Deepfakes on Trial: A Call To Expand the Trial Judge's Gatekeeping Role To Protect Legal Proceedings from Technological Fakery

REBECCA A. DELFINO[†]

Deepfakes—audiovisual recordings created using artificial intelligence (AI) technology to believably map one person's movements and words onto another—are ubiquitous. They have permeated societal and civic spaces from entertainment, news, and social media to politics. And now deepfakes are invading the courts, threatening our justice system's truth-seeking function. Ways deepfakes could infect a court proceeding run the gamut and include parties fabricating evidence to win a civil action, government actors wrongfully securing criminal convictions, and lawyers purposely exploiting a lay jury's suspicions about evidence. As deepfake technology improves and it becomes harder to tell what is real, juries may start questioning the authenticity of properly admitted evidence, which in turn may have a corrosive effect on the justice system.

No evidentiary procedure explicitly governs the presentation of deepfake evidence in court. The existing legal standards governing the authentication of evidence are inadequate because they were developed before the advent of deepfake technology. As a result, they do not solve the urgent problem of how to determine when an audiovisual image is fake and when it is not. Although legal scholarship and the popular media have addressed certain facets of deepfakes in the last several years, there has been no commentary on the procedural aspects of deepfake evidence in court. Absent from the discussion is who gets to decide whether a deepfake is authentic. This Article addresses the matters that prior academic scholarship on deepfakes obscures. It is the first to propose a new addition to the Federal Rules of Evidence reflecting a novel reallocation of fact-determining responsibilities from the jury to the judge, treating the question of deepfake authenticity as one for the court to decide as an expanded gatekeeping function under the Rules. The challenges of deepfakes—problems of proof, the "deepfake defense," and juror skepticism—can be best addressed by amending the Rules for authenticating digital audiovisual evidence, instructing the jury on its use of that evidence, and limiting counsel's efforts to exploit the existence of deepfakes.

[†] Clinical Professor of Law, Loyola Law School. I would like to thank Ali Paine, Cameron Silbar, Kenneth Park, Ahnna Chu, and Ani Boyadjian for their research assistance; Loyola Law School for the support of faculty scholarship; and as always, my family for their patience, indulgence, and encouragement. This Article is dedicated to Edward J. Imwinkelried, the Edward L. Barrett, Jr. Professor of Law Emeritus at UC Davis Law School, who graciously shared his perspectives on my proposal. Professor Imwinkelried is as patient and gifted a teacher of evidence now as he was three decades ago when I was his student in law school.

TABLE OF CONTENTS

The question is not what you look at, but what you see.

— Henry David Thoreau[1]

INTRODUCTION

A surveillance video shows a robbery in progress—it clearly shows a person armed with a gun walk into a store, rob the store clerk, and then escape. The video is sharp, and the audio is clear—it is *your* voice and face caught on the video. To the naked eye, it appears that you are the robber. After you are arrested and confronted with the video, you protest that although it looks and sounds like you, the person shown is not you—that the video is fake. You are the victim of a deepfake. But how can your lawyer prove it? And how should the court handle the deepfake? And who—the judge or the jury—gets to decide whether the evidence is real or fake?

A portmanteau of "deep learning" and "fake," so-called "deepfake" programs use artificial intelligence (AI) to produce fake videos of people that appear genuine.[2] This new technology, developed and unleashed on the internet beginning in late 2017, now allows anyone with a smartphone to believably map another's movements and words onto someone else's face and voice to make them appear to say or do anything.[3] And the more video and audio of the person is fed into the computer's deep-learning algorithms, the more convincing the result. Deepfakes pose dangers and risks to our society and democratic institutions, including our judicial system, through their connection to fake news and false images depicting public figures, government officials, and private individuals.[4] Legal literature and academic scholarship are just beginning to examine the social and legal ramifications of deepfakes,[5] and deepfake evidence in court proceedings is a new and emerging phenomenon.[6]

This Article explores the advent of deepfakes by focusing on their effect in court. Ways that deepfakes could infect a court proceeding run the gamut and include parties fabricating evidence to win a civil action, government actors

[1] 2 HENRY DAVID THOREAU, THE JOURNAL OF HENRY DAVID THOREAU: 1850 – SEPTEMBER 1851, at 373 (Bradford Terry & Francis H. Allen eds., 2016).

[2] Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 888, 889, 892–93, 929 (2019).

[3] *Id.* at 889–94; *see* Russell Spivak, *"Deepfakes": The Newest Way To Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 339 (2019) (referencing deepfake technology's pornographic beginnings).

[4] *See generally* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1776–85 (2019) (identifying a litany of potential exploitative uses for deepfake technology to cause individual harm (including sabotage, blackmail, and exploitation) and to cause societal harm (including harm to democratic institutions, civil discourse, public safety, and national security)); *see also* David Lee, *Deepfakes Porn Has Serious Consequences*, BBC NEWS (Feb. 3, 2018), https://www.bbc.com/news/technology-42912529 [https://perma.cc/R4TY-LJQK] (claiming that deepfake technology "could down the road be used maliciously to hoax governments and populations, or cause international conflict").

[5] *See, e.g.*, Chesney & Citron, *supra* note 4; Delfino, *supra* note 2, at 889–94; Spivak, *supra* note 3, at 339.

[6] *See infra* Part I.D.

wrongfully securing criminal convictions, and lawyers purposely exploiting a lay jury's suspicions about evidence.

Deepfakes will soon make trial attorneys' and judges' jobs significantly more challenging. They will require courts to take additional measures to determine the authenticity of images before admitting them into evidence. These images will also complicate ordinary trial proceedings. They will require a reevaluation of how to determine authentication, the role of the jury in that process, and how to instruct the jury on handling the evidence. As deepfake technology improves and it becomes harder to tell what is real, juries may start questioning the authenticity of properly admitted evidence, which in turn may have a corrosive effect on the justice system.

No federal evidentiary procedure explicitly governs the presentation of these images in court. The existing legal standards governing the authentication of evidence, despite providing some guidance, fall short, because the Federal Rules of Evidence were developed before the advent of deepfake technology. The current Rules will need to be adapted to solve the problem of how to show when a video is fake and when it is not.

In the last four years, academic scholarship, legal literature, and popular media have generated a mountain of commentary on various aspects of deepfakes, including a handful of articles exploring the challenges of deepfake evidence in legal proceedings.[7] However, there has been no commentary on the procedural facets of deepfake evidence in court; absent is a discussion of who, the judge or the jury, gets to decide whether a deepfake is authentic. This Article is the first to address these issues by proposing the solution of a new Rule of Evidence reallocating fact-determining responsibilities from the jury to the judge on the question of authenticity.

Part I defines and explains the rise of deepfakes. It considers the civil and criminal remedies offered thus far to contain deepfakes and recognizes that, despite efforts to regulate them, deepfakes have begun to invade legal proceedings either as the central focus of litigation or as an item of evidence to prove another claim. Part I also identifies and explores the three separate challenges that deepfakes pose to legal proceedings: (1) proving whether audiovisual evidence is genuine or fake, (2) responding to claims that genuine evidence is a deepfake, and (3) addressing growing distrust and doubt among jurors over the authenticity of audiovisual evidence.

Part II explores the complexity of the legal and prudential issues implicated by deepfakes in legal proceedings. It begins with a discussion of how courts have historically dealt with the admission of new kinds of evidence, and what can be learned from that history. This Part also examines the Rules governing authenticity and the common-law theories applied to determine the admissibility of audiovisual evidence. Part II then identifies the shortcomings and limitations

[7] *See infra* Part I.D.

of the evidentiary mechanisms currently available to litigants and courts. It argues that none of the current mechanisms can fully address the challenges posed by deepfakes in the courtroom, including problems with proof, corruption of the trial process, and heightened juror bias and skepticism. These shortcomings demonstrate how the current approaches must be revised.

Finally, Part III offers solutions to challenges surrounding the presentation of deepfakes in legal proceedings under the current Federal Rules of Evidence. This Article argues that addressing these challenges requires considering the existing, non-exhaustive means of authentication set forth by Federal Rule of Evidence 901(b). Trial courts should be afforded the flexibility to rely on multiple means to determine authenticity based on a combination of sources, including percipient witness authentication, as well as digital forensics evidence and expertise. Additionally, the Rules must be amended to involve a novel redistribution of the authority to determine authenticity from the jury to the court, so that juror skepticism and bias do not corrupt the judicial system's truth-determining process.[8]

Although courts and lawyers have faced authentication challenges before in dealing with photographs, x-rays, DNA, and audio, visual, digital, and social media evidence, deepfakes present something exceptional, complex, and urgent. The challenges deepfakes pose in legal proceedings demand that courts and lawyers creatively navigate pitfalls of proof and manage jurors' doubts and distrust. This Article explores these issues and offers a concrete solution to guide the way forward for lawyers and courts as they traverse this new technological landscape.

