need to look elsewhere for a sufficient finding that photographic evidence is what its proponent claims it is."[78] He argues for a proposed new Rule 901(b)(11) that would specifically govern "the unique challenges that digital photography in the modern age present."[79] The new Rule 901(b)(11) would provide: "Before a court admits photographic

[71] *See* Delfino, *supra* note 1, at 307.

[72] *See* FED. R. EVID. 702; *see also* Delfino, *supra* note 1, at 348.

[73] The 2000 amendment to Rule 702 imposed reliability requirements on expert testimony but did not specify the standard of proof to be employed by the judges. *See* FED. R. EVID. 702 advisory committee's note to 2023 amendment. Many courts ended up holding that the admissibility requirements in the rule are actually questions of weight. *See id.* The 2023 amendment clarifies the applicable standard of proof by requiring in text that the proponent has the burden of showing that the requirements are met "more likely than not." *Id.*

[74] Delfino, *supra* note 1, at 341.

[75] *See* FED. R. EVID 104(b).

[76] John P. LaMonaca, Note, *A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes*, 69 AM. U. L. REV. 1945, 1984 (2020).

[77] *See id.* at 1977.

[78] *Id.* at 1984.

[79] *Id.* at 1985.

evidence under this rule, a party may request a hearing requiring the proponent to corroborate the source of information by additional sources."[80]

LaMonaca explains that the new rule "essentially codifies an existing means of authentication and requires it for photographic evidence."[81] There is no proposal to change the existing allocation of authority between the court and the jury. Rather, what it essentially does is (1) change the "distinctive characteristics" ground of Rule 901(b)(4) into a foundation *requirement* and (2) state that the classic ground of authentication under Rule 901(b)(1)—that the video accurately represents what it purports to show—is never a sufficient ground of admissibility. LaMonaca concludes that "a preliminary hearing process [requiring corroboration] would bolster the confidence in video evidence for a jury to consider, rather than allowing all photographic evidence to pass the foundational stage with a testimonial witness who lacks the requisite personal knowledge to attest to the evidence's validity."[82]

This is an interesting proposal, in that one of the major ways that deepfakes can be *debunked* is actual evidence casting doubt on what is portrayed. For example, a witness could state: "The video shows me at the bank but I was in the hospital that day." So, it might not be asking too much for a proponent to provide some corroboration of the event if there is a legitimate question of authenticity. But one major problem is that, like the Delfino proposal, LaMonaca's proposal applies to *all* visual evidence, including video evidence that has been well-handled by the courts for 100 years. It seems unwarranted to require the proponent to go to the expense of providing corroboration for every surveillance video and every wedding photograph, simply because of the potential risk of deepfakes. Courts have not required an advance showing of corroboration for digital evidence and, although deepfakes present new challenges, the case has not yet been made to justify an automatic corroboration requirement for all photographic evidence.

The better solution is the reverse—that the court should entertain a deepfake inquiry only when the proponent provides some evidence indicating the possibility of a deepfake—either some electronic analysis or a showing through evidence that the event presented is implausible. And then, at that point, the proponent would be required to provide corroboration or some other additional showing before the court can find it authentic. That reverse solution is essentially employed today with regard to electronic evidence—the "it was hacked" claim is not treated seriously until the opponent comes up with something to indicate that an inquiry is warranted.[83]

[80] *Id.* Professor Agnieszka McPeak makes a similar argument that circumstantial evidence should be presented to the jury "whenever the jury is asked to ascertain the authenticity of digital video and audio evidence." Agnieszka McPeak, *The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood*, 23 VAND. J. ENT. & TECH. L. 433, 450 (2021).

[81] LaMonaca, *supra* note 76, at 1985.

[82] *Id.* at 1987–88.

