Deepfakes in Court: How Judges Can Proactively Manage Alleged AI-Generated Material in National Security Cases

Abhishek Dalal[†]
Chongyang Gao[††]
Hon. Paul W. Grimm (ret.)[†††]
Maura R. Grossman[††††]
Daniel W. Linna Jr.[†††††]
Chiara Pulice[††††††]
V.S. Subrahmanian[†††††††]
Hon. John Tunheim[††††††††]

[†] Pritzker School of Law, Northwestern University.

[††] Northwestern University.

[†††] David F. Levi Professor of the Practice of Law & Director, Bolch Judicial Institute, Duke Law School.

[††††] Research Professor, David R. Cheriton School of Computer Science, University of Waterloo & Adjunct Professor, Osgoode Hall Law School, York University.

[†††††] Senior Lecturer and Director of Law and Technology Initiatives, Pritzker School of Law & McCormick School of Engineering, Northwestern University.

[††††††] Dept. of Computer Science & Buffett Institute for Global Affairs, Northwestern University.

[†††††††] Walter P. Murphy Professor of Computer Science, Buffett Faculty Fellow at the Buffett Institute for Global Affairs, Northwestern University.

[††††††††] United States District Court for the District of Minnesota.

ABSTRACT

Dall-E. ChatGPT. GPT-4. Words that did not exist in the English lexicon just a few years ago are now commonplace. With the widespread availability of Artificial Intelligence (AI) tools, specifically Generative AI, whether in the context of text, audio, video, imagery, or even combinations of these, it is inevitable that trials related to national security will involve evidentiary issues raised by Generative AI. We must confront two possibilities: first, that evidence presented is AI-generated and not real and, second, that other evidence is genuine but alleged to be fabricated. Technologies designed to detect AI-generated content have proven to be unreliable,[1] and also biased.[2] Humans have also proven to be poor judges of whether a digital artifact is real or fake.[3] There is no foolproof way today to classify text, audio, video, or images as authentic or AI-generated, especially as adversaries continually evolve their deepfake generation methodology to evade detection. Thus, the generation and detection of fake evidence will continue to be a cat-and-mouse game. These are not challenges of a far-off future; they are already here. Judges will increasingly need to establish best practices to deal with a potential deluge of evidentiary issues.

We will discuss the evidentiary challenges posed by Generative AI using a civil lawsuit hypothetical. The hypothetical describes a scenario involving a U.S. presidential candidate seeking an injunction against her opponent for circulating disinformation in the weeks leading up to the election. We address the risk that fabricated evidence might be treated as genuine and genuine evidence as fake. Through this scenario, we discuss the best practices that judges should follow to raise and resolve Generative AI issues under the Federal Rules of Evidence.

We will then provide a step-by-step approach for judges to follow when they grapple with the prospect of alleged AI-generated fake evidence. Under this approach, judges should go beyond a showing that the evidence is merely more likely than not what it purports to be. Instead, they must balance the risks of negative consequences that could occur if the evidence turns out to be fake. Our suggested approach ensures that courts schedule a pretrial evidentiary hearing far in advance of trial, where both proponents and opponents can make arguments on the admissibility of the evidence in question. In its ruling, the judge should only admit evidence, allowing the jury to decide its disputed authenticity, after considering under Rule 403 whether its probative value is substantially outweighed by danger of unfair prejudice to the party against whom the evidence will be used.[4] Our suggested approach thus illustrates how judges can protect the integrity of jury deliberations in a manner that is consistent with the current Federal Rules of Evidence and relevant case law.

---

[1] *See* Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, & Hafiz Malik, *Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward*, 53 APPLIED INTEL. 3984–3985 (June 2022).

[2] *See generally* Ying Xu, Philipp Terhörst, Kiran Raja, & Marius Pedersen, *Analyzing Fairness in Deepfake Detection With Massively Annotated Databases*, 5 IEEE TRANSACTIONS TECH. & SOC'Y 93 (2024).

[3] Nils C. Köbis, Barbora Doležalová, & Ivan Soraperra, *Fooled Twice: People Cannot Detect Deepfakes but Think They Can*, 11 ISCIENCE 24 (Nov. 2021).

[4] FED. R. EVID. 403.

I. INTRODUCTION

Deepfakes and other AI-generated materials (AIM) are no longer novelties. Deepfakes have entered popular discourse due to their use (or alleged use) in entertainment, war, elections,[5] and other settings. Until recently, only relatively experienced technologists could create AIM. But now, anyone with an Internet connection and basic technology skills can access online tools to generate convincing fabricated video, audio, image, and text materials. The quality of AIM is rapidly improving, such that we should expect that very soon nearly anyone will be able to create convincing fake materials. The public will not be able to identify the materials as fake, and even experts will struggle to accurately distinguish genuine materials from fake. While technological solutions such as watermarking have been proposed, many experts believe that there will not be a definitive technological solution to the deepfake problem—at least anytime soon—especially when deepfakes are created by sophisticated actors, including by state actors.[6]

Deepfakes will present new challenges for courts, particularly in high-stakes cases involving elections, foreign actors, and other matters of national security. Courts are well equipped to handle the evidentiary issues of the past—such as those posed by social media. Currently, parties proffer expert witnesses, judges act as gatekeepers to ensure that experts are qualified, and juries determine the credibility of expert and fact witnesses, "find facts," and provide verdicts. However, social science research suggests that even if a person is aware that evidence is AIM, the fake evidence may still have a substantial impact on the person's perception of the facts of a situation.[7]

A 2022 study described this phenomenon as the "continued influence effect."[8] According to the study, once information is encoded in the

---

[5] *See, e.g.*, Em Steck & Andrew Kaczynski, *Fake Joe Biden Robocall Urges New Hampshire Voters Not to Vote in Tuesday's Democratic Primary*, CNN (Jan. 22, 2024, 5:44 PM), https://www.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html [https://perma.cc/C3HU-UR9V].

[6] *Many AI Researchers Think Fakes Will Become Undetectable*, ECONOMIST (Jan. 17, 2024), https://www.economist.com/science-and-technology/2024/01/17/many-ai-researchers-think-fakes-will-become-undetectable [https://perma.cc/K6LQ-8FHU]; *see also* Siddarth Srinivasan, *Detecting AI Fingerprints: A Guide to Watermarking and Beyond*, BROOKINGS INST. (Jan. 4, 2024), https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ [https://perma.cc/Q69E-VP9F] ("Watermarking is the process of embedding an identifying pattern in a piece of media in order to track its origin.").

[7] Maura R. Grossman, et al., *The GPTJudge: Justice in a Generative AI World*, 23 DUKE L. & TECH. REV. 1, 19 (2023); *see also* 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, 25–27 (2023) (discussing studies showing impact of audiovisual evidence on juror perception and memory).

[8] Ullrich K. H. Ecker, et al., *The Psychological Drivers of Misinformation Belief and its Resistance to Correction*, 1 NATURE REVS. PSYCH. 1, 13–29 (2022).

memory, it remains in the memory to be reactivated and retrieved later.[9] When the information is corrected, the brain performs some knowledge revision, but that prior information is not simply erased but now "coexist[s] and compete[s] for activation."[10] The credibility of the purported source of misinformation may also influence how the fake evidence impacts a jury member. A 2020 study found that a correction of misinformation is less effective if that misinformation was attributed to a credible source and was "repeated multiple times prior to correction."[11] As it becomes easier to generate fake visual evidence, parties will be inclined to attempt to offer it as evidence. A research study conducted in 2009 concluded that because jurors may get confused and frustrated when attorneys or witnesses explain technical or complex material, visual aids help them retain information much better.[12] Their study showed that jurors retained up to "85% of what they learn[ed] visually" as opposed to only about 10% of what they heard.[13]

This research illustrates why judges will need to exercise more control over whether alleged AIM goes to a jury. But do the Federal Rules of Evidence provide sufficient flexibility to handle AIM?

We posit that judges—if adequately educated about the unique challenges such deepfake evidence presents—can proactively manage evidentiary challenges related to alleged AIM under the existing Federal Rules of Evidence. Ordinarily, to introduce evidence, a party merely needs to show that it is relevant and authentic as set forth in Federal Rules of Evidence 401 and 901, respectively. This presents a low bar. If the alleged AIM is central to a matter, it will easily satisfy the relevance requirement, and satisfying the authenticity standard at this stage merely requires a showing that it is more likely than not that the evidence "is what the proponent claims it is."[14]

We propose that judges proactively address potential problems in this process by requiring that the parties raise potential AIM issues in pretrial conferences under Federal Rules of Civil Procedure 16 and

---

[9] *Id.* at 16.

[10] *Id.* at 16; *see also* Nathan Walter & Sheila T. Murphy, *How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation*, 85 COMM. MONOGRAPHS 423–441 (2018) for a study showing that correction of constructed misinformation, such as a fictional accident, is easier than correction of real-world misinformation, such as "denial of climate change," because of previous exposure to the real-world misinformation, and perhaps high involvement, that triggers a defensive processing.

[11] Nathan Walter & Riva Tukachinsky, *A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?*, 47 COMM. RSCH. 4, 155–177 (2020).

[12] Zachariah B. Parry, Note, *Digital Manipulation and Photographic Evidence: Defrauding the Courts One Thousand Words at a Time*, 2009 U. ILL. J.L. TECH. & POL'Y 175, 185 (2009).

[13] *Id.* at 184–85.

[14] FED. R. EVID. 901.

26(f). This will allow the parties to obtain discovery of evidence that corroborates or rebuts allegations that certain evidence is AIM and hire expert witnesses to address AIM. By being proactive, judges can also ensure that there is sufficient time to hold a hearing focused on the AIM, rather than having to handle the issues on the eve of or during trial without the parties having fully developed the factual and legal record.

Federal Rule of Evidence 403 provides another tool to manage AIM. It allows judges to "exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence."[15] Research suggests that when contested audiovisual deepfakes go to the jury, even if the jury understands that they may be or are likely fake, the deepfake can nonetheless dramatically alter jurors' perceptions.[16] This could lead to unfair prejudice, misleading the jury, or another Rule 403 problem that could substantially outweigh the probative value of the evidence, providing a basis for excluding contested AIM.

In this Article, we present a hypothetical election interference case to show how judges and lawyers can proactively manage AIM issues under the existing Federal Rules of Evidence. There is a long history of foreign nation-states interfering in the elections of other nation-states.[17] Recent examples include the alleged Russian interference in the 2016 and 2020 U.S. presidential elections[18] as well as in the 2017 French presidential election.[19] Since then, deepfakes have been used in the 2023 Turkish[20] and Slovak[21] elections. In the 2023 Chicago mayoral election, a deepfake portrayed mayoral candidate Paul Vallas making statements that he did not make.[22] There is therefore strong reason to

---

[15] FED. R. EVID. 403.

[16] *See supra* note 7.

[17] Vasu Mohan & Alan Wall, *Foreign Electoral Interference: Past, Present, and Future*, 20 GEO. J. INT'L AFFS. 110–119 (Fall 2019).

[18] *See, e.g.*, Pippa Norris, *Electoral Integrity in the 2020 U.S. Elections*, ELECTORAL INTEGRITY PROJ. 17 (Dec. 2020).