I. Deepfakes in Legal Proceedings: Scope and Challenges

A. Deepfakes, Defined

Deepfakes are fabricated audiovisual content created or altered to appear to a reasonable observer to be a genuine account of the speech, conduct, image, or likeness of an individual or event.[9] They create a fake reality by superimposing a person's face on another's body, or by changing the content of one's speech.[10] The term "deepfake" is derived from a combination of "deep

[8] In addition to the Federal Rules of Evidence, the challenges posed by deepfakes affect the procedural aspects of legal proceedings, which warrant a reexamination of the Federal Rules of Civil and Criminal Procedure. The problem of deepfake evidence in legal proceedings also requires consideration of the conduct of the lawyers who exploit suspicions about the authenticity of evidence even when they have reason to believe it is genuine. However, deepfakes' implications for the Rules of Civil and Criminal Procedure and legal ethics are beyond the scope of this Article.

[9] Delfino, *supra* note 2, at 889, 892–93, 929.

[10] *Id.* at 889–94; Spivak, *supra* note 3, at 339.

learning" and "fake."[11] "Deep learning" refers to the training process by which AI technology becomes increasingly intelligent through the continued introduction of information into the system.[12] Deepfake software applications operate by uploading digital images into a "machine-learning algorithm that's trained itself to stitch one face on top of another."[13]

The first deepfakes were generated based on a single neural network system, but with subsequent technological advances, the fabricated imagery is now made using generative adversarial networks,[14] a two-part AI system that generates altered audio or video that is then compared to the real content the technology is trying to mimic.[15] The two algorithms compete against each other to improve each system; the discriminator (the authenticator of the video) improves itself by spotting fakes, and the generator (the system generating the fake content) improves from the feedback that the discriminator provides to produce a more realistic fake version of the content.[16] This adversarial process aims to generate images so convincing that the discriminator believes they belong with the "real" dataset.[17] The outcome is a convincing deepfake that is impossible to distinguish from reality with the naked eye.[18]

B. THE RISE OF DEEPFAKES

Deepfakes offer a kind of self-improving technology that is readily accessible, increasingly inexpensive, and exceedingly difficult to detect.[19] Deepfakes first surfaced on the internet in 2017 when an anonymous Reddit user applied the technology to create realistic pornographic videos featuring famous female celebrities.[20] Following the release of these face-swapped porn videos, another anonymous Reddit user created and released FakeApp, a free application

[11] Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99, 99–100 (2019) (quoting Sundar Pichai of Google, whose company invented TensorFlow to develop AI tools for the public).

[12] *Id*.

[13] Delfino, *supra* note 2, at 893.

[14] *See* Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 B.U. PUB. INT. L.J. 245, 249 (2020) (referencing the difference between deepfakes created from a single neural network and those created from a generative adversarial network).

[15] *See* Kyle Wiggers, *Generative Adversarial Networks: What GANs Are and How They've Evolved*, VENTUREBEAT (Dec. 26, 2019, 1:45 PM), https://venturebeat.com/2019/12/26/gan-generative-adversarial-network-explainer-ai-machine-learning/ (explaining the "architecture" behind Generative Adversarial Networks (GANs)).

[16] *See* Pfefferkorn, *supra* note 14.

[17] *Id.* (referencing the purpose of GANs).

[18] David Dorfman, *Decoding Deepfakes: How Do Lawyers Adapt When Seeing Isn't Always Believing?*, 80 OR. ST. BAR. BULL. 18, 20 (explaining the threat that deepfakes pose to society).

[19] *See* Pfefferkorn, *supra* note 14, at 247. Early commentary reports that even technology experts have struggled to distinguish genuine videos from deepfakes. *See* Drew Harwell, *Top AI Researchers Race To Detect 'Deepfake' Videos: 'We Are Outgunned*,*'* WASH. POST (June 12, 2019, 4:44 PM), https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/ [https://perma.cc/2V2V-SVMK].

[20] *See* Delfino, *supra* note 2, at 893 (explaining the development of deepfake technology).

enabling users to create deepfakes with ease.[21] Before FakeApp's development, the production of realistic doctored videos was an expensive and technically complicated process confined to Hollywood movie studios.[22] FakeApp's creator achieved the goal of "mak[ing] deepfake[] technology available to people without a technical background or programming experience."[23]

The advent and swift spread of user-friendly applications such as FakeApp have unleashed a flood of benign, entertaining celebrity deepfakes. For example, using this technology in a television interview, comedian Bill Hader appeared to shapeshift to take on the faces of famous actors from Arnold Schwarzenegger to Al Pacino.[24] Actor Steve Buscemi's face is imposed on Jennifer Lawrence's body in a deepfake video clip in which the actress discusses her favorite "Real Housewife."[25] Former President Barack Obama appears to give an expletive-laced speech about the threat of misinformation to democracy ending with the plea to "stay woke, bitches."[26] Viral, doctored videos show Elon Musk's face superimposed on babies.[27] And a search for Nicholas Cage on Reddit produces numerous videos of his face added to clips of famous films in which he never appeared, such as *The Sound of Music* and *Raiders of the Lost Ark*.[28]

Although some deepfakes are harmless and amusing artistic expressions, more than ninety percent of deepfakes on the internet are pornographic

[21] Derek Hawkins, *Reddit Bans 'Deepfakes,' Pornography Using the Faces of Celebrities Such as Taylor Swift and Gal Gadot*, WASH. POST (Feb. 8, 2018), https://www.washingtonpost.com/news/morning-mix/wp/2018/02/08/reddit-bans-deepfakes-pornography-using-the-faces-of-celebrities-like-taylor-swift-and-gal-gadot/ [https://perma.cc/2TPE-AYBG].

[22] *See* Kevin Roose, *Here Come the Fake Videos, Too*, N.Y. TIMES (Mar. 4, 2018), https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html [https://perma.cc/6QA8-78NL]; *see also* Hawkins, *supra* note 21 (explaining that FakeApp "put deepfake technology into a user-friendly package").

[23] *See* Samantha Cole, *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*, VICE (Jan. 24, 2018, 10:13 AM), https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley (describing how, in late fall 2017, an anonymous Reddit user posted several porn videos on the internet under the pseudonym "Deepfakes," including a video of actress Daisy Ridley's face superimposed on the body of a porn actress).

[24] Jon Blistein, *Watch Bill Hader Become Tom Cruise, Seth Rogen in Eerie Deepfake Video*, ROLLING STONE (Aug. 13, 2019), https://www.rollingstone.com/culture/culture-news/bill-hader-tom-cruise-seth-rogen-deepfake-871154/ [https://perma.cc/5L99-PXD8].

[25] *See* Birbfakes, *Jennifer Lawrence-Buscemi on Her Favorite Housewives [Deepfake]*, YOUTUBE (Jan. 14, 2019), https://www.youtube.com/watch?v=r1jng79a5xc [https://perma.cc/VDT2-L54N].

[26] Todd Spangler, *Jordan Peele Teams with BuzzFeed for Obama Fake-News Awareness Video (Watch)*, VARIETY (Apr. 17, 2018, 8:45 AM), https://variety.com/2018/digital/news/jordan-peele-obama-fake-news-video-buzzfeed-1202755517/ [https://perma.cc/S8G8-WRR4]. The video was created by BuzzFeed and comedian Jordan Peele, who did the voice impersonation used in the video. *Id.* University of Washington researchers have developed an AI tool that allows them to easily manipulate a video of Barack Obama to swap out the speech he is giving, producing a realistic deepfake video. Adam Mann, *Deepfake AI: Our Dystopian Present*, LIVE SCI. (Sept. 30, 2019), https://www.livescience.com/deepfake-ai.html [https://perma.cc/HU3C-XP7B].

[27] *See* Amanda Kooser, *This Elon Musk Deepfake Baby Video Shattered My Brain*, CNET (May 10, 2019, 9:23 AM), https://www.cnet.com/news/this-elon-musk-deepfake-baby-video-shat-tered-my-brain/ [https://perma.cc/QP33-PWM7].

[28] *See* Sam Haysom, *People Are Using Face-Swapping Tech To Add Nicholas Cage to Random Movies and What Is 2018*, MASHABLE (Jan. 31, 2018), https://mashable.com/2018/01/31/nicolas-cage-face-swapping-deepfakes/.

depictions of women.[29] Female celebrity faces have been digitally added to pornographic content, creating deepfake porn videos.[30] Scarlett Johansson,[31] Meghan Markle,[32] and Taylor Swift[33] have all been victims of deepfake pornography. But the deepfake pornography phenomenon is not limited to celebrities; even private citizens have been victimized.[34] Victims of nonconsensual pornographic deepfakes may endure much harm, including emotional trauma, stigmatization, reputational harm, harassment, and even blackmail.[35] And as the technology advances, the deep-learning AI software will require fewer and fewer images to create a believable deepfake. As a result, some people with only a handful of images on the internet may find themselves the star of a deepfake video, thereby increasing the number of potential deepfake victims.[36]

Deepfakes have also proliferated beyond the confines of entertainment and pornography into the political sphere, creating fake news and false images involving political leaders and government actors.[37] AI-assisted technology can be used to create fake videos of politicians accepting bribes, soldiers committing war crimes, presidential candidates engaging in criminal behavior, and emergency officials announcing an impending terrorist attack.[38] All of these examples have the potential to harm our democracy, particularly because the

[29] HENRY AJDER, GIORGIO PATRINI, FRANCESCO CAVALI & LAURENCE CULLEN, THE STATE OF DEEPFAKES: LANDSCAPE, THREATS, AND IMPACT 2 (2019), https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[30] Cleo Abram, *The Most Urgent Threat of Deepfakes Isn't Politics. It's Porn.*, VOX (June 8, 2020, 12:10 PM), https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell.

