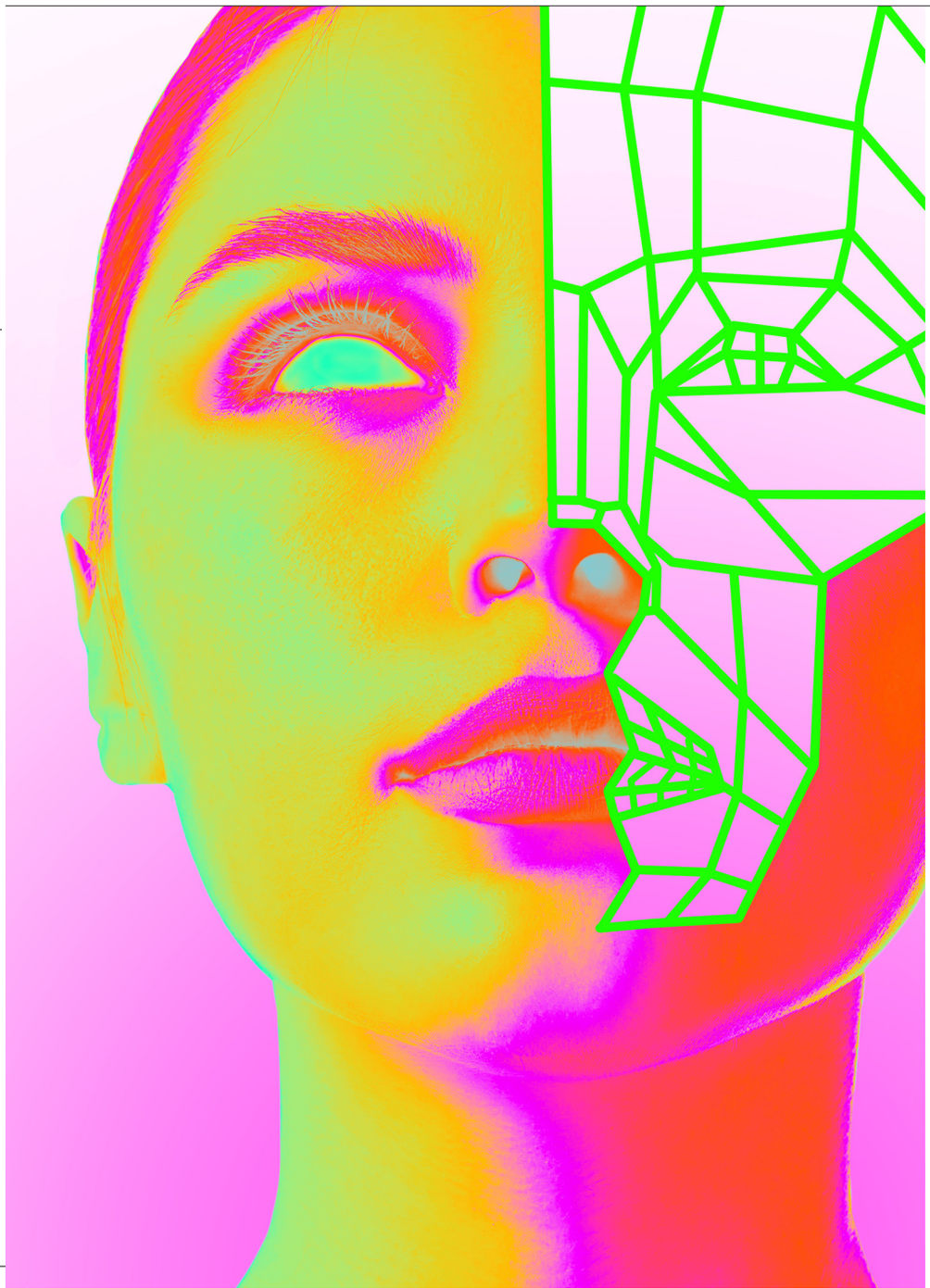
BY TIMOTHY D. EDWARDS

Challenging the Foundations of "Deepfake" Evidence



Most people have seen or heard deepfakes – fake images of human beings, including their speech, that appear real – in the past few years. Deepfakes present special challenges in criminal and civil litigation (beyond the frustration people might feel when they don't know whether they're being fooled). This article provides an analytical framework for the admissibility and exclusion of deepfake evidence, concluding with strategies for lawyers in addressing the most common evidentiary challenges presented by deepfake evidence.