[19] *See, e.g.*, Emilio Ferrara, *Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election*, 22 FIRST MONDAY (2017).

[20] Pelin Ünker & Thomas Sparrow, *Fact Check: Turkey's Erdogan Shows False Kilicdaroglu Video*, DW (May 24, 2023), https://www.dw.com/en/fact-check-turkeys-erdogan-shows-false-kilicdaroglu-video/a-65554034 [https://perma.cc/2F3B-QGM5].

[21] Daniel Zuidijk, *Deepfakes in Slovakia Preview How AI Will Change the Face of Elections*, BLOOMBERG (Oct. 4, 2023, 11:00 AM), https://www.bloomberg.com/news/newsletters/2023-10-04/deepfakes-in-slovakia-preview-how-ai-will-change-the-face-of-elections [https://perma.cc/KF8C-KUB5].

[22] Donie O'Sullivan, *This Deepfake Surfaced in a Tight Mayoral Race. It's Just the Beginning*, CNN (Feb. 7, 2024), https://www.cnn.com/videos/politics/2024/02/07/deepfake-artificial-intelligence-elections-chicago-paul-vallas-orig.cnn [https://perma.cc/4AVX-95K6].

believe that deepfakes will be used in future U.S. elections and that they will be the subject of allegations and counter-allegations that, at least in some cases, will end up being contested in court.

II. CREATING AND DETECTING AI-GENERATED MATERIAL

A. Creating AIM

There are several tools available today for creating fake media. For instance, fake images can be generated in response to a textual prompt by systems such as Microsoft's Bing Image Creator and OpenAI's DALL-E. Synthetic audio in the voice of a specific person can be generated using online services such as Speechify, with less than a minute of training audio of the target's voice. Synthetic video can be generated using online services such as Synthesia. A more recent product in this space is OpenAI's Sora, which can generate video from a text prompt. These are just a few of the well-known systems that can generate synthetic media.

Reputable services for creating synthetic media typically impose prohibitions on users creating malicious deepfakes, such as by requiring users to certify that they have permission to use the audio and video that they have uploaded.[23] But users can misrepresent their rights to use media and circumvent guardrails on such platforms. There are many examples of users generating prompts that create violent or sexual content by using prompts that the AIM generation platform did not expect.[24] To enhance traceability, some AIM-generation platforms embed watermarks or digital signatures within any AIM that they create.[25] The idea is that third parties can check for the presence of such watermarks. But these methods are far from foolproof and there is evidence that such watermarks can be removed, at least in some cases, without much difficulty.[26] Even if all online services could prevent malicious uses and added watermarks to outputs, people with moderate

---

[23] *See, e.g.*, "Terms & Conditions," *Speechify*, § 7.6(a), (May 25, 2023), https://speechify.com/terms [https://perma.cc/T55Q-VD62] ("[Y]ou represent and warrant . . . [y]ou own your User Material or have the right to submit it, and in submitting it you will not be infringing any rights of any third party, including intellectual property rights (such as copyright or trade mark), privacy or publicity rights, rights of confidentiality or rights under contract.").

[24] Katyanna Quach, *Attempts to Demolish Guardrails in AI Image Generators Blamed for Lewd Taylor Swift Deepfakes*, REGISTER (Feb. 5, 2024), https://www.theregister.com/2024/02/05/deepfakes_taylor_swift_4chan_competition/ [https://perma.cc/AC2W-M644].

[25] Beatrice Nolan, *OpenAI is Adding Digital Watermarks to its AI-Generated Images—But It's Not a Perfect Solution*, BUS. INSIDER (Feb. 7, 2024, 7:07 AM), https://www.businessinsider.com/openai-adding-digital-watermarks-ai-images-deepfakes-2024-2 [https://perma.cc/3LVL-3PYS].

[26] Barry Collins, *The Ridiculously Easy Way to Remove ChatGPT's Image Watermarks*, FORBES (Feb. 7, 2024, 5:34 AM), https://www.forbes.com/sites/barrycollins/2024/02/07/the-ridiculously-easy-way-to-remove-chatgpts-image-watermarks/ [https://perma.cc/YWT8-S3VT].

technical skills can access software that would allow them to create deepfakes without watermarks. Today, publicly accessible code-repositories such as GitHub include large amounts of software source code that can be used to create fake audio clips, images, and videos. Such code repositories rarely embed watermarks. Even in the rare cases when they do, the watermarks can be easily removed by programmers.

We will briefly describe a widely used technique and tool to create AIMs today: Generative Adversarial Networks (GANs)[27] and Stable Diffusion (SD), respectively.[28] Both GANs and SD can be applied to generate fake audio and video, and even fake multimodal content.

A GAN consists of two algorithms working together: a generator and a discriminator.[29] Suppose we want to generate a synthetic (i.e., "fake") image. In the first iteration, the generator creates an image by randomly selecting pixel values from some probability distribution. The result will be akin to the result of a monkey using a set of paintbrushes on a canvas. A batch of such images will be created and sent to the discriminator (a deep-learning classifier), which will likely discover that most, if not all the generator's images are fake. The prediction made by the discriminator is provided as feedback to the generator, which now knows that the images it had previously generated were detected as fakes. A second iteration repeats the process, but this time, the generator uses the feedback from the previous iteration to avoid past mistakes. This new batch of fake images is fed back to the discriminator, which again makes its predictions and provides feedback to the generator. After thousands or even millions of iterations, an equilibrium is reached: the generator creates sufficiently realistic fake images so that over several consecutive iterations, the discriminator is unable to improve its ability to detect the images as fake. At this point, the images generated by the generator are the best possible versions.

Stable Diffusion[30] starts with an image, *I* (e.g., a 512 x 512 x 3 image, i.e., a 512 x 512 pixel image with three channels: red, green, and blue), and converts it into a latent representation, *L1,* which is much smaller in size (e.g., 64 x 64 x 4 image). An example of this is provided in Figure 1 below. The first two dimensions of the original image (i.e.,

---

[27] *See* Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio, *Generative Adversarial Networks*, 63 *COMMUNICATIONS OF THE ACM* 139–44 (Nov. 2020).

[28] *See* Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, & Ming-Hsuan Yang, *Diffusion Models: A Comprehensive Survey of Methods and Applications*, 56 ACM COMPUTING SURVS. 38 (Nov. 2023).

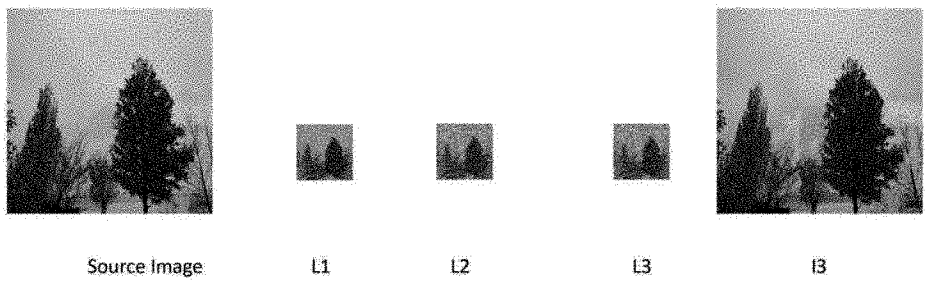
[29] *See* Goodfellow et al., *supra* note 27.

[30] *See* Yang et al., *supra* note 28; Andreas Blattman, et al., *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets* (Nov. 2023) (unpublished paper, arXiv:2311.15127).

the 512 x 512 part) represent an image as a two-dimensional matrix of pixels. These two dimensions represent the length and width of the image. The three channels represent the intensity of red, green, and blue colors in each pixel. Thus, when we see a standard 512 x 512 pixel image, we can think of this as three such images combined together—one corresponding to the red channel, one corresponding to the green channel, and one corresponding to the blue channel. The latent representation (e.g., a 64 x 64 x 3 image) is a technical representation that contains the "essence" of the original image but is much smaller. It is important to note that the latent representation does not have to be a 64 x 64 x 3 sized image. It could just as well be a (64 x 64 x 4) image or some other size. The latent representation does need to be much smaller than the original image (e.g., 64 x 64 x 4 = 16,384, which is much smaller than 512 x 512 x 3 = 786,432) to improve computational efficiency, such as runtime and computational resources used. The smaller the size of the latent image, the less representative it will be of the original image. The larger the size, the more representative it is. However, a smaller-sized image can be more efficiently processed, while a larger-sized image requires more runtime and computational resources (e.g., GPU computing resources). Thus, there is a tradeoff between the size of the latent representation and the runtime and computational resources required. Next, "noise" is iteratively added to the latent representation, yielding a new latent representation, *L2*. One can think of "noise" as modifications to the red, green, and blue values for some of the pixels of the latent representation. The latent representation *L2* should still contain the "essence" of the original image *I* but will look different from *L1* because of the added noise. A denoising process[31] is now used to remove the noise from *L2*, but this is not done perfectly, leading to a new latent representation, *L3*. *L3* will look different from *L1* because the denoising process is not perfect. At this point, the process that converted *I* into *L1* is reversed, but this reversal is applied to *L3* to get a new image, *I3*, that has the same size as *I*. In our example, *I3* will still bear a resemblance to the source image *I* but will look different. A rendering of this process is provided in Figure 1 below. The generated image, *I3*, has some trees with more snow on them than the original image.

[31] Denoising is a process that attempts to clean up imperfections in an image. For instance, an image taken with a traditional camera might have spots or a bright ray of sunshine that over-illuminates a portion of the image. A similar phenomenon can occur in audio (e.g. when there is a crackle in a recorded phone call). Denoising methods attempt to correct such imperfections in captured media. *See generally* Chunwei Tian, et al., *Deep Learning on Image Denoising: An Overview*, 131 NEURAL NETWORKS 251 (2020).

Figure 1: Stable Diffusion Image Generation Process Applied to an Image Taken by One of the Authors



B. Detecting AIM

A number of methods have been developed to detect deepfake media. In July 2023, several AI companies reached an agreement with the Biden administration[32] to place an embedded code (a "watermark") within any AIM. To ascertain whether a digital artifact is real or fake, all one would only need to look for the embedded code. Camera manufacturers are also trying to embed cryptographic signatures into images that are taken using that camera.[33] But, as explained in the last section, for a variety of reasons most experts doubt that watermarks will solve the deepfake problem.[34]

Early deepfake images were easily detected by humans because of "dumb" mistakes: an image of a person showing them having six fingers or a misshapen ear. In other cases, perfectly intelligible words (e.g., on a street sign) might have been mangled, such as a "STOP" sign reading "SWOT." Today's deepfakes are much more sophisticated than those of the past and such mistakes are less common. Instead, deepfake detectors (DDs) look for visual discontinuities in images. For instance, is the transition between a person's blue shirt sleeve and their dark skin a clear separation (as would be the case in a real image) or is there a region where there is a transition (with some portion near the border of

---

[32] Matt O'Brien & Zeke Miller, *White House Sets AI Safeguard Agreements with Amazon, Google, Meta, Microsoft and Other Tech Firms*, PBS (Jul. 21, 2023, 12:09 PM), https://www.pbs.org/newshour/nation/white-house-sets-ai-safeguard-agreements-with-amazon-google-meta-microsoft-and-other-tech-firms [https://perma.cc/4Y2Q-WVXP]; Exec. Order No. 14,110, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Fed. Reg. 24,283 (Oct. 30, 2023).