[31] Isobel Asher Hamilton, *Scarlett Johansson Says Trying To Stop People Making Deepfake Porn Videos of Her Is a 'Lost Cause,'* BUS. INSIDER (Dec. 31, 2018, 2:51 AM), https://www.businessinsider.com/scarlett-johansson-stopping-deepfake-porn-of-me-is-a-lost-cause-2018-12 [https://perma.cc/HE4M-S92V].

[32] Ian Morris, *Deepfake Porn Banned by Reddit and Pornhub After Taylor Swift and Meghan Markle Clips Emerge Online*, FORBES (Feb. 7, 2018, 4:42 PM), https://www.forbes.com/sites/ianmorris/2018/02/07/deepfake-porn-banned-by-reddit-and-pornhub-after-taylor-swift-and-meghan-markle-clips-emerge-online/#5a32524a48ea [https://perma.cc/APX7-7CR3].

[33] *Id.*

[34] *See, e.g.*, Daniella Scott, *Deepfake Porn Nearly Ruined My Life*, ELLE (Feb. 6, 2020), https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/ [https://perma.cc/4D27-2JDL] (describing an Australian law graduate who discovered that her public social media images were used to create explicit photos and videos of her).

[35] Danielle Keats Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1891–92, 1915, 1924–28 (2019).

[36] Gregory Barber, *Deepfakes Are Getting Better, but They're Still Easy To Spot*, WIRED (May 26, 2019, 7:00 AM), https://www.wired.com/story/deepfakes-getting-better-theyre-easy-spot/?utm_source=onsite-share&utm_medium=email&utm_campaign=onsite-share&utm_brand=wired.

[37] Chesney & Citron, *supra* note 4, at 1769–77; *see* Lee, *supra* note 4 (claiming that deepfake technology "could down the road be used maliciously to hoax governments and populations, or cause international conflict").

[38] *See* Chesney & Citron, *supra* note 4, at 1776–85; *see also* John Donovan, *Deepfake Videos Are Getting Scary Good*, HOW STUFF WORKS, https://electronics.howstuffworks.com/future-tech/deepfake-videos-scary-good.htm (Mar. 2, 2021).

technology used to generate these videos is rapidly advancing.[39] For instance, in 2019, millions of people viewed a video with altered audio showing Speaker of the House Nancy Pelosi slurring her speech in a recorded interview.[40] The video was widely circulated on social media, including in a tweet by former President Donald Trump.[41] In Malaysia, a politician was mired in a sex tape scandal after a deepfake video surfaced purporting to show him engaging in illegal homosexual activity.[42] Experts have warned that, if unchecked, deepfakes can undermine democracy by amplifying falsehoods and sowing discord.[43]

C. The Legal Responses to Deepfakes

As deepfakes have exploded beyond the confines of entertainment and pornography into the political realm, policymakers have shifted into containment-and-response mode, attempting to prevent, mitigate, and punish abusing deepfake technology for harmful purposes.[44] Scholars likewise have explored existing civil and criminal sanctions that could redress those harms, and have proposed new laws and legal frameworks to constrain bad actors.[45]

[39] *See Rise of the Deepfakes*, WEEK (June 9, 2018), http://theweek.com/articles/777592/rise- deepfakes [https://perma.cc/2P9A-DRWF] (noting that "deep learning" technology used to power deepfakes is improving fast).

[40] Russell Berman, *For Nancy Pelosi, This Is All Just Déjà Vu*, THE ATLANTIC (May 24, 2019), https://www.theatlantic.com/politics/archive/2019/05/trump-pelosi-video/590233/ [https://perma.cc/C2LR-KES9]. The video of Nancy Pelosi has been characterized as a "cheap" or "shallow fake" because it did not rely on deep-learning technology; instead, it involved slowing down the speed by seventy-five percent to make it appear that Pelosi was slurring her words and by manipulating the audio so that the pitch of the voice remained realistic. *See* Jason Abbruzzese, *Doctored Pelosi Videos Offer a Warning: The Internet Isn't Ready for 2020*, NBC NEWS (May 24, 2019, 10:56 AM), https://www.nbcnews.com/tech/tech-news/doctored-pelosi-videos-offer-warning-internet-isn-t-ready-2020-n1010011 [https://perma.cc/WU4Y-HYNV].

[41] Abbruzzese, *supra* note 40. President Trump's tweet sharing the video received 30,000 retweets and 90,000 likes. *Id. See* Lauren Feiner, *Facebook Says the Doctored Nancy Pelosi Video Used To Question Her Mental State and Viewed Millions of Times Will Stay Up*, CNBC (May 25, 2019, 9:40 AM), https://www.cnbc.com/2019/05/24/fake-nancy-pelosi-video-remains-on-facebook-and-twitter.html [https://perma.cc/ZGZ2-G6S4] (emphasizing the impact of the deepfake video of Nancy Pelosi on public perception). Rudy Giuliani, Donald Trump's attorney at the time, shared the false video and stated: "What is wrong with Nancy Pelosi? Her speech pattern is bizarre." *Id.*

[42] Jarni Blakkarly, *A Gay Sex Tape Is Threatening To End the Political Careers of Two Men in Malaysia*, SBS NEWS (June 17, 2019, 6:28 PM), https://www.sbs.com.au/news/a-gay-sex-tape-is-threatening-to-end-the-political-careers-of-two-men-in-malaysia [https://perma.cc/KN2J-ZAX2]. The video scandal caused significant political fallout. A. Ananthalakshmi, *Malaysian Police Say Political Leader Behind Gay Sex Tape Allegations*, REUTERS (July 17, 2019, 11:54 PM), https://www.reuters.com/article/us-malaysia-politics/%20alaysian-police-say-political-leader-behind-gay-sex-tape-allegations-idUSKCN1UD0OF [https://perma.cc/3RW6-G9GB] (noting that the video appears to be authentic, but that facial recognition could not confirm the identity of all parties).

[43] *See* Chesney & Citron, *supra* note 4, at 1769, 1777 (describing the scope of domestic, social, economic, and foreign policy implications of deepfakes).

[44] *See, e.g.*, Andrew Grotto, *Andrew Grotto Testimony on 'Cybersecurity and California Elections,'* STAN. UNIV. (Mar. 7, 2018), https://cisac.fsi.stanford.edu/docs/andrew-grotto-testimony-cybersecurity-and-california-elections; Kaveh Waddell, *Lawmakers Plunge into 'Deepfake' War*, AXIOS (Jan. 31, 2019), https://www.axios.com/deepfake-laws-fb5de200-1bfe-4aaf-9c93-19c0ba16d744.html (reporting that federal legislators have begun "invit[ing] legal scholars to privately brief their staff on deepfakes").

[45] *See* Chesney & Citron, *supra* note 4, at 1777; Delfino, *supra* note 2, 903–25.

However, the legal responses have lagged behind the technology.[46] Few federal and state laws target deepfakes. The first federal law to address deepfakes was enacted in late 2019. The National Defense Authorization Act of 2020[47] requires the U.S. Intelligence Community to conduct annual assessments regarding the foreign weaponization of deepfakes, particularly by China and Russia, and to notify Congress whenever a foreign power deploys a deepfake to interfere with an American election.[48] The Act also authorized the creation of a competition to encourage research on deepfakes.[49]

Other federal legislation regarding deepfakes has been proposed since 2018, but none has passed out of committee.[50] The most significant was H.R. 3230, the Deepfakes Accountability Act of 2019, which would have mandated that most classes of deepfakes contain digital watermarks and prominent written or audio statements disclosing the extent of the alterations.[51] The proposed Act would also have provided criminal penalties and a civil right to enforce these requirements. In addition, it would have updated false personation laws to encompass digital impersonation, created an in rem litigation procedure to sue unidentifiable deepfake creators (or open up the discovery process to aid in their identification), and established a national security task force to develop detection tools and share them with online platforms.[52] H.R. 3230, like the other federal

[46] Delfino, *supra* note 2, at 903–04.

[47] *See* Matthew Ferraro, Jason C. Chipman & Stephen W. Preston, *First Federal Legislation on Deepfakes Signed into Law*, WILMERHALE (Dec. 23, 2019), https://www.wilmerhale.com/en/insights/client-alerts/2019 1223-first-federal-legislation-on-deepfakes-signed-into-law [https://perma.cc/C9CV-FQFG] (announcing that Donald Trump signed the first law concerning deepfakes on December 20, 2019).

[48] *See id.* (summarizing the report delivered to Congress in 2020 on deepfakes by the Director of National Intelligence). The report to Congress included: (1) the technological capabilities of foreign countries to create deepfakes, (2) how disinformation from foreign governments could harm the United States' elections, (3) what technology the United States can develop to combat deepfake attacks, (4) current deepfake capabilities of the United States, (5) an explanation of what is currently being done regarding deepfakes in the United States, and (6) recommendations for additional needs to combat deepfakes. *Id.*; *see* National Defense Authorization Act for Fiscal Year 2020, S. 1790, 116th Cong. (2019) (enacted) (authorizing the appropriations and policies for the Department of Defense).

[49] S. 1790.