[83] *See* Paul W. Grimm, Lisa Yurwit Bergstrom & Melissa M. O'Toole-Loureiro, *Authentication of Social Media Evidence*, 36 AM. J. TRIAL ADVOC. 433, 459 (2013) ("A trial judge should admit the evidence if there is plausible evidence of authenticity produced by the

And that solution—placing the burden of going forward on the opponent—is what was employed in one of the few court cases that has discussed the deepfake possibility. The Colorado Court of Appeals in *Colorado v. Gonzales*[84] stated that although software has made it easy for laypeople to manipulate recordings, "the fact that the falsification of electronic recordings is always possible does not, in our view, justify restrictive rules of authentication that must be applied in every case *when there is no colorable claim of alteration*."[85] The court explained that "[w]hen a plausible claim of falsification is made by a party opposing the introduction of a recording, the court may and usually should apply additional scrutiny" to determine whether a reasonable jury could conclude that the item is what it purports to be.[86]

There are two additional rulemaking points that should be made about the LaMonaca proposal. First, the suggested provision should not be placed as a new Rule 901(b)(11). Rule 901(b) provides examples of authenticated items.[87] This new provision purports to require an extra admissibility requirement for evidence that will be offered under an existing rule, such as Rule 901(b)(9). It is not a new example of authentication. Therefore, it is better placed as an addition to Rule 901(b)(9), like in the Grimm-Grossman proposal discussed below, or as a separate subdivision, such as Rule 901(c).[88] Second, the proposed rule refers to "photographic" evidence, which seems too narrow to cover all deepfakes. A term such as "audiovisual" is preferable.[89] The Grimm-Grossman proposal, discussed below, simply ties into Rule 901(b)(9), which governs items resulting from a process or system—this subsection is probably the best tie-in for deepfakes.

C. The Grimm-Grossman Proposal

Professor Grossman and Judge Grimm conclude that the existing authenticity rules are flexible enough to address any problems arising from deepfakes.[90] They see no need for a higher standard of proof at the admissibility level.[91] They do believe, however, that the difficulty in determining the authenticity of deepfakes justifies some procedural structure

proponent of the evidence and only speculation or conjecture—not facts—by the opponent of the evidence about how, or by whom, it 'might' have been created.").

[84] 474 P.3d 124 (Colo. App. 2019).

[85] *Id.* at 130 (emphasis added); *see also* Shannon Bond, *People Are Trying to Claim Real Videos Are Deepfakes. The Courts Are Not Amused*, NPR (May 8, 2023, 5:01 AM), https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused [https://perma.cc/CP74-MZU9] (noting that courts have rejected out-of-hand, broad claims that videos could be deepfakes).

[86] *Gonzales*, 474 P.3d at 130.

[87] *See* FED. R. EVID. 901(b).

[88] *See infra* note 93.

[89] *See* LaMonaca, *supra* note 76, at 1987.

[90] *Symposium on Scholars' Suggestions for Amendments, and Issues Raised by Artificial Intelligence*, 92 FORDHAM L. REV. 2375, 2430–32 (2024).

[91] *See id.* at 2434–34.

and protection at an admissibility hearing.[92] They propose an amendment to Rule 901(b)(9) that would provide as follows:

> (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 a reliable result; and
>
> (B) if the proponent concedes that—or the opponent provides a factual basis for suspecting that—the item was generated by artificial intelligence, additional evidence that:
>
> (i) describes the software or program that was used; and
>
> (ii) shows that it produced reliable results in this instance.[93]

This proposal provides a helpful way to structure an authenticity question in light of deepfakes. It imposes no safeguards in the first instance when a proponent seeks to admit an audiovisual item—meaning that the mere claim of "deepfake" by the opponent is treated as a nonevent. However, if the opponent provides a factual basis for believing that there is a deepfake or if the proponent concedes that AI has been used, the proponent must describe how the item was prepared and show that it is a reliable account of what it portrays.

The proposed procedural requirements would be placed in Rule 901(b)(9), which would be the rule under which an audiovisual presentation made with AI would probably have to be authenticated.[94] Though another possibility is to have a freestanding Rule 901(c), labeled something like "Procedures for Items Generated by Artificial Intelligence."

The proposal is also useful in emphasizing that the search is for *reliability*. The term "reliability" is used in Rule 702, and the same types of concerns posed by expert testimony arise when an item is prepared with AI—i.e., the jury will not be able to determine that a deepfake is inauthentic, so procedural safeguards are required at the admissibility level. Moreover, the essential problem of AI is that it leads to an unreliable presentation of an event.

Notably, one could combine the procedural requirements of the above proposal with the addition of a heightened standard of proof, such as a preponderance of the evidence. Obviously, the piling on of safeguards is dependent on the perceived degree of risks posed by deepfakes.

[92] *See id.* at 2433.

[93] *Id.* at 2431 n.203.