[33] *See* Matthew S. Smith, *This Leica Camera Stops Deepfakes at the Shutter*, IEEE SPECTRUM (Nov. 17, 2023), https://spectrum.ieee.org/leica-camera-content-credentials [https://perma.cc/Q9F4-P72H].

[34] *See* Vittoria Elliott, *Big AI Won't Stop Election Deepfakes With Watermarks*, WIRED (Jul. 27, 2023, 7:00 AM), https://www.wired.com/story/ai-watermarking-misinformation/ [https://perma.cc/99GE-9KU4].

the shirt sleeve and the skin looking different)? Similarly, DDs can look for improperly formed shadows (e.g., are the shadows consistent with the expected number of light sources?). In the case of videos, are there inconsistencies in the movement of the facial muscles and lips, and the rendered speech? In the case of audio, does the audio sound monotonous? Or does it have the usual ups and downs of ordinary human speech? These important questions underlie some of the DDs available today.

In addition to DD methods that seek to detect genuinely new deepfakes, there are also specialized systems that are capable of finding copies or variations of images already known to be deepfakes. Systems such as PhotoDNA,[35] from places such as Dartmouth College and Microsoft, have been used for well over a decade to find near-copies of images known to depict illegal content, such as terrorist imagery and child sexual abuse material (CSAM). Such systems can be used to find copies of deepfake images after someone has already found an initial version that was separately found to be a deepfake.

Unfortunately, DD algorithms are far from perfect. Several authors of this Article (Gao, Pulice, Subrahmanian) have conducted tests on a small suite of videos, both real and deepfake. Table 1 shows their findings. One hundred real videos were collected from the Internet, as well as 100 well-known deepfakes. The authors also generated 100 deepfakes in the Northwestern Security & AI Lab (NSAIL). They tested four well-known deepfake detectors (DD1 through DD4), which included the winner of the Meta Deepfake Detection Challenge.[36] DD1 labeled every real video as real, but also labeled almost every fake video as real. Simply put, DD1 labeled almost everything as real and found almost no fake videos, showing a high false-positive rate. DD2 arguably did the best, getting an error rate of only 3% on the real videos, but still huge error rates (76% and 89%) on the two fake datasets. DD3 was slightly better at detecting fakes (error rates of 87% and 71%) but made more errors on real videos (15%). DD4's performance was close to that of DD3, with a 10% error rate on real videos and 85% and 87% error rates on fake ones. While many of the deepfakes would have been easily detected by a human, these detectors were biased toward labeling videos as real, thereby making few errors on real videos and many on fake videos.

---

[35] Hany Farid, *An Overview of Perceptual Hashing*, 1 J. ONLINE TR. & SAFETY, Oct. 2021.

[36] *See* Cristian Canton Ferrer et al., *Deepfake Detection Challenge Results: An Open Initiative to Advance AI*, META (Jun. 12, 2020), https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ [https://perma.cc/UEC9-YR3B].

Table 1: Deepfake Detector Error Rates

SOURCE	METRIC	DD1	DD2	DD3	DD4
Real Videos	False Positive Rate	0	0.03	0.15	0.10
NSAIL Deepfakes	False Negative Rate	0.92	0.76	0.87	0.87
Fake Internet Videos	False Negative Rate	0.99	0.89	0.71	0.85

These results do not provide confidence that today's DDs can reliably distinguish between real and fake videos. Given concerns about the validity of DDs (i.e., the accuracy of a DD's predictions of whether media is real or fake) as well as their reliability (i.e., the consistency of making accurate predictions about whether media is real or fake),[37] the introduction of DDs in a legal matter will likely face hurdles under the Federal Rules of Evidence.[38]

III. ELECTION-INTERFERENCE HYPOTHETICAL

The possibility of foreign interference in elections by using deepfakes presents a serious challenge to national security. When alleged deepfakes are deployed, there will be concerns about possible foreign interference. No matter the specific facts, the related uncertainty regarding election integrity is a threat to national security in various ways.

First, when there are allegations of deepfake use in an election, there is the possibility that one or more candidates will challenge the legitimacy of the election. As an example, just two days before the 2023 Slovak election, an audio deepfake depicted anti-Russian candidate Michal Šimečka discussing how he rigged the election.[39] Slovakia's election rules forbid candidate statements within 48 hours (two days) of the poll—making it near-impossible for the candidate to question the

---

[37] Paul W. Grimm, et al., *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. 9, 89 (2021).

[38] *See id.*

[39] *See* Curt Devine, et al., *A Fake Recording of a Candidate Saying He'd Rigged the Election Went Viral. Experts Say It's Only the Beginning*, CNN (Feb. 1, 2024), https://www.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html [https://perma.cc/ZW6L-P9P2].

authenticity of the deepfake. The outcome of the election may have been a casualty of the deepfake, leading to a more pro-Russian government.[40]

Second, allegations of deepfake use may sow distrust amongst the population in the elected government, even if the allegations of deepfake use were false.[41]

Third, deepfakes might deter certain voters from going to the polls. As an example, February 2024 witnessed the release of an audio deepfake which falsely impersonated President Biden telling voters not to go to the polls.[42] Should such deepfakes not be quickly debunked in the future, the outcome of an election could be compromised. These are just three possible ways in which deepfakes could be used to compromise the security of an election and help impose an improperly elected government on a democracy.

A. Hypothetical Scenario

Our scenario involves a hypothetical election-interference case brought in federal court by a presidential candidate, Connie, against PoliSocial, a political social-media strategy company, and Eric, Connie's opponent.

Assume that it is August 4, 2028, three months before the 2028 U.S. presidential election. Connie alleges that Eric and entities affiliated with Eric, including PoliSocial, published fake AIM that defamed Connie. Since the party conventions, which took place in early July 2028, a massive social-media campaign involving a network of 6,000 bot accounts has unleashed a wave of disinformation targeting Connie. Connie's lawyers have obtained evidence that these bot accounts all posted content from the same IP address, a third-party data operations center in New Hampshire. A news investigation subsequently showed close coordination between Eric's campaign and the third party, a political action committee (PAC) that supports Eric.

More important to national security, Connie alleges that the fake AIM is designed to interfere with the upcoming election by attempting to intimidate Connie's supporters and prevent them from voting for her because of her alleged connections to Chinese officials. Connie also alleges that the AIM is designed to mobilize Eric's supporters to further

---

[40] *See* Morgan Meaker, *Slovakia's Election Deepfakes Show AI Is a Danger to Democracy*, WIRED (Oct. 3, 2023), https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/ [https://perma.cc/2VZ8-R73G].

[41] Daniel Byman, Daniel W. Linna Jr. & V.S. Subrahmanian, *Governments Use of Deepfakes*, CTR. FOR STRATEGIC & INT'L STUD., Mar. 2024 at 3.

[42] *See* Shannon Bond, *AI Fakes Raise Election Risks as Lawmakers and Tech Companies Scramble to Catch Up*, NPR (Feb. 8, 2024), https://www.npr.org/2024/02/08/1229641751/ai-deepfakes-election-risks-lawmakers-tech-companies-artificial-intelligence [https://perma.cc/69FE-KJHY].

intimidate and threaten Connie's supporters so that they do not exercise their constitutional right to vote for Connie. Connie seeks an injunction requiring Eric to retract the defamatory statements, admit that the video and audio recordings released were fake AIM, and cease engaging in defaming Connie and taking actions that seek to intimidate and threaten voters and interfere with the exercise of their right to vote.

Allegations posted by these bot accounts included videos of Connie in a variety of compromising poses with a man later identified as a Chinese embassy official. These videos, from security cameras and private cell phones, were geolocated to expensive resorts, beaches, and spas. In addition, many audio clips appeared to show Connie soliciting money from an unnamed man whose phone was later geolocated in Beijing at the time of the call.

The messages and allegations were amplified by frequent social media messages posted by the 6,000 bot accounts using a variety of hashtags such as #CheatingConnie, #CrookedConnie, and #VoteConnieOut. Despite desperate denials from Connie's campaign, the stories spread like wildfire, first on social media, and then on mainstream news and broadcast media. The posts were also sent to social media groups frequented by Connie's supporters.

Polls suggest a close race between the two candidates. Both campaigns expect a razor-thin margin, setting up the potential for several rounds of recounts in key battleground states; thus, every vote will matter.

Several of the circulating videos show ballot boxes being stuffed in pro-Connie districts during the primary elections in New Hampshire and South Carolina. These videos also went viral, first on social media, and later in mainstream news and broadcast TV channels. Subsequent videos show voters attending Connie's rallies being confronted by "election integrity" groups threatening her supporters with violence at the polls. Members of these groups have been linked to Eric's campaign.

Meanwhile, Connie's campaign received a tip in the form of a recording of PoliSocial CEO John Doe apparently speaking with Eric just after a campaign event in late July 2028. In this clip, John Doe is caught saying "We buried her, Eric. The videos are doing the trick. This country is not ready for a woman in the Oval Office. Congrats." Geolocation data showed that the tipster's phone was near the phones of both John Doe and Eric at the time the alleged audio recording was made.

At the same time, the campaign received another audio recording of Eric during a prior Senate election three years ago. The recording allegedly captures Eric and a campaign staffer discussing the use of deepfake videos to implicate their opponent in a similar corruption scandal. Eric is caught saying "Wow! This technology is so good now it would be impossible for anyone to spot it as a fake." When confronted

with this evidence, John Doe and Eric both insisted that the audios are deepfakes. Connie's campaign claims that the audio recordings are smoking guns.

IV. CLAIMS UNDER THE VOTING RIGHTS ACT AND FOR DEFAMATION

This hypothetical presents several critical pieces of audio and audiovisual evidence that both parties will seek to introduce to buttress their claims and defenses. These include the alleged videos showing Connie in compromising poses with a Chinese embassy official and showing ballot boxes being stuffed by her constituents in the primaries; the alleged audio recordings of conversations between Connie and an individual geolocated in Beijing; the conversation between Eric and PoliSocial's CEO, John Doe; and the recording of Eric and his campaign staffer from three years ago.

Given the highly public nature of the presidential election campaign, there are several considerations that would frame a trial. For one, there is the question of irreparable harm to both campaigns from the widespread dissemination of the audiovisual evidence in the public domain through both conventional and social media platforms. Such dissemination would likely have a lasting impression on public perception of these candidates before the trial begins. Relatedly, given the high-profile nature of presidential campaigns, it is also likely that potential jurors will have been exposed to the same evidence. Moreover, both candidates have an incentive to launch an inauthentic "deepfake defense" in which they challenge genuine audiovisual evidence as being AIM to support their respective claims and defenses, benefitting from the "Liar's Dividend."[43] The Liar's Dividend describes the phenomenon where some actors will seek to "escape accountability for their actions by denouncing authentic video and audio as deep fakes."[44] They would attempt to invoke the public's growing skepticism of audio and video evidence as it learns more about the power of AIM.[45]

A. Voting Rights Act Claim

To bring a claim under the Voting Rights Act, Connie must show that Eric and his affiliates intimidated, threatened, or coerced her supporters or attempted to do so "for voting or attempting to vote."[46] Connie will seek to introduce the videos showing her supporters being

---

[43] *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1785 (2019).