[50] *See* Delfino, *supra* note 2, at 918–25 (summarizing Congress' initial efforts to legislate deepfakes); *see also* Deep Fakes Accountability Act, H.R. 3230, 116th Cong. (2019) (providing guidelines for regulating and marking tampered videos); Deepfake Report Act of 2019, S. 2065, 116th Cong. (2019) (requiring the Department of Homeland Security to report on the state of "digital content forgery technology" during specified periods); Identifying Outputs of Generative Adversarial Networks Act, H.R. 4355, 116th Cong. (2019) (asking for federal support in manipulated media research).

[51] *See* H.R. 3230 (naming proposed legislation by Congress that protects the public from disinformation spread through deepfakes); Dorfman, *supra* note 18, at 21 (asserting that the Deepfakes Accountability Act is "[t]he most significant bill in Congress"); *see also* Chesney & Citron, *supra* note 4, at 1758 (highlighting the issues of subject consent that arise with the creation of deepfakes). "Although deep fakes can be created with the consent of people being featured, more often they will be created without it." Chesney & Citron, *supra* note 4.

[52] *See* H.R. 3230, § 1041 (a)–(e) (specifying ratifications needed for altered videos under Texas law). The new regulations include providing a digital watermark on the altered image, as well as an audiovisual, visual, or audio disclosure. *Id*; *see also* Daniel Lipkowitz, *Manipulated Realty, Menaced Democracy: An Assessment of the DEEP FAKES Accountability Act of 2019*, 2020 N.Y.U. J. LEGIS. & PUB. POL'Y QUORUM 30, 31 (critiquing the reforms outlined in the Deep Fakes Accountability Act). Watermarks can be easily removed, and it is difficult

efforts to legislate deepfakes, did not receive the full endorsement of Congress and expired when the congressional session closed at the end of 2020.[53]

States have also been slow to enact deepfake legislation. Only a small handful of states have realized the imminent threat of deepfakes and taken action. In 2019, Virginia amended its revenge-porn law, which criminalizes the use of images "with the intent to coerce, harass, or intimidate" another person, to encompass falsely created videos.[54] That same year, Texas amended its Election Code to criminalize the creation and distribution of deepfakes intended to harm a political candidate.[55] Under the Texas law, it is a criminal offense to knowingly post a manipulated video of a political candidate within thirty days before an election.[56] California passed a similar election law, making it illegal for anyone to knowingly post a deepfake video relating to a political candidate within sixty days before an election.[57] Other efforts to criminalize deepfake pornography in California have not met with success.[58] However,

to find the creators of false content. Lipkowitz, *supra*. However, the legislation is a step in the right direction toward regulating deepfakes because it (1) draws a clear line between criminal and non-criminal deepfakes, and because (2) current criminal law and tort law do not adequately address harms caused by deepfakes. *Id.*

[53] *See* H.R. 3230, § 1041 (a)–(e) (documenting that H.R. 3230 never advanced out of the House Judiciary Committee during the 116th Congress).

[54] VA. CODE. ANN. § 18.2-386.2 (2020) (criminalizing falsely created pornographic images); H.B. 2678, 2019 Leg., Reg. Sess. (Va. 2019) (providing the amended language to Virginia's original law criminalizing the malicious distribution of pornographic images without the subject's consent); Robert Volker & Henry Ajder, *Analyzing the Commoditization of Deepfakes*, 2020 N.Y.U. J. LEGIS. & PUB. POL'Y QUORUM 22, 27 (crediting Virginia as the first state to criminalize "nonconsensual, 'falsely created,' explicit images and videos [as] . . . a Class 1 misdemeanor").

[55] *See* TEX. ELEC. CODE § 255.004(e) (West 2019) (defining "deep fake video" as "a video created" through AI "with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality"); *see also* Matthew Ferraro, *Texas Law Could Signal More State, Federal Deepfake Bans*, WILMERHALE (Sept. 10, 2019), https://www.wilmerhale.com/en/insights/publications/20190910-law360-texas-law-could-signal-more-state-federal-deepfake-bans (observing that Texas is the first state to enact legislation banning the creation of deepfake videos and the second state to enact criminal penalties for the distribution of deepfake videos).

[56] TEX. ELEC. CODE § 255.004.

[57] CAL. ELEC. CODE § 20010 (West 2020). *See* Will Fischer, *California's Governor Signed New Deepfake Laws for Politics and Porn, but Experts Say They Threaten Free Speech*, BUS. INSIDER (Oct. 10. 2019, 9:51 AM), https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10 [https://perma.cc/D2QQ-HF5F] (describing the California deepfake legislation on elections and reactions to it). Critics of the deepfake legislation say it may hurt free speech principles under the First Amendment because "[t]he law is overbroad, vague, and subjective." *Id.* Assemblyman Bernman countered this by stating, "your words into my mouth, or to use AI technology to take my body and make it look like I did something I never did." *Id.*

[58] In January 2020, the California Assembly introduced a criminal deepfakes bill. Assembly Bill 1903 would have made it a criminal offense to "knowingly, and without the consent of the depicted individual," prepare, produce, or develop any deepfake "that depicts an individual personally engaging in sexual conduct," and to distribute or exchange it with others or offer to do so. A.B. 1903, 2019 Leg., Reg. Sess. (Cal. 2020). The proposed penalty would have included a fine of up to $1,000 and/or up to one year in a county jail. *Id.* The fine would increase to a maximum of $10,000 for depictions of minors, and the jail sentence, if imposed, could be determined pursuant to section 1170(h) of the Penal Code. The proposed Penal Code section would have contained exceptions, including for the use of a "clear disclosure" that "the audio or visual media is not a record of a real event." *Id.* The bill never proceeded beyond committee review. *See id.*

California enacted a new civil right of action allowing victims of deepfake porn to pursue tort remedies.[59]

D. Deepfakes in Court Proceedings

Because of the limited number and scope of national and state laws to contain deepfakes, the increasing ease with which deepfakes are created, and the frequency of their appearance on the internet, deepfakes are not likely to go away any time soon. Moreover, the urgency surrounding deepfakes is underscored by their recent invasion of our legal system—the very fabric of democracy and institutions of truth and justice.

Deepfake evidence has already turned up as a critical issue in two cases that attracted significant attention. In one case in the United Kingdom of a mother who sought to use her husband's threatening audio comments against him in a child-custody trial,[60] the husband was able to show that the audio file was fake, created through software that falsified his voice using metadata analysis.[61] Although the falsification of the audio file was ultimately detected in that case,[62] it serves as a cautionary tale of the power of deepfakes as a source of false evidence, and a preview of what is to come as deepfakes invade legal proceedings.

Another recent headline-grabbing case from the United States illustrates the extent of damage done by even the mere allegation of a deepfake in a legal proceeding. In March 2021, a Pennsylvania woman, Raffaela Spone, was arrested and charged with multiple counts of harassment for allegedly creating deepfakes to frame her daughter's cheerleading rivals.[63] The prosecutor made national[64] and international news[65] at the time asserting that Spone—who became known as "deepfake cheerleader mom"—created a fake video of one teenage girl vaping; that she altered social media accounts of the victims to make

[59] *See* Cal. Civ. Code § 1708.86 (West 2020); A.B. 730, 2019 Leg., Reg. Sess. (Cal. 2019) (setting forth how both the California Code of Civil Procedure and Elections Code were amended by the legislation). The bill's protections are active until January 1, 2023. A.B. 730. *See* A.B. 1903 (introducing a bill to criminalize certain acts using deepfakes and defining a deepfake as "a recording that has been created or altered in a manner that it would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of the individual depicted in the recording"). Assembly Bill 1903 "would . . . criminally prohibit a person from preparing, producing, or developing, without the depicted individual's consent, a deepfake depicting sexual conduct." *Id.*

[60] *See* Matt Reynolds, *Courts and Lawyers Struggle with Growing Prevalence of Deepfakes*, Am. Bar Ass'n: Am. Bar Ass'n J. (June 9, 2020), https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes [https://perma.cc/7ALF-PKVW].

[61] *Id.*

[62] *Id.*

[63] Kim Bellware, *Cheer Mom Used Deepfake Nudes and Threats To Harass Daughter's Teammates, Police Say*, Wash. Post (Mar. 13, 2021, 8:16 PM), https://www.washingtonpost.com/nation/2021/03/13/cheer-mom-deepfake-teammates/.

[64] *Cheerleader's Mom Accused of Making "Deepfake" Videos of Daughter's Rivals*, CBS News (Mar. 15, 2021, 6:39 AM), https://www.cbsnews.com/news/raffaela-spone-cheerleader-mom-deepfakes/.

[65] *Mother 'Used Deepfake To Frame Cheerleading Rivals,'* BBC News (Mar. 15, 2021), https://www.bbc.com/news/technology-56404038.

them appear nude, drinking, and vaping; and that she sent them texts and voicemails telling them to kill themselves.[66] One of the victims even appeared on ABC's *Good Morning America* to recount the harm and distress she had endured as a result of the deepfake.[67]

Spone denied creating the deepfakes, and a firm of technology experts volunteered to help her.[68] Those digital forensics experts and other experts who saw the video determined that it appeared to be authentic rather than a deepfake.[69] They noted, however, that the poor video quality and the lack of other evidence made it impossible to draw any firm conclusions.[70]

In May 2021, the prosecutor's office announced it was no longer pursuing the deepfake video as a basis for the charges after it was revealed that the lead officer in the case had concluded that the video was fake based only on a "naked eye" assessment.[71] Nonetheless, by the time the prosecution had changed course, the damage to Spone had already been done. Spone found herself overwhelmed by negative attention; she received death threats and was ridiculed and harassed in her community and online.[72] According to her lawyer, her reputation was destroyed.[73]

The fact that the deepfake evidence was not used against the victims in both of these cases might lead one to conclude that the concern about deepfake evidence is overblown. But the father in the United Kingdom case and Spone would most certainly disagree with that conclusion. While the deepfake evidence in both cases was not admitted into evidence, the ensuing harm accrued in emotional, reputational, and legal costs for the victims, and disrupted the orderly administration of justice.