For decades, the laws of evidence have addressed the admissibility of photographic images and audio recordings, including images that are incomplete, altered, or otherwise unreliable due to falsification or errors in the processing stage.

The challenges posed by such evidence have been compounded by computer programs that can create fake images of human beings, including their speech, that appear real to the naked eye or ear. The result is a false or edited visual image or audio recording – a "deepfake" in artificial intelligence (AI) terms – that creates an image no different than a photograph or audio replication that has been altered. To be admitted as evidence in legal proceedings, such images must be authenticated while clearing specific evidentiary hurdles that are designed to ensure the reliability of photographic or audio representations of an actual, relevant event.

This article addresses a foundational question: How does the proponent or opponent of evidence consisting of images or audio recordings alleged to be deepfakes establish or refute the admissibility of the evidence based on the Federal and Wisconsin Rules of Evidence? This article provides an analytical framework for the admissibility and exclusion of deepfake evidence, concluding with strategies for lawyers in addressing the most common evidentiary challenges presented by deepfake evidence.

An Introduction to Deepfakes: Defining the Problem

Typically, deep learning is executed by a type of algorithm called a neural network, which is designed to replicate the way a human brain learns information. "Deepfakes are synthetic media often in the form of videos, audio, or images generated through artificial intelligence (AI) and deep learning algorithms."[1] Deepfake software applications operate by uploading digital images into a machine-learning algorithm that has trained itself to stitch one image on top of another. Deepfakes use two algorithms – a generator and a discriminator – to create and refine fake content.

It is important to understand this process before analyzing the admissibility of deepfake evidence because indicia of unreliability are often hidden in the underlying technical details of the technology. Here, the generator builds a training data set based on the desired output, creating the initial fake digital content, while the discriminator analyzes how realistic or fake the initial version of the content is. The combination of the generator and discriminator algorithms creates a generative adversarial network, which uses deep learning to recognize patterns in real images and then uses those patterns to create the fakes.[2] When creating a deepfake photograph, the algorithm views photographs of the target from an array of angles to capture all the details and perspectives. This information is then run through the discriminator multiple times to fine-tune the realism of the final image or video. The deepfake image is created through this process.

Evidentiary Challenges Presented by Deepfakes

Deepfakes present three distinct challenges in court proceedings: 1) proving whether a digital image or audio evidence is authentic; 2) responding to the "deepfake defense" – the allegation that genuine digital-image or audio evidence is a deepfake; and 3) growing distrust among jurors over the authenticity of all digital-image and audio evidence.[3]

The *Lorraine* Decision. Case law regarding the admissibility of electronically stored information (ESI) provides an important baseline in addressing the admissibility of deepfake evidence. In a leading case, *Lorraine v. Markel American Insurance Co.*, the U.S. District Court for the District of Maryland acknowledged that the authentication of documents from a computer may require greater scrutiny than traditional approaches to the authentication of hard-copy documents. The *Lorraine* court surveyed ESI authentication cases from across the country and concluded that admissibility of ESI is "complicated by the fact that ESI comes in multiple evidentiary 'flavors,' including e-mail, website ESI, internet postings, digital photographs, and computer-generated documents and data files."[4]

Despite this added complexity, courts have adopted a flexible and comprehensive framework to address authenticating ESI based on the same core concepts previously set forth in the rules of evidence, which include identifying the author or creator, identifying the person who received the document, and determining whether the

document is an accurate representation of a person, place, or thing.

Relevance. Relevance is usually the first consideration in determining the admissibility of deepfake technology and other forms of ESI. The Federal Rules of Evidence dictate that only relevant evidence is admissible. Rule 104(b) provides that the preliminary admissibility standard for relevance depends on a fact and states that "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." According to the Advisory Committee, "authentication and identification represent a special aspect of relevancy" because evidence must be authentic for it to be relevant. The special part of relevancy "falls in the category of relevancy dependent upon fulfillment of a condition of fact and is governed by Rule 104(b)." Thus, under the Federal Rules of Evidence, evidence can be deemed relevant and admissible only if it is authentic.[5]

There are two theories of relevance under which video evidence is 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).[6] Under the pictorial-evidence theory, video evidence may be authenticated by any witness who was present when the video was made and perceived the events depicted (that is, the percipient witness).[7] Generally, the percipient witness need only offer testimony that the video evidence fairly and accurately represents the events perceived by the witness.[8] The video might be admissible even if the witness is not aware of who created the video or when the video was created.