[44] *Id.*

[45] *Id.*

[46] 52 U.S.C. § 10307(b).

confronted and harassed at rallies and warned against voting for her, and will argue that the videos are fakes created by Eric's campaign designed to intimidate people so that they do not go to the polls and vote for her. She will also seek to introduce the posts targeting her supporters on social media and allege that Eric and his supporters are using such posts to spread disinformation and discourage her supporters from exercising their right to vote. Eric will deny responsibility for the videos and contend that he has no reason to believe that the videos have been faked, but he is still not responsible for the conduct of any individuals depicted in the videos. Eric will further argue that citizens have constitutional rights, including under the First Amendment, to protest undemocratic activities designed to undermine fair elections.

B. Defamation Claim

In addition to the Voting Rights Act claim, Connie will bring a claim for defamation to buttress her request for relief that Eric retract his claims, admit that the content of the audio and videos showing her in potentially illegal or compromising positions are fake, and be enjoined from publishing additional false claims. To prove defamation, Connie must show that (i) a false statement was made by Eric's campaign, (ii) the statement was communicated to a third party, (iii) Eric's campaign acted with actual knowledge that the statement was false, and (iv) Connie suffered harm.[47] Since Connie is a public figure, Connie will also have to show that with regard to the allegedly false material, Eric acted with "actual malice—that is, with knowledge that it was false or with reckless disregard of whether it was false or not."[48]

Eric will argue that the compromising audio and video recordings of Connie are real and attempt to use them as evidence to show that no false statements were made. If the alleged statements are true, Connie cannot state a defamation claim. Connie, in turn, will argue that they are fake AIM. Connie will seek to introduce the audio recording appearing to capture a conversation between Eric and PoliSocial's CEO to show that Eric's campaign created and spread false information about her campaign. Eric will dispute the authenticity of the audio recording, arguing that it is AIM. Connie will also try to introduce the audio recording from three years ago of Eric talking to a staffer about using deepfakes against his opponent in a previous senate election. As such, both Eric and Connie are likely to challenge the admissibility of

[47] *See Defamation*, LEGAL INFO. INST. (June 2023), https://www.law.cornell.edu/wex/defamation [https://perma.cc/LK4Q-GVVD].

[48] New York Times Co. v. Sullivan, 376 U.S. 254, 280 (1964).

audiovisual evidence on the basis that it is AIM and therefore does not meet the authentication requirements under the Federal Rules of Evidence.

In this hypothetical, the admissibility of the audiovisual evidence is central to the disposition of the case—even more so if the candidates are unable to provide other forms of corroborating evidence to support their claims or defenses. In determining the admissibility of the evidence, the court will also have to consider the risk of unfair prejudice to either party if the disputed evidence is admitted and turns out to be fake, but sways a jury nonetheless.

V. FEDERAL RULES OF EVIDENCE FRAMEWORK

When a party introduces non-testimonial evidence, they must meet the admissibility requirements of relevance and authenticity under Federal Rules of Evidence 401 and 901, respectively. "Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence, and (b) the fact is of consequence in determining the action."[49] "[E]ven evidence that has a slight tendency to . . . resolve a civil or criminal case meets the standard."[50] In addition, Rule 402 states that "[r]elevant evidence is admissible unless any of the following provides otherwise: the United States Constitution; a federal statute; these rules [of evidence]; or other rules prescribed by the Supreme Court. Irrelevant evidence is not admissible."[51] "In essence, Rule 402 creates a presumption that relevant evidence is admissible, even if it is only minimally probative, unless other rules of evidence or sources of law require its exclusion."[52]

A. Rule 403 Allows the Exclusion of Relevant Evidence When Probative Value is Substantially Outweighed by Unfair Prejudice

Federal Rule of Evidence 403, however, states that "[t]he court may exclude relevant evidence if its probative value is substantially outweighed by one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence."[53] The Advisory Committee Notes accompanying Rule 403 state that:

---

[49] FED. R. EVID. 401.

[50] Grimm et al., *supra* note 37, at 87.

[51] FED. R. EVID. 402.

[52] Grimm et al., *supra* note 37, at 87.

[53] FED. R. EVID. 403.

> The case law recognizes that certain circumstances call for the exclusion of evidence which is of unquestioned relevance. These circumstances entail risks which range all the way from inducing decision on a purely emotional basis, at one extreme, to nothing more harmful than merely wasting time, at the other extreme. Situations in this area call for balancing the probative value of and need for the evidence against the harm likely to result from its admission.[54]

Rule 403 therefore establishes a balancing test which tilts in favor of admissibility and permits the exclusion of relevant evidence upon a sufficient showing of unfair prejudice to the party against whom the evidence is introduced, or some other specific problematic outcome.[55]

Mere prejudice alone is insufficient to permit the exclusion of relevant evidence; the prejudice must be sufficiently unfair to warrant exclusion under Rule 403.[56] Specifically, Rule 403 provides, "Unfairness may be found in any form of evidence that may cause a jury to base its decision on something other than the established propositions in the case."[57] It further provides, "Prejudice is also unfair if the evidence was designed to elicit a response from the jurors that is not justified by the evidence."[58]

Relevant evidence may also be excluded under Rule 403 when it might confuse the issues or mislead the jury. Just as with unfair prejudice, this analysis is highly fact-dependent.[59] One recurring basis for excluding evidence as confusing the issues or misleading a jury is when plausible evidence would be very difficult to rebut.[60] "Courts are reluctant to admit evidence that appears at first to be plausible, persuasive, conclusive, or significant if detailed rebuttal evidence or complicated judicial instructions would be required to demonstrate that the evidence actually has little probative value."[61]

Courts have excluded some types of scientific and statistical evidence under Rule 403, particularly "if the jury may use the evidence for purposes other than that for which it was introduced."[62] Courts have also excluded legal materials such as statutes, cases, and constitutional

---

[54] FED. R. EVID. 403 Advisory Committee Notes (emphasis added).

[55] *See* Grimm et al., *supra* note 37, at 88.

[56] *See* JACK B. WEINSTEIN & MARGARET A. BERGER, WEINSTEIN'S FEDERAL EVIDENCE § 403.04 (2d ed. 2024).

[57] *See id.*

[58] *See id.*

[59] *See id.* at § 403.05(2).

[60] *See id.* at § 403.05(3)(b).

[61] *See id.*

[62] *See id.* at § 403.05(3)(c).

provisions.[63] Finally, courts may exclude relevant evidence "if its probative value is substantially outweighed by danger of undue delay, wasting time, or needlessly presenting cumulative evidence."[64]

Rule 403 does not confer judges with the power to determine witness credibility, which remains the domain of the jury.[65] Judges may "exclude testimony that no reasonable person could believe, where it flies in the face of the laws of nature or requires inferential leaps of faith rather than reason."[66] Apart from such outliers, Rule 403 leaves the power to determine credibility with the jury and does not let the judge supplant the jury's views on credibility.

Appellate courts afford trial courts wide discretion in exercising their Rule 403 powers, and trial-court decisions are only reversed where there has been an abuse of discretion.[67] Nevertheless, appellate courts have recognized that this power should be exercised sparingly given that the balancing test weighs in favor of admissibility.[68] Furthermore, appellate courts have indicated a preference for trial judges to state their findings on the record rather than simply presenting their conclusions.[69] "The greater the risks, the more vital the evidence, the more thorough should be the consideration given to objections under Rule 403, and the more need there is for trial judges to give some indication of the bases for their decisions."[70] Furthermore, "sidebar conferences in which the matter is raised and discussed should be on the record."[71]

B. Authenticity Under Rule 901 is a Low Bar

The second requirement for admissibility of non-testimonial evidence is that it must meet the authenticity requirement under Federal Rule of Evidence 901. Rule 901 states that "the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is."[72] "This low threshold allows a party to fulfill its obligation to authenticate non-testimonial evidence by a mere preponderance, or slightly better than a coin toss."[73] Rule 901(b) lists ten non-

---

[63] *See id.* at § 403.05(3)(a).

[64] *See id.* at § 403.06 (citations omitted).

[65] *See* CHRISTOPHER B. MUELLER & LAIRD C. KIRKPATRICK, FEDERAL EVIDENCE § 4.12 (4th ed. 2023).

[66] *See id.*

[67] *See id.*

[68] *See id.*

[69] *See id.*; *see also* FED. R. EVID. 103(c) ("The court may make any statement about the character or form of the evidence, the objection made, and the ruling.").

[70] *See* MUELLER & KIRKPATRICK, *supra* note 65.

[71] *See id.*

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

[73] Taurus Myhand, *Once the Jury Sees It, the Jury Can't Unsee It: The Challenge Trial Judges*

exclusive ways in which a proponent can demonstrate authenticity.[74] In the context of audio evidence, authentication can be satisfied using an "opinion identifying a person's voice—whether heard firsthand or through mechanical or electronic transmission or recording—based on hearing the voice at any time under circumstances that connect it with the alleged speaker."[75]

There are two theories under which video evidence can be admitted: "either as illustrative evidence of a witness's testimony (the "pictorial-evidence theory") or as independent substantive evidence to prove the existence of what is depicted (the "silent-witness theory")."[76] Under the pictorial-evidence theory, video evidence can be authenticated by any witness present when it was made who perceived the events depicted.[77] Videos admitted under the silent-witness theory are subject to additional scrutiny since there are no independent witnesses to corroborate their accuracy.[78] In instances where videos are the products of security surveillance cameras, they could be authenticated as the accurate product of an automated process.[79] Ultimately, the threshold for admissibility remains low, and once the proponent of the evidence shows that a reasonable jury could find the video authentic, the burden shifts to the opponent to demonstrate that it is clearly inauthentic.[80]

C. The Federal Rules of Evidence Allocate Greater Factfinding Power to Juries

Finally, it is also worth examining how the Federal Rules of Evidence delegate adjudication responsibilities on evidentiary admissibility between the judge and the jury. Ultimately, the Federal Rules of Evidence were designed to allocate greater preliminary factfinding power to juries and reflected a turn away from the traditional English common law approach that gave judges unfettered power in determining the admissibility of evidence.[81] This can be seen in the interplay

---

Face When Authenticating Video Evidence in the Age of Deepfakes, 29 WIDENER L. REV. 171, 177 (2022).

[74] FED R. EVID. 901(b).

[75] FED R. EVID. 901(b)(5).

[76] *See* Myhand, *supra* note 73, at 177.

[77] *See id.*

[78] *See id.*

[79] *See* FED. R. EVID. 901(b)(9).

[80] Myhand, *supra* note 73, at 179.