Even though our legal system is as vulnerable to content manipulation as any other area of civic life, legislative efforts have not specifically addressed the impact of deepfakes on the court system. Today, video and audio recordings are an indispensable element of some criminal and civil actions. Still, the shadow of uncertainty lingers over each of these proceedings until deepfakes' impact on court proceedings is addressed.

At its essence, the common-law adversarial system depends upon legal advocates' pursuit of their clients' interests by presenting competing versions of

[66] E-mail from Robert J. Birch, Esq., Law Offs. of Robert J. Birch, to Rebecca Delfino, Professor of L., Loy. L. Sch. (Feb. 2, 2022) (on file with author) [hereinafter Birch E-mail].

[67] Drew Harwell, *Remember the 'Deepfake Cheerleader Mom'? Prosecutors Now Admit They Can't Prove Fake-Video Claims*, WASH. POST (May 14, 2021, 9:19 PM), https://www.washingtonpost.com/technology/2021/05/14/deepfake-cheer-mom-claims-dropped/.

[68] Birch E-mail, *supra* note 66.

[69] *Id.*; Harwell, *supra* note 67.

[70] Harwell, *supra* note 67.

[71] *Id.*; Birch E-mail, *supra* note 66. At a subsequent hearing in July 2021, the detective on the case testified that they had no evidence that the video was deepfaked, that there was no evidence that Spone manipulated any social media images, and that there were no threats. Birch E-mail, *supra* note 66.

[72] Harwell, *supra* note 67.

[73] Birch E-mail, *supra* note 66.

their respective cases through the presentation of evidence. Before a court permits evidence to be presented to the trier of fact—the jury—the proponent is required to prove the evidence is real; in other words, that it is an authentic representation of what it purports to show.[74] Lawyers, courts, and juries have been considering evidence in this manner for hundreds of years. This process of determining the truth has functioned effectively because physical evidence can be quickly, efficiently, and reliably evaluated. The system worked, relying on the effectiveness of the innate human ability to determine what is real by trusting one's senses under the theory that "seeing is believing." Deepfake evidence upends this process. As deepfake technology improves and inches closer to becoming indistinguishable from reality, judges and jurors will be hard-pressed to determine with the naked eye whether the evidence is genuine—that is, whether the evidence is what the proponent claims it is.[75]

Preliminarily, it is important not to overstate the problem. To be sure, the threat of deepfake evidence will not touch every court proceeding; not all cases involve digital images or audio evidence. Moreover, as new types of evidence have emerged over the years, lawyers and courts have adapted and managed to resolve new authenticity issues.[76] Eventually, they may also resolve the authenticity challenges presented by deepfakes.[77] However, because deepfakes are more sophisticated than other forms of image manipulation, and because the means of detecting deepfakes has not kept pace with the technology used to create them,[78] the introduction of deepfake evidence in the courtroom raises new, profound issues for the administration of justice in both civil and criminal proceedings.[79] This Subpart identifies the context in which deepfakes may appear in legal proceedings and then explores the unique challenges that deepfakes present to the justice system.

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

[75] *See* Agnieszka McPeak, *The Threat of Deepfakes in Litigation: Raising the Authentication Bar To Combat Falsehood*, 23 VAND. J. ENT. & TECH. L. 433, 443 (2021) (commenting on the speed at which deepfakes are developing and the corresponding responsibility attorneys have to challenge any evidence that is a deepfake).

[76] *See* Pfefferkorn, *supra* note 14, at 255; Brian Barakat & Bronwyn Miller, *Authentication of Digital Photographs Under the "Pictorial Testimony" Theory: A Response to Critics*, FLA. BAR J., Oct. 2004, at 38 (explaining the continued role of "pictorial testimony" in authenticating digital photography).

[77] *See* Paul W. Grimm, Maura R. Grossman & Gordon V. Cormack, *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. 9, 85 (2021) (arguing that the existing Rules of Evidence have the inherent flexibility to address deepfakes).

[78] Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 25–26 (2020) (discussing the belief among the community of computer science and digital forensics experts that detection methods cannot keep pace with the innovations aimed at evading detection).

[79] *See* Kathryn S. Lehman, Scott M. Edson & Victoria Smith, *5 Ways To Confront Potential Deepfake Evidence in Court*, LAW360 (July 26, 2019, 4:59 PM), https://www.law360.com/articles/1181306/5-ways-to-confront-potential-deepfake-evidence-in-court (explaining what deepfake technology is and how it is distinct from general manipulated audio and video); Pfefferkorn, *supra* note 14, at 254.

1. The Context in Which Deepfakes Appear in Legal Proceedings

Deepfake evidence arises in legal proceedings in two distinct circumstances: as the subject of a crime or civil claim, or as an item of evidence offered in a case to prove a separate claim.[80] As discussed elsewhere, federal law does not currently recognize any civil action or crime based on deepfakes.[81] However, a few states have enacted crimes or provided civil causes of action based on the creation and distribution of deepfakes depicting nonconsensual pornography[82] and interfering with elections.[83] The universe of those crimes and torts is still relatively small. It will likely expand as deepfakes continue to abound and cause more harm,[84] and in such proceedings, deepfakes will likely be the central focus of the litigation.

The second and larger context in which deepfake evidence may appear is in every legal proceeding where digital and audio images are presented as a part of the proof in the case. In addition to the deepfake cheerleader mom and United Kingdom custody cases, deepfakes have already arisen as evidence in defamation,[85] child pornography,[86] and assault with attempt to murder[87] cases, as well as in a federal civil rights action.[88] These cases illustrate the dangers that deepfakes currently pose to our justice system and its stakeholders: problems with proving whether evidence is real and defending against allegations that an image is a deepfake. They also highlight how these issues arise even before a trial begins. In either context—whether forming the basis of a claim or used as evidence—deepfakes pose unique challenges to how legal proceedings are conducted, some of which may prove dispositive to a case's outcome.

2. The Challenges Deepfakes Pose in Court.

Deepfakes present three distinct challenges in court proceedings: (1) proving whether a digital image or audio evidence is authentic; (2) responding

[80] Pfefferkorn, *supra* note 14, at 254.

[81] *See supra* Part I.C.

[82] *See supra* Part I.C.

[83] *See supra* Part I.C.

[84] Note also that the deepfake laws in Virginia, Texas, and California that impose criminal or civil liability have yet to be interpreted by the courts.

[85] *See In re* Woori Bank, No. 21-mc-80084-, 2021 WL 2645812, at *1–2 (N.D. Cal. June 28, 2021) (granting the plaintiff's ex parte application to subpoena a social media platform to support his defamation action based on allegations that a "deepfake" image of him engaging in an improper intimate act had been posted on the social media platform).

[86] *See* Schaffer v. Shinn, No. CV 20-08157, 2021 WL 6101435, at *7 (D. Ariz. June 4, 2021) (recommending denial of writ of habeas where petitioner attacked sufficiency of the evidence supporting a sentencing enhancement, arguing that a pornographic image at issue was a deepfake).

[87] *See* People v. Smith, 969 N.W.2d 548, 548 (Mich. Ct. App. 2021) (holding that the trial court did not abuse its discretion in admitting certain Facebook posts into evidence that purportedly included the defendant's image and gang moniker where the defendant alleged that the posts were fake).

[88] *See* Hohsfield v. Staffieri, No. 21-19295, 2021 WL 5086367, at *1 (D.N.J. Nov. 1, 2021) (granting the plaintiff's *in forma pauperis* application where the plaintiff brought a § 1983 action against police officers, claiming that they created a deepfake photo of him engaging in a lewd act to frame him and justify his arrest).

to the "deepfake defense," or the allegation that genuine digital image or audio evidence is a deepfake; and (3) addressing growing distrust and doubt among jurors over the authenticity of all digital image and audio evidence.

a. Proof

The first significant challenge of deepfakes is proving that a piece of digital image or audio evidence is genuine. As explored more fully in the subsequent Part, the Federal Rules of Evidence and case law recognize various ways of authenticating evidence, or "produc[ing] evidence sufficient to support a finding that the item is what the proponent claims it is."[89] However, as deepfake technology advances, those traditional means of authentication may prove time-consuming, costly, and unworkable. Indeed, the naked eye will eventually be unable to discern the subtle clues that indicate that an image, video, or audio clip is fake.[90] Because deepfakes are usually a mash-up of real and fake images, even a percipient fact witness to an event may be unable to authenticate an image they witnessed.[91] Even experts struggle with ascertaining a potential deepfake.

Consequently, proving the genuineness of an image or audio recording may require multiple complicated, expensive proofs to corroborate the evidence at issue. Such proofs may also require more time to develop and secure, causing potential delays in the discovery process and any subsequent trial. The additional resources required of the parties and the court may ultimately impact the parties' success in pursuing their claims.