Under the silent-witness theory, video evidence is subjected to more scrutiny because there is no percipient witness to testify as to its accuracy. The recorded video becomes the "witness" to the events depicted. An example would be a video submitted from an automatic surveillance camera, which could be authenticated as the accurate product of an automated process under Federal Rule of Evidence 901(b)(9).[9] As a matter of authentication, laying the foundation, and determining relevancy, a court should always listen to audiotapes to determine their admissibility.[10]

Even if the video or photograph is authentic and relevant, the opponent of such evidence may object to its admission on the ground that its probative value is substantially outweighed by the danger of undue prejudice. For example, in *Huddleston v. United States*, the U.S. Supreme Court noted that "[u]nfair prejudice emanates ... first, from the requirement of Rule 404(b) that the evidence be offered for a proper purpose; second, from the relevancy requirement of Rule 402 — as enforced through Rule 104(b); third, from the assessment the trial court must make under Rule 403 to determine whether the probative value of the similar acts evidence is substantially outweighed by its potential for unfair prejudice ...."[11] Thus, even a properly authenticated image can be excluded from evidence if it fails to meet the relevancy standard or is subject to the prohibition against the introduction of unduly prejudicial evidence set forth in Federal Rules of Evidence 401-403 and their Wisconsin counterparts.

Authentication. Federal Rule of Evidence 104(b) mirrors the standard for authentication in Rule 901(a); to satisfy the authentication or identification requirement, "the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is." The authenticity of evidence is ultimately a factual determination for the trier of fact — traditionally a jury — to evaluate. However, before a court admits evidence for the jury to consider, the court must first "determine whether its proponent has offered a satisfactory foundation from which the jury could reasonably find that the evidence is authentic."

"The process by which a judge determines whether the foundation for authentication is proper does not establish the evidence as authentic. The jury is still responsible for the ultimate determination of authenticity and, therefore, credibility; arguments concerning the unreliability of the evidence go to the weight of the evidence and not admissibility. Courts have recognized that the threshold for making the prima facie showing of authenticity to the court is not high, and the burden on the proponent to prove authenticity is slight."[12]

The primary rules that must be considered in addressing the admissibility of evidence challenged as deepfake are Rule 901(a)[13] (which sets forth the requirement that the proponent of nontestimonial evidence must first authenticate or identify it to show that it is what the proponent claims it is), Rule 901(b) (which sets out 10 nonexclusive examples of how evidence can be authenticated), and Rule 902 (which lists ways in which evidence may be authenticated without extrinsic evidence).

The party undertaking to authenticate evidence under Rule 901(a) need only make that showing by a preponderance of the evidence, that is, showing that the evidence is more likely authentic than not. Under Rule 901(b), the proponent has several ways in which to make this threshold showing, including subparts (b)(1), testimony of a witness with personal knowledge; (b)(3), comparison of an example known to

Timothy D. Edwards, Wayne State 1989, is the owner of Edwards ESI LLC, Fitchburg, where he provides e-discovery consulting services and litigates construction, employment, and business disputes. He is a member of the State Bar of Wisconsin's Intellectual Property & Technology Law Section and Labor & Employment Law Section. He is a Fellow of the Wisconsin Law Foundation. Access the digital article at www.wisbar.org/wl.

edwards@tedmadison.com

be authentic with one in which authenticity is challenged; (b)(4), distinctive characteristics of the evidence; and (b)(5), opinion as to voice.

Under Rule 901(b)(9), digital evidence can be authenticated with evidence of a process or system that "produces an accurate result." This authentication method anticipates the presentation of testimony of someone with technical, scientific, or specialized knowledge of the issue to explain why the evidence is valid and reliable, and it should be considered in the context of possible expert testimony, as referenced below. Rule 901(b) is not exhaustive and is intentionally broad; it also offers examples of authenticating specific forms of evidence, including handwriting, a voice, and telephone communication. Rule 902 provides that certain items of evidence are "self-authenticating; they require no extrinsic evidence of authenticity to be admitted."