[81] *See* Edward J. Imwinkelried, *Trial Judges: Gatekeepers or Usurpers? Can the Trial Judge Critically Assess the Admissibility of Expert Testimony Without Invading the Jury's Province To Evaluate the Credibility and Weight of the Testimony?*, 84 MARQ. L. REV. 1, 8 (2000); *see also* EDWARD J. IMWINKELRIED, 45 AM. JUR. TRIALS 1 §§ 6–9 (2024) (tracing the history of Federal Rule of Evidence 104).

between Rule 104(a), which defines the role of the trial judge in making preliminary determinations regarding the admissibility of evidence, the qualification of witnesses, and the existence of an evidentiary privilege,[82] and Rule 104(b), the so-called "conditional-relevance rule," which provides that when the relevance of evidence depends on the existence of a fact, "proof must be introduced sufficient to support a finding that the fact does exist."[83] While it may not be apparent from the text of Rule 104(b) itself, what this means in essence is that when one party claims that evidence is authentic, and therefore relevant and admissible, but the opposing party claims that it is fake, and therefore not relevant to prove any disputed fact, it is the *jury*, not the judge, that must resolve the fact dispute and decide which version of the facts it accepts, so long as the judge finds that sufficient proof has been introduced for the jury to be able to reasonably find that the evidence is authentic.[84]

Proponents of evidence that is challenged as AIM could argue that the incriminating evidence's ultimate authenticity should be determined by the jury as a part of its role as a decider of contested facts under Rule 104(b). They could further argue that letting the jury make such factual determinations would not undermine the jury deliberation process.[85] For instance, Rule 104(b) allocates to the jury the responsibility of determining the authenticity of an exhibit, such as a letter.[86] The jury could simply "disregard the letter's contents during their deliberations" if they determined it was a forgery.[87] The rules show a particular concern with letting the judge entirely exclude evidence because they fear doing so would greatly restrict the jury's function as a trier of fact, and in some cases, virtually eliminate it.[88]

---

[82] FED. R. EVID. 104(a) ("The court must decide any preliminary question about whether a witness is qualified, a privilege exists, or evidence is admissible. In so deciding, the court is not bound by evidence rules, except those on privilege.").

[83] FED. R. EVID. 104(b) ("When the relevance of evidence depends on whether a fact exists, proof must be introduced sufficient to support a finding that the fact does exist. The court may admit the proposed evidence on the condition that the proof be introduced later.").

[84] Fed. R. Evid. 104(b) advisory committee notes ("If preliminary questions of conditional relevancy were determined solely by the judge, as provided in subdivision (a) [of Rule 104], the functioning of the jury as a trier of fact would be greatly restricted and in some cases virtually destroyed. These are appropriate questions for juries. Accepted treatment, as provided in the rule, is consistent with that given fact questions generally. The judge makes a preliminary determination whether the foundation evidence is sufficient to support a finding of fulfillment of the condition. If so, the item is admitted. If after all the evidence on the issue is in, pro and con, the jury could reasonably conclude the fulfillment of the condition is not established, the issue is for them. If the evidence is not such as to allow a finding, the judge withdraws the matter from their consideration.").

[85] *See* Delfino, *supra* note 7, at 324.

[86] *See* FED. R. EVID. 104(b) advisory committee notes.

[87] *See* Imwinkelried, *Trial Judges*, *supra* note 81, at 11.

[88] *See* FED. R. EVID. 104(b) advisory committee notes.

D. Potential Bases for Excluding Possible Deepfakes Under Rule 403

The critical evidentiary issue that the judge must decide is whether Federal Rule of Evidence 104(b) requires the judge to admit the contested audiovisual evidence and let the jury determine the disputed fact of its authenticity, or whether the judge, as gatekeeper under Rule 104(a), may (or must) exclude the audiovisual evidence under Rule 403 if the judge finds that the unfair prejudice to the opponent of admitting the evidence substantially outweighs its probative value.

When considering audiovisual evidence, studies have shown that once the jury has seen disputed videos, they are unlikely to be able to put them out of their minds whether or not they are told they are fake.[89] Our research has not disclosed any caselaw addressing this dilemma in the context of AIM. However, in other contexts, there is ample precedent supporting the authority of the trial judge to exclude evidence under Rule 403 as unfairly prejudicial even if the judge has concluded that a reasonable jury could find by a preponderance of the evidence that it is authentic, and therefore relevant.

In *Johnson v. Elk Lake School District*,[90] the Third Circuit discussed the appropriate roles for the trial judge and the jury under Federal Rules of Evidence 104(a) and 104(b) with respect to admissibility of evidence under Rule 415. Rule 415 governs when evidence of the adverse party's prior uncharged sexual assault or child molestation may be admitted, in a civil case involving a claim for relief based on a party's alleged sexual assault or child molestation.[91] Specifically, the Third Circuit considered whether the trial judge had to first make "a preliminary finding by a preponderance of the evidence under Federal Rule of Evidence 104(a) that the act in question qualifies as a sexual assault and that it was committed by the defendant."[92] The Third Circuit ruled that the trial court need not make such a preliminary finding. Instead, it held that "the court may admit the evidence so long as it is satisfied that the evidence is relevant, with relevancy determined by whether a jury could reasonably conclude by a preponderance of the evidence that the past act was a sexual assault and that it was committed by the defendant," citing Federal Rule of Evidence 104(b).[93]

Importantly, however, the court added:

---

[89] *See* Grossman et al., *supra* note 7; Delfino, *supra* note 7; *see also* Nash v. United States, 54 F.2d 1006, 1007 (2d Cir. 1932) (Judge Learned Hand referencing limiting instructions) ("[T]he recommendation to the jury of a mental gymnastic which is beyond, not only their powers, but anybody's else.").

[90] 283 F.3d 138, 144–45 (2002).

[91] *Id.*

[92] *Id.* at 143–44.

[93] *Id.* at 144.

> We also conclude . . . that even when the evidence of a past sexual offense is relevant [(i.e., satisfies Rule 104(b)[94]], the trial court retains discretion to exclude it under Federal Rule of Evidence 403 if the evidence's probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.[95]

Similarly, in *Huddleston v. United States*,[96] which was cited by the Third Circuit in *Johnson*, the Supreme Court considered the proper roles of (1) the trial judge under Federal Rule of Evidence 104(a), and (2) the jury as the decider of disputed facts. In particular, the Court addressed the jury's role as trier of fact under Rule 104(b) in connection with the admissibility of evidence of "other crimes, wrongs, or acts" under Rule 404(b).[97] Writing for the Court, Chief Justice Rehnquist rejected the petitioner's argument that the trial judge was required by Rule 104(a) to make a preliminary determination that the defendant committed a similar act before allowing it to be admitted.[98] Rather, the Court held that the only requirement for admission of Rule 404(b) evidence is that it be relevant, which only occurs "if the jury can reasonably conclude that the act occurred and that the defendant was the actor. . . . Such questions of relevance conditioned on a fact are dealt with under Federal Rule of Evidence 104(b)."[99]

Chief Justice Rehnquist then addressed the issue of whether "unduly prejudicial evidence might be introduced under Rule 404(b)," concluding that the protection against such an outcome lies in "the assessment the trial court must make under Rule 403 to determine whether the probative value of the similar acts evidence is substantially outweighed by its potential for unfair prejudice."[100] Thus, as in *Johnson*, the *Huddleston* Court recognized that the trial judge retained the authority under Rule 403 to exclude evidence even when a reasonable jury could conclude, by a preponderance of the evidence under Rule 104(b), that it was relevant.[101]

Johnson and *Huddleston* involved evidence about prior acts and convictions,[102] which, it can be argued, is of relatively low probative

---

[94] *Id.* at 155.

[95] *Id.*

[96] 485 U.S. 681 (1988).

[97] *Id.* at 682; *see also Johnson*, 283 F.3d at 144–45.

[98] *Huddleston*, 485 U.S. at 682.

[99] *Id.* at 689.

[100] *Id.*

[101] *Id.* at 687.

[102] *Id. at* 682.

value. But when challenged AIM goes to the heart of the matter, such as is the case with most of the alleged deepfakes in our hypothetical, a strong argument can be made that the probative value is much greater if the evidence is found to be authentic.

Regarding unfair prejudice, evidence admitted under Rule 403 of prior actions and convictions presents cause for concern that a jury will find against a party based not on the facts of the current action, but rather on a character inference based on past actions. This is particularly troubling because it is difficult for a party to rebut or mitigate the jury's tendency to use evidence of prior actions and convictions in this way. Thus, not only can the evidence be prejudicial, but the prejudice can also be "unfair."

In the case of audiovisual evidence, as discussed above, studies routinely show that it can have a tremendous impact on juror perception and memory, even when a juror understands that the evidence may be or is likely fake.[103] One experiment showed that people were more likely to confess to acts that they had not committed when they were presented with doctored videos ostensibly showing them engaging in the act.[104] In a subsequent study, researchers found that participants presented with doctored videos were more likely to sign witness statements accusing their peers of cheating than those who were simply told about the alleged infractions.[105] Moreover, the participants in the study were aware that their statements would be used to punish their peers.[106] These studies clearly show that "video evidence powerfully affects human memory and perception of reality."[107]

Thus, there is potentially substantial prejudice to the party objecting to the deepfake, and given the potential impact even when the deepfake is strongly suspected to be fake, it is straightforward to find that this is unfair. On the other hand, as will be discussed in later sections, the defendant will have the opportunity to challenge the evidence using various tools of discovery and expert witness testimony, possibly mitigating the unfairness of the alleged deepfake being presented to the jury. But in some circumstances, it might be very difficult to rebut or

---

[103] *See* Grossman, et al., *supra* note 7; Delfino, *supra* note 7.

[104] Yael Granot, Neal Feigenson, Emily Balcetis & Tom Tyler, *In the Eyes of the Law: Perception Versus Reality in Appraisals of Video Evidence*, 24 PSYCH. PUB. POL'Y & L. 93, 97–98 (2017) (citing Robert A. Nash & Kimberley A. Wade, *Innocent but Proven Guilty: Eliciting Internalized False Confessions Using Doctored-Video Evidence*, 23 APPLIED COGNITIVE PSYCH., 624–637 (2008) (describing a study where innocent participants were shown doctored videos of them illicitly taking money placed in front of them were more likely to confess having done so and internalize the belief that they did so).

[105] Kimberley A. Wade, Sarah L. Green & Robert A. Nash, *Can Fabricated Evidence Induce False Eyewitness Testimony?*, 24 APPLIED COGNITIVE PSYCH. 899, 900 (2010).

[106] *Id.*

[107] Delfino, *supra* note 7, at 311.

mitigate the jury's tendency to be swayed by a possible deepfake, even one that the jury determines is not real, thus supporting the argument that allowing it to go to the jury would be unfair.

In summary, it is our contention that with respect to AIM deepfakes, the judge should not submit the challenged AIM evidence to the jury if the judge determines that its probative value is substantially outweighed by unfair prejudice to the objecting party. This holds true even if the judge determines that a reasonable jury could determine that the challenged evidence is authentic by a preponderance of the evidence. In addition to the text of the Rules, there is ample case law to support the proposition that judges can exclude unfairly prejudicial evidence, even where relevance and authenticity are established. Additionally, Federal Rule of Evidence 102 states that the rules of evidence should be construed to promote fairness, develop evidence law, ascertain the truth, and secure a just determination.[108] There is ample authority for our approach to handling alleged AIM under the Federal Rules of Evidence.