In addition to the challenge of proving that evidence is real in the era of deepfakes, lawyers may also struggle to demonstrate that evidence is *not* real. Deepfakes will affect counsel's ability to object to the authenticity of evidence. Shallow-fakes, like a video of Nancy Pelosi slurring her words, are manipulations of authentic videos created by slowing down or speeding up sections of footage. Shallow-fakes are easy to debunk, because original videos exist against which the shallow-fake can be compared. However, deepfakes cannot be debunked because no original may exist, and thus deepfakes may not be easily exposed and discredited. On the other hand, if attorneys anticipate an authentication challenge, they may decide that their video's probative value to the case is outweighed by the costs of getting that evidence admitted.[92] The expense and delay may not be feasible for the client. Thus, if the jury is presented with evidence that purports to be real but is, in fact, an undetected deepfake, an individual could be found liable based on fabricated evidence.[93]

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

[90] McPeak, *supra* note 75, at 443.

[91] *See generally* Nils C. Köbis, Barbora Doležalová & Ivan Soraperra, *Fooled Twice: People Cannot Detect Deepfakes but Think They Can*, ISCIENCE, Oct. 29, 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8602050/pdf/main.pdf (demonstrating that people are not able to detect deepfakes reliably).

[92] *Id.*

[93] *See* Marie-Helen Maras & Alex Alexandrou, *Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos*, 23 INT'L J. EVID. & PROOF 255, 255 (2019).

This detrimental impact on justice only amplifies in criminal cases. A criminal defendant's life and liberty could depend on whether that individual can marshal the resources to test the authenticity of evidence offered against them. If a deepfake is believed to be authentic and admitted without challenge, it could well undermine the trial's truth-seeking mission. Even if the defendant has an alibi witness who is willing to testify, in the absence of some way to disprove the clear video evidence depicting the defendant at the scene of the crime, it is hard to imagine the jury interpreting the alibi testimony as anything other than self-serving. Or suppose, on the other side, that the existing mechanisms and tools prove unwieldy or expensive to verify the authenticity of the evidence, leading the prosecution to decide not to prosecute even if the digital evidence is in fact "real."[94] Thus, although videos and audio remain potent pieces of evidence, their authenticity will be hard to gauge in the era of deepfakes.

b. The Deepfake Defense

Deepfakes also present challenges in court proceedings because they create an opportunity to raise new questions, objections, and arguments to even genuine evidence. The fact that deepfakes exist invites parties (and their lawyers) to exploit their existence—to plant seeds of doubt in jurors' minds over the authenticity of all digital audio and images, even when the lawyer knows the evidence is genuine. This "deepfake defense" will debut in court in the foreseeable future, if it has not already.[95] A version of this phenomenon occurred in the deepfake cheerleader mom case—the alleged victim told the police that a genuine video depicting her vaping was fake to support her allegation of criminal harassment.[96] This idea, also known as "the Liar's Dividend,"[97] is built around the premise that genuine audiovisual material can be undermined by a claim that it is fake.

c. Juror Skepticism and Bias

The emergence of deepfakes also presents new challenges around jurors' perception of audiovisual evidence. In general, humans tend to accept images

[94] *See* Agnes E. Venema & Zeno J. Geradts, *Digital Forensics, Deepfakes, and the Legal Process*, SCITECH LAW., July 1, 2020, at 17.

[95] A lawyer defending Guy Reffitt, a leader of the January 6, 2020, insurrection on the United States Capitol, against federal criminal charges argued the deepfake defense to challenge audiovisual evidence demonstrating Reffitt's participation in the insurrection. Dana Verkouteren, *In the First Jan. 6 Trial, a Jury Found Capitol Riot Defendant Guy Reffitt Guilty*, DIGIS MAK (Mar. 8, 2022), https://digismak.com/in-the-first-jan-6-trial-a-jury-found-capitol-riot-defendant-guy-reffitt-guilty/. Lawyers' use of the deepfake defense, its impact on court proceedings, and solutions to curb its use are beyond the scope of this Article.

[96] Birch E-mail *supra* note 66; Harwell, *supra* note 67.

[97] The "Liar's Dividend," coined by professors Robert Chesney and Danielle Citron, is the concept that the accused can create doubt about the accusation simply by questioning its authenticity, or by using altered video or audio evidence that appears to contradict the claim. Chesney & Citron, *supra* note 4, at 1785–86.

and other forms of digital media at face value.[98] The field of psychology teaches that humans value visual perception above other indicators of truth.[99] Thus, the legal system has historically favored admitting audiovisual evidence.[100] Studies demonstrate that jurors who hear oral testimony along with video testimony are 650% more likely to retain the information.[101] Indeed, studies have demonstrated that video evidence powerfully affects human memory and perception of reality.[102]

The internet has further elevated videos and images as sources of factual information. More Americans now get their news from social media rather than print media.[103] With Twitter, YouTube, Facebook, and other similar platforms gaining market share, social media continues to elevate audiovisual content—

[98] Richard K. Sherwin, Neal Feigenson & Christina Spiesel, *Law in the Digital Age: How Visual Communication Technologies Are Transforming the Practice, Theory, and Teaching of Law*, 12 B.U. J. SCI. & TECH. L. 227, 246 (2006). Studies have shown over and over again that people tend to believe what they see, despite knowing that videos can misrepresent facts. *See* Yael Granot, Neal Feigenson, Emily Balcetis & Tom Tylr, *In the Eyes of the Law: Perception Versus Reality in Appraisals of Video Evidence*, 24 PSYCH. PUB. POL'Y & L. 93, 97–98 (2017) (warning how powerful video evidence can be in convincing people that a fake event occurred). In a study conducted by a bank where no participants illicitly took money, the bank was still able to convince participants that they stole money after showing them a doctored video depicting them doing so. *Id.* After watching the video, despite knowing that they did not steal, participants would confess to taking money from the bank. *Id.*

[99] *See* Carolyn Purnell, *Do We All Still Agree That "Seeing Is Believing"?*, PSYCH. TODAY (June 23, 2020), https://www.psychologytoday.com/us/blog/making-sense/202006/do-we-all-still-agree-seeing-is-believing [https://perma.cc/GK2H-B7Q7].

[100] *See* Granot et al., *supra* note 98, at 93 (describing the holding in *United States v. Watson*, 483 F.3d 828 (D.C. Cir. 2007)). Prosecutors striking blind persons from the jury demonstrates the idea that one must be able to see in order to fully comprehend all of the evidence in a case. *Id.* Visual evidence is so highly regarded in the justice system that it is seen as a way to mitigate juror bias toward other pieces of evidence in a case. *Id.*

[101] *See* Karen Martin Campbell, *Roll Tape—Admissibility of Videotape Evidence in the Courtroom*, 26 U. MEM. L. REV. 1445, 1447 (1996) (providing statistics on how jurors retain videotaped information at trial). Jurors who received visual testimony were 100% more likely to retain information than jurors who received only oral testimony. *Id.*; *see also* Zachariah B. Parry, *Digital Manipulation and Photographic Evidence: Defrauding the Courts One Thousand Words at a Time*, 2009 U. ILL. J.L. TECH. & POL'Y 175, 185 (citing statistics on the impact of visual evidence on jurors). "Jurors often are bored, confused, and frustrated when attorneys or witnesses try to explain technical or complex material," and having visual aids can help them retain information much better. Parry, *supra*, at 184. Jurors can retain up to 85% of visual information; by contrast, they retain only about 10% of what they hear. *Id.* at 185.

[102] Kimberly A. Wade, Sarah L. Green & Robert A. Nash, *Can Fabricated Evidence Induce False Eyewitness Testimony?*, 24 APPLIED COG. PSYCH. 899, 900 (2010). In 2010, researchers at the University of Warwick conducted a study on the psychological effect that video has on reconstructing personal observations. *Id.* The researchers placed sixty college students in a room to engage in a computerized gambling task. *Id.* at 901–02. Following completion of the task, researchers individually showed each subject a digitally altered video depicting a co-subject cheating, when in fact none of the subjects had cheated. *Id.* at 903–04. Nearly half of the subjects were willing to testify that they had personally witnessed a co-subject cheating after seeing the fake video; only one in ten was willing to testify to the same effect after the researcher merely told the subject about the cheating, rather than showing the fake video evidence. Hadley Leggett, *Fake Video Can Convince Witnesses To Give False Testimony*, WIRED (Sept. 14, 2009, 6:02 PM), https://www.wired.com/2009/09/falsetestimony [https://perma.cc/M88G-8TKJ] ("[R]esearchers emphasized that no one should testify unless they were 100% sure they had seen their partner cheat.").