[94] Currently, Rule 901(b)(9) provides that authenticity of a process or system can be established if the proponent puts forth "[e]vidence describing a process or system and showing that it produces an accurate result." FED. R. EVID. 901(b)(9). Under the Grimm-Grossman proposal, this provision is retained as the first subdivision of an amended Rule 901(b)(9), with one important difference: the results of the system must be shown to be "reliable" rather than "accurate." *Symposium on Scholars' Suggestions for Amendments, and Issues Raised by Artificial Intelligence*, *supra* note 90, at 2431 n.203.

D. Another View: No Change Is Necessary

Not all commentators believe that a change to the rules is necessary for dealing with deepfakes. Riana Pfefferkorn notes that the courts have previously handled technological changes under the existing rules, and she argues that deepfakes can be handled in the same way.[95] She asserts that the courts are "no stranger to doctored photographs" and that "generations of technologies with truth-subversive potential have become commonplace in society over the years."[96] "While the resulting fakes have inevitably gained traction at times in the public consciousness, the sky has not fallen."[97] She states that "[t]he existence of the mere possibility of manipulation, without more, does not call for a high bar of authentication today any more than it did 150 years ago."[98] She concludes that "[t]he nation's courts are robust institutions that have shown themselves capable of handling each new variant of the age-old problem of fakery" and that the "[c]ourts' track record of resilience should assuage" much of the concern about deepfakes.[99] Pfefferkorn's view is that the rise of deepfakes will probably increase the costs of authentication, perhaps by requiring expert testimony in more cases than previously.[100] But that does not mean that the rules need to be amended.

Similarly, Grant Fredericks, the owner and operator of Forensic Video Solutions and a pioneer in the field of deepfake technology, is confident that fake videos will be kept out of evidence, both because they can be discovered using the advanced tools of his trade and because the video's proponent would be unable to answer basic questions (such as who created the video, when, and with what technology) to authenticate it.[101]

IV. CONCLUSION

The Committee must decide whether it is necessary to develop a change to the Federal Rules of Evidence in order to deal with deepfakes. If some rule is to be proposed, it probably should not be a specific rule setting forth the methods in which visual evidence can be authenticated—as those methods are already in Rule 901 and the overlap would be problematic. More

[95] *See* Pfefferkorn, *supra* note 21, at 259.

[96] *Id.* at 256, 258.

[97] *Id.* at 258.

[98] *Id.* at 266–67.

[99] *Id.* at 258; *see also* Russell Brandom, *Deepfake Propaganda Is Not a Real Problem*, The VERGE (Mar. 5, 2019, 12:25 PM), https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax [https://perma.cc/4R2S-NZ5G] ("We've had the tools to fabricate videos and photos for a long time. . . . AI tools can make that process easier and more accessible, but it's easy and accessible already. . . . [D]eepfakes are already in reach for anyone who wants to cause trouble on the internet. It's not that the tech isn't ready yet. . . . [I]t just isn't useful."); Jeffrey Westling, *Deep Fakes: Let's Not Go Off the Deep End*, TECHDIRT (Jan. 30, 2019, 12:05 PM), https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-gooff-deep-end.shtml [https://perma.cc/8NUZ-98T2].

[100] Pfefferkorn, *supra* note 21, at 309.

[101] Robin Hoffman, *Forensic Video Experts: Fake Videos Not Threat to Courtroom Evidence*, PIPELINE COMMC'NS (June 24, 2019), https://www.pipecomm.com/forensic-video-experts-fake-videos-not-threat-to-courtroom-evidence/ [https://perma.cc/6R8T-F3E6].

productive solutions include heightening the standard of proof or requiring an additional showing of reliability. However, any such requirements should kick in only after some showing by the opponent has been made. A contention such as "it might be a deepfake" or "deepfakes are easy to do" has to be a nonevent.

Any possible change to treat deepfakes must be evaluated with the perspective that the authenticity rules are flexible and have been flexibly and sensibly applied by the courts to treat other forms of technological fakery. Moreover, the Committee is well aware that any change has to be broad and general (and thus perhaps less helpful), as an amendment with specific standards and terminology is likely to be outmoded by the time it ever becomes effective. Chasing fast-developing technology with a rulemaking procedure that takes a minimum of three years is a challenge, to say the least.