VI. APPLYING THE GPTJUDGE FRAMEWORK FOR RESOLVING AUTHENTICITY DISPUTES

How should a court handle the allegedly fake AIM evidence at the heart of Connie's and Eric's claims and defenses? Recall that it is August 4, 2028, merely three months before the 2028 U.S. presidential election. Connie and her lawyers want to push the case forward rapidly. Eric may have less interest in moving quickly, as, for the most part, the status quo seems to benefit him and his campaign. If the court follows traditional scheduling practices, the risk is that the evidentiary issues related to the allegedly fake AIM evidence will not be fully developed, potentially derailing a trial, and producing a less-than-optimal outcome from the perspective of the parties, the judge, and the public. How can the court prevent this?

Below, we analyze aspects of the hypothetical litigation between Connie and Eric by applying the framework for addressing allegedly fake AIM evidence developed by Maura R. Grossman, Hon. Paul W. Grimm (ret.), Daniel G. Brown, and Molly (Ximing) Xu in their article, *The GPTJudge: Justice in a Generative AI World*.[109] This framework provides the parties and the court with a step-by-step roadmap for administering and ruling on admissibility challenges to alleged AIM evidence. While Grossman and co-authors provide a general framework for

[108] FED. R. EVID. 102 ("These rules should be construed so as to administer every proceeding fairly, eliminate unjustifiable expense and delay, and promote the development of evidence law, to the end of ascertaining the truth and securing a just determination.").

[109] Grossman, et al., *supra* note 7, at 19.

dealing with AIM as evidence, in this article, we focus specifically on authenticity challenges to the introduction of *audiovisual* evidence—evidence that one party puts forth as genuine and the other contests as being a "deepfake."

Recognizing the problems posed by audiovisual evidence that is alleged to be AIM, we show how judges can be proactive in addressing admissibility challenges under the current Federal Rules of Evidence. In as much as our hypothetical deals with a civil trial in federal court, we suggest the use of pretrial conferences under Federal Rule of Civil Procedure 26(f) (between the parties) and Rule 16 (with the court) to allow the parties to disclose their intention to proffer audiovisual evidence and to raise evidentiary challenges thereto so that the parties can seek discovery to obtain the relevant facts to address their competing views about the authenticity of the possibly AIM evidence.

We also suggest that judges schedule such an evidentiary hearing well in advance of trial, so that the challenging party can present the factual basis for their evidentiary challenge and the proponent of the evidence can respond. Finally, we suggest that judges rule on the admissibility of the potential AIM evidence by drawing on the factual record presented at the hearing while also asking the parties to address the potential applicability of Rule 403 to exclude relevant evidence when it creates a risk of unfair prejudice to the opposing party, confusion of the issues, misleading the jury, or delay.

A. Pretrial Conferences Pursuant to Federal Rules of Civil Procedure 26(f) and 16

There are two types of pretrial conferences under the Federal Rules of Civil Procedure, which set the stage for the court to manage the AIM admissibility issues presented in the hypothetical: (i) the Rule 26(f) conference between the parties, and (ii) the Rule 16 conference with the court. These conferences serve largely administrative purposes leading up to trial and provide an opportunity for the parties to discuss their respective plans for discovery. These conferences also allow the parties and the court to determine, well before trial, the appropriate scope and timelines for discovery, including discovery to resolve any evidentiary challenges that will be made. The conferences also provide the court with an early indication of the types of evidence both parties intend to present and to identify any related evidentiary challenges, including the assertion that any evidence is fake AIM.

Rule 26(f) requires that "the parties must confer as soon as practicable—and in any event at least twenty-one days before a scheduling

conference is to be held or a scheduling order is due under Rule 16(b)."[110] Notably, Rule 26(f)(1) provides an exception for the court to order otherwise,[111] but delay is inadvisable in most circumstances involving extensive discovery or complex evidentiary issues, such as those involving AIM. The court has considerable flexibility to adjust the timing of conferences, submissions by the parties, and exchanges of discovery materials, including information that the parties intend to use to support their claims and defenses.

During the Rule 26(f) conference, Connie will likely discuss the basis for her Voting Rights Act and defamation claims, and Eric will likely raise his related defenses. Both parties will raise the audiovisual evidence that they are likely to use to support their respective claims and defenses. This will serve as an early indication of potential admissibility challenges and will likely also trigger additional discovery requests for corroborating evidence to support each of their positions.

The parties are required to prepare a discovery plan, which, among other things, calls on them to agree on the timing for initial disclosures under Rule 26(a), areas where discovery is needed, and when such discovery should be completed.[112] Initial disclosures include the production or description of electronically stored information (ESI) that will be used by the parties to support their claims and defenses (other than for impeachment), which in turn must be made within fourteen days of the initial Rule 26(f) conference.[113] Therefore, in connection with their initial disclosures, Connie and Eric would be required to disclose the existence and location, or produce copies, of the video and audio recordings they intend to use at trial, including those alleged to be fake AIM.

In most cases, the court should schedule one or more pretrial conferences with the parties to establish "early and continuing control so that the case will not be protracted because of lack of management."[114] The court's scheduling order will outline deadlines for Connie and Eric to disclose the nature of the evidence that supports their claims and defenses and should also outline any deadlines to challenge such evidence and seek additional discovery to support such a challenge. During the conference, the judge (undoubtedly alerted to the existence and importance of the audiovisual evidence from having read the allegations in the pleadings) can ask both parties if they intend to challenge the other's introduction of audiovisual evidence on the grounds that it is fake AIM. At this point, both parties would have already had a chance

---

[110] FED. R. CIV. P. 26(f)(1).

[111] *Id.*

[112] FED. R. CIV. P. 26(f)(3).

[113] FED. R. CIV. P. 26(a)(1)(A)(ii); *see also* FED. R. CIV. P. 26(a)(1)(C).

[114] FED. R. CIV. P. 16(a)(2).

to confer with each other and should be aware of the possibility of challenges to each other's introduction of audiovisual evidence.

Connie will attempt to satisfy her burden of showing that Eric and his supporters knowingly published false and defamatory statements about her by publishing the allegedly fake videos and audio of her. Eric will challenge Connie's introduction of the audio conversation between John Doe and Eric, which allegedly supports Connie's allegation that the two coordinated the publication of false and defamatory statements about Connie. Connie will also introduce the recording from three years prior of Eric speaking to a staffer about the possibility of using deepfakes in a senate campaign, which Eric will challenge as fake. Both parties will seek additional discovery to support their admissibility challenges, and the judge's scheduling order should include a deadline for the completion of such discovery and the filing of motions to exclude evidence.[115] The court should set an evidentiary hearing date to rule on admissibility challenges.[116] The hearing should allow both parties "to develop the facts necessary to rule on the admissibility of the challenged evidence."[117]

B. Developing a Factual Record

Both Connie and Eric will use various discovery tools available to them under the Federal Rules of Civil Procedure to establish a factual basis to support the introduction of their own audio and audiovisual evidence and to challenge the evidence that is detrimental to their claims. Connie's goal will be to find corroborating evidence to show that the audio and video recordings of her at election sites are fake AIM. Similarly, Eric will try to establish that the audio and video recordings are genuine and accurately reflect the events they purport to memorialize. The reverse is true for the case of the audio recording of the conversation between Eric and John Doe and Eric and his staffer three years prior.

Connie can use several different tools to establish a factual record for the evidentiary hearing. One approach would be to supply corroborating evidence suggesting that the audio and video recordings are fake. She can support this contention with alibi testimony or geolocation data showing that she was not at the location suggested at the time the videos and audio were allegedly recorded. For instance, evidence that places her at a different location at the time the video was recorded, as memorialized by its metadata, would be particularly helpful to her in

---

[115] *See* Grossman, et al., *supra* note 7, at 15.

[116] *See id.* at 16.

[117] *Id.*

supporting her authenticity challenge against the introduction of Eric's evidence.

She can also seek to introduce expert testimony or evidence from deepfake detection (DD) tools to lend credibility to her arguments that the evidence in question is not authentic (keeping in mind the likely difficulty of showing that the DD tool used is valid and reliable). Connie can also subpoena Eric's phone records to establish a connection between him and John Doe to lend further credibility to her argument that the conversation indicating collusion between the two of them was genuine since the telephone records show the two men speaking at the time suggested by the metadata of the recording.

Similarly, Eric will try to establish that the incriminating videos of Connie with the Chinese embassy official are authentic. He might try to subpoena photos of the two together at other times and places. Eric will also try to find witnesses to authenticate Connie's voice in the purported audio recording of her soliciting bribes. He may subpoena Connie's phone records and bank statements to try to establish connections between Connie and the foreign agents with whom she is alleged to have connections. Eric may also try to depose Connie's close aides and campaign officials to establish a relationship between Connie and the Chinese embassy official she is seen with in the compromising video. If Eric can develop adequate corroborating evidence to show that Connie had a relationship with the Chinese embassy official, or that shows her at the locations at the time when the video was alleged to have been recorded, he will be able to make a strong argument in favor of the authenticity of the audio and video evidence. Like Connie, Eric will likely also make use of the results of AIM DD tools and expert testimony to argue that the audiovisual evidence is authentic, although given questions about the validity and reliability of deepfake detectors, the court will need to closely scrutinize any such evidence and the experts who offer it.

Finally, it is inevitable that both Connie and Eric will retain experts to support their positions. This will result in expert disclosure of their opinions and their factual bases, materials reviewed, past testimonial experience and publications,[118] and almost certainly their depositions.[119] When highly technical and specialized evidence is central to the case, it also can be expected that Connie and Eric will assess whether they believe they can exclude all or important portions of the other's experts' testimony, by filing *Daubert*[120] motions challenging the

[118] FED. R. CIV. P. 26(a)(2).

[119] FED. R. CIV. P. 30.

[120] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 578 (1993); *see also* FED. R. EVID. 702.

qualifications, factual sufficiency, methodology, and conclusions of the opposing experts.

C. Evidentiary Hearing

1. Scheduling

Trial judges should schedule an evidentiary hearing well in advance of trial to allow both parties ample opportunity to develop a factual record and to challenge the opposing parties since the ultimate resolution of these issues will likely play a vital role in the disposition of the litigation.[121] Both Eric and Connie need to show that the audiovisual evidence they are introducing meets the evidentiary requirements of both relevance and authenticity. Eric will introduce the incriminating audio and video evidence linking Connie to the Chinese embassy official and to the individuals from whom she is alleged to have solicited bribes. Connie will introduce the audio recording between Eric and John Doe and Eric and his staffer from three years prior. Eric will object that the audio recordings are fake AIM.