[103] Elisa Shearer, *Social Media Outpaces Print Newspapers in the U.S. as a News Source*, PEW RSCH. CTR. (Dec. 10, 2018), https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/ [https://perma.cc/5B94-6B46].

and videos in particular—over other content formats.[104] But social media also spreads deepfakes and drives news coverage of deepfakes.[105] And as the public discovers the potential to be fooled by deepfakes, it will also come to doubt authentic videos. For instance, Gabon's President Ali Bongo suffered a stroke in late 2017 and was out of the public eye for months. It was rumored that Bongo was critically ill or dead.[106] In response, the government released a video of Bongo, meant to quell the rumors and alleviate public concern. However, the video was attacked as a deepfake that exacerbated speculation over Bongo's condition.[107] Controversy ignited by the video even led to an unsuccessful military coup.[108] Speculation persists over whether the video of Bongo is a deepfake. The mistrust and instability that resulted illustrate the public's inability to gauge authenticity in the age of deepfakes and the danger posed by related skepticism.[109]

As public knowledge of deepfakes continues to grow and people become increasingly skeptical about the credibility of audiovisual images, jurors will be primed to doubt the authenticity of even real audio and video content.[110] Juror skepticism may lead to plummeting juror confidence in video evidence absent a sponsoring witness, even if a judge authenticates the video.[111] Moreover, juror skepticism is problematic because it may allow bad actors to escape accountability simply because the jury has no means of determining that a piece of content is not, in fact, a deepfake.[112] If this tactic of challenging authenticity is used successfully in multiple cases, video evidence may ultimately lose its

[104] *See* Deep Patel, *12 Social Media Trends To Watch in 2020*, ENTREPRENEUR (Dec. 20, 2019), https://www.entrepreneur.com/article/343863 [https://perma.cc/7ZKR-MV89].

[105] *See*, *e.g.*, Ali Breland, *The Bizarre and Terrifying Case of the "Deepfake" Video That Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [https://perma.cc/9VSS-FD8G].

[106] *Id.*

[107] *Id.*

[108] *Id.*

[109] *See* Janosch Delcker, *Welcome to the Age of Uncertainty*, POLITICO (Dec. 17, 2019, 7:50 PM), https://www.politico.eu/article/deepfake-videos-the-future-uncertainty/ [https://perma.cc/LSD5-RXB3].

[110] *Id.*; *see* Reynolds, *supra* note 60; Brown, *supra* note 78, at 25–26 (asserting that even if a realistic deepfake is publicly identified as fake, it is unclear that the public would believe it).

[111] *See* Mika Westerlund, *The Emergency of Deepfake Technology: A Review*, 9 TECH. INNOV. MGMT. REV. 39, 42–43 (2019) (describing how the public may begin to distrust authorities deemed reliable in the past because of deepfakes); Nicholas Mirra, *Putting Words in Your Mouth: The Evidentiary Impact of Emerging Voice Editing Software*, 25 RICH. J.L. & TECH. 1, 3 (2018) (cautioning that courts must be prepared for how "new technology may threaten existing and well established forms of evidence"); Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing*, 27 CATH. U. J.L. & TECH. 51, 58 (2018) (contending that video may lose its value because "[t]he same accountability video that brings action can now be abused in a number of ways"); Drew Harwell, *Top AI Researchers Race To Detect 'Deepfake' Videos: 'We Are Outgunned*,*'* WASH. POST (July 12, 2019, 4:44 PM), https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/ [https://perma.cc/3N8Y-C484] (warning how the public may begin to generally distrust video footage because "[i]t's too much effort to figure out what's real and what's not").

[112] Pfefferkorn, *supra* note 14, at 269–70.

persuasive power and, if taken far enough, degrade public trust in the very institution of the courts.[113]

II. Existing Mechanisms To Address the Challenges of Deepfakes in Legal Proceedings and Why They Are Inadequate

The law offers various mechanisms to deal with the admission of novel evidence, many of which might be applied to deepfakes. Specifically, as explored in this Part, courts and lawyers will look to the Rules of Evidence to address the challenges presented by deepfakes in legal proceedings. However, as this Part ultimately argues, contemporary legal frameworks are insufficient to address the potential harm that deepfakes pose.

Historically, common-law standards governed the admissibility of scientific, photographic, and video evidence. Initially, the legal standards governing the admissibility of this evidence were strict. In many instances, the trial court determined the issue of authenticity. But as the public became exposed to photographs, x-rays, audiovisual recordings, and new scientific evidence like DNA, and as that evidence became more common, lawyers, jurors, and courts grew more comfortable with their admission in legal proceedings.[114] Eventually, most common-law practices on the admission of evidence gave way to the Federal Rules of Evidence.[115] However, even after enacting the Federal Rules of Evidence, courts continued to look to two common-law theories—the pictorial communication theory and the silent witness theory—to authenticate photographic and video evidence under Rule 901(b).[116]

As our society has transformed into one dominated by new, complicated technologies, courts have had to consider whether the existing rules of admissibility are sufficient to address new categories of evidence.[117] This Part discusses the current approaches to admissibility standards under the Federal Rules of Evidence and the two common-law theories, while placing them in the

[113] *Id.* at 270–71.

[114] Jill Witkowski, *Can Juries Really Believe What They See? New Foundational Requirements for the Authentication of Digital Images*, 10 Wash. U. J.L. & Pol'y 267, 279 (2002) ("Over time, however, the courts replaced the strict foundational requirements concerning the process of taking motion pictures with the admission of witness testimony that the film was a fair and accurate representation of what actually happened."); *see also* Edward J. Imwinkelried, Evidentiary Foundations § 4.09[2] (9th ed. 2015) (stating that although "the courts were initially very conservative in their treatment of motion pictures," "[t]he law governing the admission of motion pictures has been liberalized in recent years").

[115] An Act To Establish Rules of Evidence for Certain Courts and Proceedings, Pub. L. No. 93-595, 88 Stat. 1926 (1975) (codified as amended at 28 U.S.C. §§ 2072–2074). The Judicial Conference responsible for implementing the Rules Enabling Act of 1934 did not formally study a uniform evidence code until 1961 and finally submitted its proposed rules to Congress for approval in 1972. Paul R. Rice & Neals-Erik William Delker, *Federal Rules of Evidence Advisory Committee: A Short History of Too Little Consequence*, 191 F.R.D. 678, 682–84 (2000).

[116] Pub. L. No. 93-595, 88 Stat. 1926 (1975) (codified as amended at 28 U.S.C. §§ 2072–2074); *see also* Imwinkelried, *supra* note 114, § 4.01[1] (outlining the procedure for authentication under Rule 901).

[117] McPeak, *supra* note 75, at 441.

historical context of the admission of other types of evidence that have, like deepfakes, been prone to manipulation and juror confusion.

A. Historical Treatment of Visual, Audio, Digital, and Scientific Evidence and Its Lessons for Dealing with Deepfakes

The rules of evidence originated in the common law.[118] Historically, courts approached the admission of novel types of evidence with caution and suspicion, imposing strict, high bars to admissibility to demonstrate the reliability and authenticity of the evidence. The traditional treatment of visual, audio, digital, and scientific evidence is illustrative of this approach and may be predictive of courts' approach to deepfakes.

1. Standards for Admission of Photographs and X-Rays

Although photographic evidence became a means of persuading the jury in legal proceedings by the end of the nineteenth century,[119] courts were initially hesitant to admit photographs into evidence.[120] This hesitancy centered on whether a witness could testify on behalf of a photograph. *United States v. Ortiz* is illustrative.[121] In *Ortiz*, the Supreme Court allowed the admission of a photograph only after the photographer testified regarding the process used to take a photograph.[122] 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."[123] The standards regarding the admissibility of photographic evidence relaxed further after the 1988 case *United States v. Rembert*.[124] The *Rembert* court concluded "that all that is necessary to meet the threshold requirement of authentication is a 'showing sufficient to permit a reasonable juror to find that the evidence is what its proponent claims.'"[125]

Courts have taken an equally flexible approach to the admission of x-ray evidence. For example, in *State v. Matheson*, the court admitted an x-ray photograph into evidence, even though a witness was not available to testify on the photograph's accuracy.[126] The court nevertheless admitted the x-ray,

[118] John H. Langbein, *Historical Foundations of the Law of Evidence: A View from the Ryder Sources*, 96 COLUM. L. REV. 1168, 1170–72 (recounting the history of the rules of evidence from the common law).

[119] Jennifer L. Mnookin, *The Image of Truth: Photographic Evidence and the Power of Analogy*, 10 YALE J.L. & HUMANS. 1, 2, 5 (1998).

[120] Catherine Guthrie & Brittan Mitchell, *The Swinton Six: The Impact of* State v. Swinton *on the Authentication of Digital Images*, 36 STETSON L. REV. 661, 677–78 (2007).

[121] 176 U.S. 422 (1900).

[122] *Id.* at 430.

[123] Guthrie & Mitchell, *supra* note 120, at 678.

[124] 863 F.2d 1023, 1026 (D.C. Cir. 1988).

[125] *Id.* at 1027 (quoting United States v. Blackwell, 694 F.2d 1325, 1330 (D.C. Cir. 1982)).

[126] 103 N.W. 137, 138–39 (Iowa 1905).

recognizing the skill of the individual who took the image and the value it provided to the case.[127]

2. *Standards for Admission of Audio and Video Recordings*

Before the Federal Rules of Evidence were enacted in 1975, courts throughout the United States imposed stringent requirements for authenticating audio evidence.[128] In 1958, in *United States v. McKeever*,[129] the court articulated seven requirements for admissibility, which became the paradigm. In *McKeever,* the defendants sought to admit an audio-recorded conversation between one of the defendants and a witness.[130] 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."[131] For decades, federal courts tested the admission of audio evidence against these requirements.[132]

Similarly, courts initially applied strict standards to the admissibility of video evidence.[133] However, as videos became more common in society, courts began to apply the *McKeever* test to determine the admissibility of video evidence, as well.[134] And as photographs, motion pictures, and recordings became more familiar and common in daily life, their use in court expanded.[135] Over time, courts relaxed the *McKeever* test and eventually set it aside in favor of more lenient standards.[136] Interpreting the *McKeever* test as "a guide rather

[127] *Id.*

[128] Clifford S. Fishman, *Recordings, Transcripts and Translations as Evidence*, 81 WASH. L. REV. 473, 478 (2006).