In 2017, amendments to Federal Rule 902 addressed ESI through the addition of Rule 902(13) and (14), which permit authentication by certification of records generated by an electronic process or system and by data copied from an electronic device, storage medium, or file.[14]

Rule 902(13) allows authentication of a record "generated by an electronic process or system that produces an accurate result" if "shown by the certification of a qualified person" that complies with specific requirements. Rule 902(13) allows ESI to be authenticated without a witness at the stand to state what is supposedly obvious and unlikely to be challenged.

Rule 902(14) allows authentication of "data copied from an electronic device, storage medium, or file, if authenticated by process of digital identification, as shown by a certification of a qualified person." Under Rule 902(14), if proponents of ESI can extract a hash value – a unique numerical identifier that functions like a digital fingerprint – then the evidence is self-authenticating and can be admitted without corroborating witness testimony.

Wisconsin Cases on Admissibility

Several Wisconsin cases address the admissibility of video evidence and social media communications and provide some guidance in the AI deepfake context. For example, in *State v. Baldwin*, a criminal defendant argued that certain text messages were not authentic and thus were inadmissible, despite case law holding that electronic correspondence, including text messages, does not warrant different or more stringent authentication rules than those that are used to authenticate other sorts of correspondence.[15] In *State v. Giacomantonio*, the defendant relied on law in other jurisdictions that "requires more than

mere confirmation that the number or address belonged to a particular person" when authenticating electronic communications.[16] The court rejected this assertion, noting that "[w]e do not need to look further than Wisconsin law, which, as Giacomantonio points out, allows circumstantial evidence for authentication."[17]

In *Giacomantonio*, the state argued that Wis. Stat. sections 909.01 and 909.015 provide the framework for authentication. Section 909.01 provides that "[t]he requirements of authentication or identification as a condition precedent to admissibility are satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims." One way to lay a foundation is through the "[t]estimony of a witness with knowledge that a matter is what it is claimed to be." Wis. Stat. § 909.015(1). Additionally, authentication can be done through circumstantial evidence. Wis. Stat. § 909.015(4) states that examples of authentication include "[a]ppearance, contents, substance, internal patterns, or other distinctive characteristics, taken in conjunction with circumstances."

Based on these principles, the *Giacomantonio* court concluded that authentication was properly established here through circumstantial evidence. "The detective testified that he saw the text messages when the victim's mother brought the phone to him, that he took the screen shots of the messages, and that the screen shots accurately depicted the text messages he viewed. That testimony [alone] … sufficiently authenticated the screen shots as to their accuracy in representing what the detective saw on the phone."[18] Additionally, Giacomantonio did not deny that the phone number was his or claim that those messages did not appear on the victim's phone. Accordingly, the text messages were properly entered into evidence because the defendant "was welcome to cross-examine the victim or any other person regarding whether the text messages had been altered or falsely manufactured."[19]

Traditional Standards for Authentication

Standards for Admission of Photographs and X-Rays. Courts were initially hesitant to admit photographs, videos, or audio recordings into evidence because of concerns as to whether a witness could testify on behalf of the item to be introduced. In an early case, *United States v. Ortiz,* the U.S. Supreme Court allowed the admission of a photograph only after the photographer testified regarding the process used to take a photograph.[20] But the initial hesitancy in admitting photographs relaxed, and courts eventually only required a witness to testify that the photograph was a fair and accurate representation of the contested object or scene.[21]

Standards for Admission of Audio and Video Recordings. Before the Federal Rules of Evidence were enacted in 1975, courts imposed stringent requirements for authenticating audio evidence. In *United States v. McKeever,*[22] defendants sought to admit an audio-recorded conversation between one of the defendants and a witness. The court held that to admit the audio recording into evidence, the proponent of the recording had to demonstrate its "accuracy, authenticity, chain of custody, relevance, and competency."[23] "Interpreting the *McKeever* test as 'a guide rather than a rule,' courts later determined that trial judges should have 'wide latitude' to determine whether a video recording's proponent had laid a sufficient foundation for a reasonable jury to conclude that it was authentic."[24]