2. The judge's ruling on the evidentiary issues

After having heard all the evidence presented by Connie, Eric, John Doe, and PoliSocial, the judge will need to evaluate the evidence and rule on the contested issues. As previously noted, this ruling (whether made orally "from the bench" or in writing) should be as factually specific and legally comprehensive as possible, both to guide the future conduct of the trial, and to provide any reviewing appellate court with a clear explanation of what the ruling is, and why.[122] We will focus first on the issues Connie likely will raise, then those by Eric, John Doe, and PoliSocial.

a. Connie's arguments that the audiovisual evidence of her relationship with the Chinese embassy official, her soliciting money, and her constituents stuffing ballot boxes are deepfakes

Reduced to its essentials, Connie will argue that the audiovisual evidence showing her illicit relationship with the Chinese official, her solicitation of money, and her constituents stuffing ballot boxes (collectively, "the Audiovisual Evidence") are deepfakes, that the events they purport to show did not occur, and that Eric, John Doe, and PoliSocial are responsible for disseminating them to the public, thereby defaming

[121] *See* Grossman, et al., *supra* note 7, at 16.

[122] FED. R. EVID. 103(C).

her. What is unique about Connie's position, however, is that she will *not* seek to exclude this evidence. Rather, she must introduce it to prove the falsity of its contents to demonstrate defamation.

In this instance, the judge will find that relevance is easily established. The audiovisual evidence is essential to proving key elements of Connie's defamation claim. Because the Audiovisual Evidence is relevant and has been challenged as fake AIM, the judge will need to assess the evidence Connie proffered during the hearing to show it is a deepfake. Although deepfake technology is rapidly evolving, and at the moment it can be nearly impossible to determine if deepfakes are legitimate or not, it is nonetheless likely that Connie will call on an expert to testify that the audiovisual evidence is fake. This will require the judge to evaluate the factors from Federal Rule of Evidence 702, as further amplified by the *Daubert* factors.[123] Importantly, as the amendments to Rule 702 from December 1, 2023 make clear, the judge must find that Connie has met her burden by a preponderance of the evidence before admitting her expert's testimony.[124]

The judge will first assess whether Connie's expert has sufficient knowledge, training, education, and experience to testify and whether their testimony will be helpful to the jury in deciding the case.[125] Assuming Connie has hired a legitimate expert, the judge will have little difficulty deciding that the expert is qualified and that their evidence will be helpful to the jury. Next, Rule 702 requires Connie to demonstrate (again, by a preponderance) that her expert considered sufficient facts or data to support their opinions, that the methodology they used to reach their opinions was reliable, and that those methods and principles (themselves reliable) were reliably applied to the facts of the case.[126]

With regard to the reliability prongs, the judge will be guided by the *Daubert* factors: whether the methodology used by the expert in reaching their opinions has been tested; if so, whether there is a known error rate associated with the methodology; whether the methods and principles relied on by Connie's expert are generally accepted as reliable

---

[123] *See Daubert*, 509 U.S. at 591–95; *see also* FED. R. EVID. 702 advisory committee notes to 2000 amendments ("*Daubert* set forth a non-exclusive checklist for trial courts to use in assessing the reliability of scientific expert testimony. The specific factors explicated by the *Daubert* Court are (1) whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the scientific community.").

[124] FED. R. EVID. 702.

[125] *Id.*

[126] *Id.*

by other experts in the same field; whether the methodology used by Connie's expert has been subject to peer review; and whether there are standard accepted procedures for using the methodology, and whether the expert complied with them.[127]

The judge will also consider any expert testimony offered by Eric, John Doe, and PoliSocial to undermine Connie's expert in deciding whether to allow Connie's expert to testify to the jury. In this regard, the judge's focus is not on the *correctness* of Connie's expert's opinions, but rather on whether they were qualified, considered sufficient facts, used reliable methodology, reliably applied it to the facts of the case, and complied with any generally accepted protocols related to the DD methodology selected.[128] The judge will evaluate the defendants' expert's testimony the same way in which they evaluate Connie's expert's.

If the judge is persuaded that Connie and the defendants have met their foundational requirements by a preponderance of the evidence, they will both be allowed to testify at trial, and it will be up to the jury to decide which expert's testimony it accepts (if any). It should be obvious that Connie and the defendants will be wise to retain qualified experts who carefully comply with the *Daubert* and Rule 702 factors, particularly since current DD methods may readily be challenged. If the judge concludes that either (or both) experts failed to meet these requirements, the judge will exclude them from testifying at trial.

Finally, it is worth noting that Connie's and the defendants' experts will be selected by them, and they are not likely to offer an expert who does not express opinions consistent with their litigation positions. This means that in real life, the parties' experts will not be testifying as independent technologists or scientists, but more realistically, as paid advocates. Since the cost of the judge appointing a court expert under Rule 706 is typically prohibitive, and the court has no funds to do so on its own, the judge may find themself in "a battle of wits unarmed"[129]—lacking sufficient knowledge of the technical issues to evaluate all the Federal Rule of Evidence 702 and *Daubert* factors effectively (including the particular DD methods applied). The best way for the judge to avoid this is to make it clear during the pretrial conference what the judge will expect the experts to address during their testimony, and the

---

[127] See FED. R. EVID. 702 advisory committee's notes to 2000 amendments, for a description of additional factors, including whether the expert accounted for any alternative explanations; whether the methodology used by the expert exists for a purpose other than litigation, and if so, was the same degree of rigor used by the expert for litigation purposes as they would use for a non-litigation purpose; and whether there any unjustified analytical leaps the expert made that are not justified by the facts considered and the methodology used.

[128] *See Daubert*, 509 U.S. 579 at 591–95.

[129] Brad Sylvester, *Fact Check: Did Shakespeare Say a Quote About a 'Battle of Wits'?*, CHECKYOURFACT (Aug. 13, 2019, 1:21 PM), https://checkyourfact.com/2019/08/13/quote-shakespeare-battle-wits-unarmed/ [https://perma.cc/P89Q-33NF].

materials they will rely on to support it. The judge should require that the materials be produced well ahead of the hearing, so that he or she will have sufficient time to review them in advance of the hearing and be as prepared as possible to question the experts during their testimony.[130] At least one judge has held a "science day" in which they were able to learn from the experts in a more informal setting[131] and we encourage that practice in the case of expert testimony concerning DD technology.

b. The Defendants' Motion to Exclude incriminating audio evidence

Defendants will argue that the incriminating audio evidence (of Eric and John Doe, and the recording of Eric and his staffer three years prior) raises the very issues we addressed above regarding conditional relevance; the judge's role under Rule 104(a); the jury's role as decider of contested facts under Rule 104(b); and, especially, Rule 403, which nonetheless allows the judge to exclude relevant evidence if its probative value is substantially outweighed by the danger of unfair prejudice.

Preliminarily, it will be Connie's burden to introduce evidence sufficient for the judge to conclude that the jury reasonably could find by a preponderance of evidence (more likely than not) that the incriminating audio recordings are authentic. And the defendants will have the burden of introducing evidence that it is not Eric's or John Doe's voice on the recordings—it is fake AIM. The judge will consider all the evidence submitted by Connie and the defendants. If the judge determines that Connie failed to meet her burden, then the recording will not be found to be authentic, and therefore not relevant, and it will be excluded. But assuming that Connie's evidence is sufficient for the jury to find that the recording reflects Eric and John Doe's voices, by a preponderance of the evidence, then the jury will hear the evidence, unless the judge finds under Federal Rule of Evidence 403 that its introduction will result in unfair prejudice that substantially outweighs its probative value.

While the Federal Rules of Evidence generally disfavor judicial gatekeeping of evidence at the admissibility stage, there are nonetheless instances where the exclusion of deepfake evidence is likely warranted under Rule 403. Ultimately, the text of Rule 403 provides trial judges with explicit, albeit limited, gatekeeping power to exclude evidence that is relevant, but whose probative value is substantially

---

[130] *See, e.g.*, FED. R. EVID. 614(b).

[131] *See, e.g.*, In re Glucagon-like Peptide-1 Receptor Agonists (GLP-1 RAS) Products Liability Litigation, No. 2:24-md-03094-GEKP, Case Management Order No. 5 (E.D. Pa. Apr. 26, 2024) (scheduling a "Science Day" on June 14, 2024, during which the parties "will, in an objective format, provide the Court with an overview of certain medical and scientific issues . . .").

outweighed by its potential for unfair prejudice. We can represent the Rule 403 analysis on a continuum for each evidentiary situation presented. On one end of the continuum, the potential for unfair prejudice is at a maximum and probative value is at a minimum, leading to exclusion of the evidence. On the other end of the continuum, the potential for unfair prejudice is at a minimum and probative value is at a maximum, leading to admission of the evidence. We can use this continuum to analyze the Rule 403 balancing test for the examples in our hypothetical.

The audio recording of Eric discussing the use of deepfakes from a prior election campaign nearly three years ago likely meets the relevance and authenticity requirements because it arguably supports Connie's claims. Eric will argue that the audio relating to the prior election should be excluded under Rule 404(b) to the extent that it is being used as character evidence to suggest that he engaged in the same conduct in the current election campaign.[132] Connie can rebut by arguing that it should be permitted under Rule 404(b)(2) because it proves that Eric had the knowledge, capability, and plan to make and use deepfakes.[133]

Additionally, Eric will argue that this audio recording is fake AIM and will attempt to present evidence to support his case. But it might be difficult to find alibi evidence, especially if it is not clear when the recording was made. In addition, Eric will counter that even if authentic, the evidence of him discussing deepfakes in a prior election does not necessarily mean that he used the deepfakes in that campaign either. For one, the audio simply captures him talking about deepfake technology generally and there is no specific discussion of using deepfakes against his opponent. He would further dispute the implication that such evidence relating to a state senate election proves that he would use deepfakes in a future presidential campaign.

With regard to the recording between Eric and a staffer three years prior, Eric can make a strong argument that the audio recording should be excluded under Rule 403 because its probative value is relatively low and is substantially outweighed by being highly unfairly prejudicial. The audio, while relevant, has lower probative value because it relates to a prior election, not facts at the heart of this case. Moreover, it is likely to be unfairly prejudicial to the defendant because admitting it will predispose the jury to draw conclusions on his actions based on prior acts and Eric's character, likely even if the jury agrees with Eric

---

[132] *See* FED. R. EVID. 404(b)(1) ("Evidence of any other crime, wrong, or act is not admissible to prove a person's character in order to show that on a particular occasion the person acted in accordance with the character.").

[133] *See* FED. R. EVID. 404(b)(2) ("This evidence may be admissible for another purpose, such as proving motive, opportunity, intent, preparation, plan, knowledge, identity, absence of mistake, or lack of accident.").

that it is fake. Both *Johnston* and *Huddleston* support the judge exercising authority to exclude evidence of prior acts and convictions when the probative value is low and unfair prejudice is high.[134]

The judge's ruling on the audio recording of the conversation between John Doe and Eric will be a closer call, falling somewhere in the middle of our continuum. Each of these decisions is highly fact-dependent, of course, and a real situation will involve rich facts that are developed by the parties through discovery. In our hypothetical, the audio of John Doe and Eric, if authentic and relevant, is a "smoking gun" supporting Connie's claims that the defendants created and published the explosive audio and videotapes that defamed her. But, at the same time, it is devastatingly prejudicial to the defendants, and arguably unfairly prejudicial if it is likely fake and sways the jury nonetheless. The judge will be especially careful to look at the totality of facts that show that it was John Doe on the tape and other evidence developed and presented by Eric and Connie. Are there credible testifying witnesses who are familiar with Eric and John Doe's voices? Does geolocation information place them at the place and at the time where the recording was made? Are there any other corroborating facts to support Connie's position that it is Eric and John Doe speaking? What is the nature and quality of Eric's and John Doe's evidence that it was not them? A mere denial? A credible alibi including witnesses that could establish that neither could possibly have made the recording at the time and place where Connie claims it was made?

For instance, if Eric has strong evidence supporting his contention that the voices on the recording could not possibly have been his or John Doe's, he can present a stronger argument for exclusion, that unfair prejudice substantially outweighs the probative value of the audio. Such evidence could be established in the form of alibis placing him and John Doe at different locations at the time the audio was purportedly recorded. An example of such a scenario might be that either Eric or John Doe was unconscious or undergoing surgery at the time of the recording and thus convincing evidence corroborates that they could not have been the subjects of the audio recording. On unfair prejudice, Eric could argue that the contents of the recording were inflammatory because they show him using disparaging remarks towards his opponent. Eric's alleged quip suggesting that women are not fit to be president arguably bolsters his argument. He can argue that the audio should be excluded because it would leave a lasting negative impression on the jury in a way that could lead them to rule based on their emotions rather than the merits of the case, even if the jury finds that the audio is

[134] *See* Johnson v. Elk Lake Sch. Dist., 283 F.3d 138, 155 (2002); Huddleston v. United States, 485 U.S. 681, 687 (1988).

likely fake. Depending on the specific facts, the better argument could go either way, although the parties and court must always consider the preference for admitting evidence and the requirement that unfair prejudice must "substantially outweigh" probative value to exclude relevant evidence under Rule 403.

Finally, what if Connie were to seek to admit an audio recording of Eric directing John Doe to create deepfakes of Connie in the upcoming election? If Eric objects that the recording is itself a deepfake, could he make a colorable argument for exclusion under Rule 403? Such evidence would be highly probative because it relates to the current matter, not an earlier election. Unlike the evidence relating to a previous election, this audio would be at the heart of the case. Eric could argue that the jury will find that it is a deepfake but nevertheless be swayed by the audio, pointing to the research showing the power of audiovisual evidence, even that which a jury determines is fake. Is this unfair prejudice, and does it substantially outweigh the probative value of such an audio recording? Alone, almost certainly not, especially if Connie produces any corroborating evidence, such as geolocation evidence and phone records. If Eric presents nothing more than the assertion that the audio recording is AIM, a serious concern would be that by deeming it inadmissible, the judge would be supplanting the jury as fact finder. Additionally, Eric will have every opportunity to take discovery to demonstrate that the audio is AIM, such as by showing that Eric and John Doe were not at the locations suggested by the audio. Eric can also hire an expert witness and take other steps to attack the chain of custody and other indicators of the genuineness of the audio recording. If Eric cannot produce strong evidence that the audio is fake, he will have only a weak argument that the audio recording is unfairly prejudicial. In scenarios like this one, there is not a strong argument to exclude the alleged AIM under Rule 403, although we emphasize that this is a fact-dependent inquiry. In the right case, the parties will be able to develop the record and present arguments that could tip the balance in favor of exclusion under Rule 403.

VII. ALTERNATIVE APPROACHES: RULE CHANGE PROPOSALS

Given the complexities and challenges presented by AIM, there are growing calls to amend the Federal Rules of Evidence. In this section, two approaches to modifying the rules are discussed. In thinking about such proposals, it is important to consider that rule changes are infrequent and often take years to materialize.

A. Burden Shifting Approach Towards Admissibility

Two authors of this paper (Grimm and Grossman) have proposed a modification to Rule 901 for possible deepfakes.[135] They suggest a separate rule because of the increasing difficulty of differentiating between authentic audiovisual evidence and fabricated or altered audiovisual evidence. This is particularly true in instances where, as in our proposed hypothetical, one party introduces audiovisual evidence and the other challenges its admissibility on the grounds that it is AIM. Under the existing Rules, the proponent of evidence challenged as AIM might choose to authenticate an audio recording under Rule 901(b)(5) (opinion as to voice) or Rule 901(b)(3) (comparison of evidence known to be authentic with other evidence the authenticity of which is questioned). These are easier routes to authentication than Rule 901(b)(9), which focuses on evidence generated by a "system or process." Grimm and Grossman propose a new Rule 901(c):

> 901(c): *Potentially Fabricated or Altered Electronic Evidence.*
>
> If a party challenging the authenticity of computer-generated or other electronic evidence demonstrates to the court that it is more likely than not either fabricated, or altered in whole or in part, the evidence is admissible only if the proponent demonstrates that its probative value outweighs its prejudicial effect on the party challenging the evidence.

The proposed rule does not use the word "deepfake," because it is not a technical term. Instead, the proposed rule describes the evidence as being either computer-generated (which encompasses AI-generated evidence) or electronic evidence, which encompasses other forms of electronic evidence that may not be AI generated (such as digital photographs, or digital recordings).

The proposed rule puts the initial burden on the party challenging the authenticity of computer generated or electronic evidence to make a showing to the court that it is more likely than not either fabricated or altered in whole or part. This standard is similar to the showing required by the proponent of scientific, technical, or specialized evidence under newly revised Rule 702. It requires the challenging party to

---

[135] *See* Advisory Comm. on Evidence Rules, Agenda for Committee Meeting (Apr. 2024), https://www.uscourts.gov/sites/default/files/2024-04_agenda_book_for_evidence_rules_meeting_final_updated_5-8-2024.pdf [https://perma.cc/KB9R-AV96]; *see also Symposium On Scholars' Suggestions For Amendments, And Issues Raised By Artificial Intelligence*, 92 FORDHAM L. REV. 2375 (2024); Daniel J. Capra, *Deepfakes Reach the Advisory Committee on Evidence Rules*, 92 FORDHAM L. REV. 2491 (2024).

produce evidence to support the claim that it is fabricated or altered. Mere conclusory allegations are insufficient. If the challenging party makes the required showing, the burden shifts to the proponent of the challenged evidence to show that its probative value outweighs its prejudicial effect on the party challenging the evidence. This is the same showing required by Rule 609(a)(1)(B) (regarding attacking character for truthfulness by introducing evidence of a criminal conviction), and is a lesser showing than a "reverse balancing" test such as used in Rule 609(b)(1) (regarding the introduction of a criminal conviction from more than ten years in the past) or Rule 703 (regarding introducing to the jury inadmissible facts or data relied upon by an expert).

If the party objecting to the evidence as being AIM fails to make the showing to the court that it more likely than not is fabricated or altered, then the court will allow the proponents' evidence and the opposing party's evidence to go to the jury under Rule 104(b). But, if the opposing party makes the required showing and the proposing party fails to show that the probative value of the challenged evidence outweighs its prejudicial effect on the challenging party, the court will exclude the evidence under Rule 104(a).

B. Expanding Judicial Gatekeeping

Rebecca Delfino proposes yet another formulation of Rule 901(c) that "would expand the gatekeeping function of the court by assigning the responsibility of deciding authenticity issues solely to the judge."[136] The primary function of the proposed rule is to recategorize the authentication of audiovisual evidence from conditional relevance under Rule 104(b) to relevancy under Rule 104(a).[137] Such an expansion of the judge's role in determining admissibility is justified primarily by the nature of the threat that deepfake evidence poses.[138] Delfino argues that such forms of evidence are "technically complex and highly prejudicial to jury deliberations."[139]

Of particular concern is that deepfakes are intentionally designed to deceive and alter the perceptions of the viewer. Add to this the problem that deepfakes are increasingly sophisticated enough that they are virtually undetectable, even by experts. Given that authenticating

---

[136] *See* Delfino, *supra* note 7, at 341. Delfino's proposed version of the rule is "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."

[137] *Id.*

[138] *Id.* at 342.

[139] *Id.* at 345.

deepfakes will likely exceed the jury's capabilities, Delfino argues for expanding the judge's role "as the preliminary factfinder to protect the integrity of the jury's deliberations."[140] As Grossman, Grimm, and Brown noted, a change that "eliminate[s] the role of the jury in determining the authenticity of digital and audiovisual evidence in response to the appearance of deepfakes . . . would involve a substantial departure from the current evidentiary framework . . . making it infeasaible as a practical solution."[141] The Grimm-Grossman proposal aims to address the same concerns without veering as far from longstanding legal precedent.

At its April 19, 2024, meeting, however, the Advisory Committee on the Federal Rules of Evidence considered both of these proposals, among others,[142] and determined not to make any rules changes at this time.[143]

VIII. CONCLUSION

Given the ease with which anyone can create a convincing deepfake, courts should expect to see a flood of cases in which the parties allege that evidence is not real, but AI generated. Election interference is one example of a national security scenario in which deepfakes have important consequences. There is unlikely to be a technical solution to the deepfake problem. Most experts agree that neither watermarks nor deepfake detectors will completely solve the problem, and human experts are unlikely to fare much better. Courts will have no option, at least for the time being, other than to use the existing Federal Rules of Evidence to address deepfakes. The best approach will be for judges to proactively address disputes regarding alleged deepfakes, including through scheduling conferences, permitted discovery, and hearings to develop the factual and legal issues to resolve these disputes well before trial.

---

[140] *Id.* at 346 ("Thus, [deepfake] evidence presents the same risk as other highly prejudicial or technical evidence, where case law and commentators have recognized that judges should act as the preliminary factfinder to protect the integrity of the jury's deliberations. The reality of deepfakes requires that we acknowledge the limits of our trust in the jury.").

[141] Grossman, et al., *supra* note 7, 16.

[142] Capra, supra note 135; *see also* Advisory Comm. on Evidence Rules, Agenda for Comm. Meeting (Apr. 2024), https://www.uscourts.gov/sites/default/files/2024-04_agenda_book_for_evidence_rules_meeting_final_updated_5-8-2024.pdf [https://perma.cc/KB9R-AV96].

[143] Nate Raymond, *US Judicial Panel Wrestles With How to Police AI-Generated Evidence*, REUTERS (Apr. 19, 2024, 5:35 PM), https://www.reuters.com/legal/transactional/us-judicial-panel-wrestles-with-how-police-ai-generated-evidence-2024-04-19/ [https://perma.cc/5PXQ-WB6E]; *see also* Advisory Comm. on Evidence Rules, Minutes of the Meeting of Apr. 19, 2024, at 110–112 (June 4, 2024), https://www.uscourts.gov/sites/default/files/2024-06_agenda_book_for_standing_committee_meeting_final.pdf [https://perma.cc/JR6W-RS7N].

Even as several scholars propose to amend the Federal Rules of Evidence in recognition of the threat posed by deepfake evidence, such changes are unlikely in the near future. Meanwhile, trial courts will require an interim solution as they grapple with AIM evidence. Rule 403 will play an important role, as the party against whom an alleged deepfake is proffered may be able to make a compelling argument that the alleged deepfake should be excluded because the probative value of the alleged deepfake is substantially outweighed by the potential for unfair prejudice. This is because social science research shows that jurors may be swayed by audiovisual evidence even when they conclude that it is fake. This argument will be strongest when the alleged deepfake will lead the jury to decide the case based on emotion rather than on the merits.