A leading case involving the admissibility of allegedly falsified photographs is *State v. Swinton*, in which a criminal defendant challenged the admissibility of photographs of bitemark evidence, some of which were software enhanced and some of which were created with photoshop software.[25] The Connecticut Supreme Court adopted the following six factors for the authentication of evidence generated or enhanced by a computer:

"(1) the computer equipment is accepted in the field as standard and competent and was in good working order, (2) qualified computer operators were employed, (3) proper procedures were followed in connection with the

input and output of information, (4) a reliable software program was utilized, (5) the equipment was programmed and operated correctly, and (6) the exhibit is properly identified as the output in question."[26]

Common-Law and Statutory Standards for Admission of Scientific Evidence

Determination of whether deepfake evidence is authentic and admissible under the Federal Rules also can be informed by the rules courts have applied for admitting scientific, technical, or other specialized information. Such information can be used to authenticate or challenge the authenticity of deepfake evidence.

For many decades, federal courts applied the standard established in *Frye v. United States*[27] to determine the admissibility of novel scientific evidence that might assist a jury in its deliberations. In 1993, the Supreme Court held in *Daubert v. Merrell Dow Pharmaceuticals Inc.* that the *Frye* standard of admissibility was incompatible with the Federal Rules of Evidence, which approached novel scientific evidence more broadly.[28] Federal Rule of Evidence 702 now requires that the introduction of evidence dealing with scientific, technical, or specialized knowledge beyond the understanding of lay jurors be based on sufficient facts or data and reliable methodology applied to the facts of the case. As described in the Advisory Committee Note to the amendment of Rule 702 that went into effect in 2000, the "*Daubert* factors" are the following:

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

Expert testimony can be used to admit or exclude deepfake evidence if the testimony satisfies the *Daubert* standard. Typically, the expert will testify regarding the use of a program or other method that reliably creates or detects deepfake evidence. If the method used is proved to be reliable based on these factors, the court can rely on the expert's testimony to evaluate the authenticity of the image or audio.[29]

Going Forward

As deepfake technology has developed, many commentators have questioned whether the existing authenticity rules are flexible enough to address any problems arising from deepfakes, including the possibility of juror bias or confusion. Some scholars see no need for a higher standard of proof at the admissibility level. "A trial judge should admit the evidence if there is plausible evidence of authenticity produced by the proponent of the evidence and only speculation or conjecture – not facts – by the opponent of the evidence about how, or by whom, it 'might' have been created."[30] They do believe, however, that the difficulty in determining the authenticity of deepfakes justifies some procedural structure and protection at an admissibility hearing. They propose an amendment to Rule 901(b)(9) that would provide as follows:

(9) *Evidence About a Process or System*. For an item generated by a process or system:

(A) evidence describing it and showing that it produces a reliable result; and

(B) if the proponent concedes that – or the opponent provides a factual basis for suspecting that – the item was generated by artificial intelligence, additional evidence that:

(i) describes the software or program that was used; and

(ii) shows that it produced reliable results in this instance.[31]

This proposal provides a helpful way to structure an authenticity question when deepfake technology is involved. It imposes no safeguards in the first instance when a proponent seeks to admit an audiovisual item, meaning that the mere allegation by the opponent that something is "deepfake" would be treated as a nonevent. However, if the opponent were to provide a factual basis for believing that there is a deepfake or if the proponent were to concede that AI has been used, the proponent would have to describe how the item was prepared and show that it is a reliable account of what it portrays.[32] **WL**

ENDNOTES

[1]Neill Jacobson, *Deepfakes and Their Impact on Society*, CPI OpenFox (Feb. 26, 2024), https://www.openfox.com/deepfakes-and-their-impact-on-society/. "Deepfake" is a combination of "deep" and "fake," referring to the fact that it uses "deep learning," a subset of machine learning that relies on artificial neural networks and is manipulated content. *See* Graham Meikle, *Deepfakes* 2 (2022). Deepfake videos are one of the most well-known forms of synthetic media. *Summary of Discussions and Next Step Recommendations from "Mal-uses of AIgenerated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening"* (Brooklyn: Witness, 2018).

[2]*See, e.g.*, Akamai, *What Is a Generative Adversarial Network (GAN)?*, https://www.akamai.com/glossary/what-is-a-generative-adversarial-network-gan (last visited Nov 7, 2025).

[3]Rebecca 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, 308-09 (Feb. 2023), https://repository.uclawsf.edu/cgi/viewcontent.cgi?article=4012&context=hastings_law_journal.

[4]*Lorraine v. Markel Am. Ins. Co.*, 241 F.R.D. 534 (D. Mary. 2007).

[5]Delfino, *supra* note 3, at 322.

[6]*See* 2 Kenneth S. Broun, *McCormick on Evidence* § 215 (7th ed. 2013) (describing the different ways that photographs are used in courts).

[7]*Id.*

[8]*Id.*

[9]*Id.*

[10]*State v. Nipple,* No. 98-1615-CR, 1999 WL 161063 (Wis. Ct. App. March 25, 1999) (unpublished).

[11]*Huddleston v. United States*, 485 U.S. 681, 691 (1988).

[12]Delfino, *supra* note 3, at 322.

[13]This article refers to the Federal Rules of Evidence and the Wisconsin Rules of Evidence as the "Rules" or, if singular, as a "Rule." To the extent that the Rules differ in each jurisdiction, they are noted accordingly.

[14]Wisconsin does not mirror these provisions of the Federal Rules.

[15]*See State v. Baldwin*, 2010 WI App 162, ¶ 55, 330 Wis. 2d 500, 794 N.W.2d 769.

[16]*State v. Giacomantonio*, 2016 WI App 62, 371 Wis. 2d 452, 885 N.W.2d 394 (citing *State v. Thompson*, 777 N.W.2d 617, 624-25 (N.D. 2010); *Commonwealth v. Koch*, 39 A.3d 996, 1004 (Pa. Super. Ct. 2011), *aff'd by equally divided court*, 106 A.3d 705 (Pa. 2014)).

[17]*Id.* ¶ 19 (citing *Baldwin*, 2010 WI App 162, ¶ 55, 330 Wis. 2d 500).

[18] *Id.* ¶ 21.

[19]*Id.* ¶ 24.

[20]*United States v. Ortiz*, 176 U.S. 422 (1900).

[21]*See, e.g.*, *United States v. Stearns*, 550 F.2d 1167, 1171 (9th Cir. 1977): "Even if direct testimony as to foundation matters is absent, however, the contents of a photograph itself, together with such other circumstantial or indirect evidence as bears upon the issue, may serve to explain and authenticate a photograph sufficiently to justify its admission into evidence."

[22]*United States v. McKeever*, 169 F. Supp. 426 (S.D.N.Y. 1958), *rev'd on other grounds*, 271 F.2d 669 (2d Cir. 1959).

[23]"Proper authentication of digital videos or photographs may require detailed evidence about the chain of custody, such as how digital content was retrieved from a defendant's computer and subsequently stored." Delfino, *supra* note 3, at 326.

[24]*Id.* at 315-16.

[25]*Id.* at 318; *State v. Swinton*, 847 A.2d 921 (Conn. 2004).

[26]*Swinton*, 847 A.2d at 942; *see* Catherine Guthrie & Brittan Mitchell, *The Swinton Six: The Impact of* State v. Swinton *on the Authentication of Digital Images*, 36 Stetson L. Rev. 661 (2007).

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

[28]*Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

[29]*See* Delfino, *supra* note 3, at 319-20, 339.

[30]*See, e.g.,* Daniel J. Capra, *Deepfakes Reach the Advisory Committee on Evidence Rules,* 92 Fordham L. Rev. 2491, 2502-03 n.83 (2024), https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=6094&context=flr (citing Paul W. Grimm, Lisa Yurwit Bergstrom & Melissa M. O'Toole-Loureiro, *Authentication of Social Media Evidence*, 36 Am. J. Trial Advoc. 433, 459 (2013)).

[31]Dr. Maura R. Grossman and the Hon. Paul W. Grimm (retired) submitted a proposed modification of Rule 901(b)(9) to address authentication issues regarding AI evidence to the Advisory Committee on Evidence Rules in advance of their presentation during the autumn 2023 meeting.

[32]There are methods for combatting deepfakes. Under the affidavit of forensic analysis (AFA) approach, proponents of video evidence are required to submit with their proffered video evidence an AFA that will be used to assist the trial judge in performing the gatekeeping function under Rule 104(b). Separate from detection methods, deepfakes can also be combatted by technology ensuring the authenticity of real media. **WL**