[129] 169 F. Supp. 426, 430 (S.D.N.Y. 1958), *rev'd on other grounds*, 271 F.2d 669 (2d Cir. 1959).

[130] *Id.* at 428; Witkowski, *supra* note 114, at 276–77.

[131] *McKeever*, 169 F. Supp. at 430 ("[B]efore a sound recording is admitted into evidence, a foundation must be established by showing the following facts: (1) That the recording device was capable of taking the conversation now offered in evidence. (2) That the operator of the device was competent to operate the device. (3) That the recording is authentic and correct. (4) That changes, additions or deletions have not been made in the recording. (5) That the recording has been preserved in a manner that is shown to the court. (6) That the speakers are identified. (7) That the conversation elicited was made voluntarily and in good faith, without any kind of inducement.").

[132] For example, *United States v. Branch* stated that the factors in *McKeever* provide guidance for district courts authenticating audio evidence. 970 F.2d 1368, 1371–72 (4th Cir. 1992). In *Branch*, however, the Fourth Circuit held that the proponent of a recording did not need to establish each of the seven requirements. *Id.* Rather, the court suggested that these requirements should be used as guidance to determine whether the recording was authentic. *Id.* Many courts have followed the example of *Branch* and applied a more flexible approach to the *McKeever* test. Witkowski, *supra* note 114, at 278.

[133] Witkowski, *supra* note 114, at 279.

[134] *Id.*

[135] *See* 2 KENNETH S. BROUN, MCCORMICK ON EVIDENCE § 215 (7th ed. 2013) (describing the different ways that photographs are used in courts). "As judges, counsel and the lay public have become accustomed to the prevalence of such recordings in court, their persuasive potential is both widely acknowledged and the subject of concern." *Id.* § 216.

[136] Witkowski, *supra* note 114, at 279 ("Over time, however, the courts replaced the strict foundational requirements concerning the process of taking motion pictures with the admission of witness testimony that the film was a fair and accurate representation of what actually happened."); *see also* IMWINKELREID, *supra* note

than a rule," courts adopted more relaxed tests and 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.[137]

As Congress enacted the Federal Rules of Evidence in the 1970s and transitioned away from the common law, the admissibility of audio and video evidence became more flexible.[138] Congress incorporated much of the same common-law standards used by courts after the relaxation of the *McKeever* test into the Rules, such as "relevance (codified in Federal Rule of Evidence Rule 401), probative value balanced against undue prejudice (codified in Federal Rule of Evidence Rule 403), and accuracy (codified in the sufficient to support a finding standard in Rule 901)."[139]

The history of the admissibility of audio and video evidence has lessons for deepfakes. It suggests that a "go-slow-and-strict" approach to the rules of authenticity might be required in the near and short-term future to allow for the development of better technologies that can detect deepfakes.[140] It also teaches that affording courts maximum flexibility to look at the totality of the evidence to determine authenticity is essential. Courts will want to have at their disposal the full, unlimited range of means to evaluate the evidence under Federal Rule of Evidence 901.

3. Standards for Admission of Social Media and Digital Images

Social media evidence has become increasingly important evidence in legal proceedings.[141] Although some courts were initially reluctant to admit social media evidence, the advancement of social media has increased the need for the admissibility of such evidence to resolve legal disputes more efficiently.[142]

However, courts across the United States have not adopted a uniform approach to admitting social media evidence.[143] Some courts have followed a

114, § 4.09[2] (stating that although "the courts were initially very conservative in their treatment of motion pictures," "[t]he law governing the admission of motion pictures has been liberalized in recent years").

[137] Witkowski, *supra* note 114, at 278; *see also Branch*, 970 F.2d at 1371–72 (finding the *McKeever* factors sufficient but not required to establish a foundation for authenticity); United States v. Biggins, 551 F.2d 64, 66–67 (5th Cir. 1977) (holding that the court "neither adopt[ed] nor reject[ed] [the *McKeever* test] as a whole" and looking to four factors as a guide without sacrificing evidence to a "formalistic adherence" to a judicially imposed standard).

[138] An Act To Establish Rules of Evidence for Certain Courts and Proceedings, Pub. L. No. 93-595, 88 Stat. 1926 (1975) (codified as amended at 28 U.S.C. §§ 2072–2074); Rice & Delker, *supra* note 115.

[139] Witkowski, *supra* note 114, at 279–80; *see* FED. R. EVID. 401, 403, 901.

[140] *See* Brown, *supra* note 78, at 25 (citing forensic digital experts and computer scientists reporting that effective technology to "comprehensively identify deepfakes is years-away"); Kaveh Waddell, *The Impending War over Deepfakes*, AXIOS (July 22, 2018), https://www.axios.com/2018/07/22/the-impending-war-over-deepfakes.

[141] Lawrence Morales II, Discoverability and Admissibility of Electronic Evidence (2017) (unpublished manuscript) (on file with author).

[142] Siri Carlson, Comment, *When Is a Tweet Not an Admissible Tweet? Closing the Authentication Gap in the Federal Rules of Evidence*, 164 U. PA. L. REV. 1033, 1034 (2016).

[143] Kathryn S. Lehman & Lindsey Macon, *Social Media in the Courtroom*, FOR DEF., May 2019, at 23.

stricter approach to authenticating social media evidence.[144] One such example is the Maryland Court of Appeals' approach in *Griffin v. State*,[145] where a printout of a Myspace profile was introduced into evidence at trial to prove that the defendant's girlfriend had threatened a witness.[146] The defendant challenged the introduction of the evidence, arguing that no evidence was offered to show how the printout was obtained and how the page was linked to the defendant's girlfriend.[147] Although the state provided information like the girlfriend's profile photograph and personal information to link her to the page, the Maryland Court of Appeals determined that such facts were not "distinctive characteristics" of the social media profile sufficient to authenticate it.[148] The court reasoned that due to the high manipulability of social media, such evidence requires a higher "degree of authentication."[149]

Courts' inconsistent responses to questions regarding the authenticity of social media evidence are exemplified by a pair of California cases. In *People v. Beckley*, the prosecution offered a photo a detective had downloaded from Myspace to rebut a defense witness's testimony.[150] The witness testified that she and her boyfriend had no association with a gang, but the photo on the Myspace page appeared to show otherwise.[151] In this situation, the detective could not testify from any personal knowledge that the person depicted in the photo was the witness, and there was no expert witness to testify that the photo was not a fake.[152] The Court of Appeal rejected the prosecution's effort to admit the image, because neither an expert nor a fact witness could authenticate it. In reversing the conviction, the appellate court referenced the potential fabrication of the evidence, observing that the websites were not monitored for accuracy nor subject to independent verification. Reasoning that photos are susceptible to alteration, the court held the photo inadmissible.[153]

Four years later, another California appellate court rejected *Beckley*'s approach. *In re KB*[154] concerned social media evidence from an Instagram post used to secure a conviction for illegal firearms possession. The court affirmed the conviction, finding that neither expert testimony nor eyewitness testimony was necessary to authenticate the post.[155] Instead, the court recognized that the

[144] Carlson, *supra* note 142, at 1046.

[145] 19 A.3d 415 (Md. 2011).

[146] *Id.* at 417.

[147] *Id.* at 417–19; *see also* Lehman & Macon, *supra* note 143, at 24.

[148] *Griffin*, 19 A.3d at 424.

[149] *Id.*

[150] 110 Cal. Rptr. 3d 362, 366 (Cal. Ct. App. 2010); *see also* Reynolds, *supra* note 60, at 2 ("In the 2010 case, . . . the California 2nd District Court of Appeal ruled that prosecutors should not have admitted a MySpace image claiming to show the girlfriend of a defendant flashing a gang sign because neither a witness nor an expert authenticated it.").

[151] *Beckley*, 110 Cal. Rptr. 3d at 367.

[152] *Id.* at 366.

[153] *Id.*

[154] 190 Cal. Rptr. 3d 287, 289 (Cal. Ct. App. 2015).

[155] *Id.* at 294.

evidence could be authenticated by other corroborating evidence, including testimony describing the basic functionality of Instagram by the investigating officer, corroborating information from the social media content itself, and the fact that the social media account was password protected.[156]

Other courts have followed a similar liberal approach to evaluate whether social media evidence is admissible by applying the standard requirement for authenticity: "whether the proponent has produced sufficient evidence for a reasonable jury to find that the proffered evidence is authentic."[157] For example, in *Tienda v. State*, the court held that the trial court did not abuse its discretion by allowing the admission of Myspace pages in which the defendant referenced a homicide.[158] The Texas Court of Appeals ruled that "[t]he preliminary question for the trial court to decide is simply whether the proponent of the evidence has supplied facts that are sufficient to support a reasonable jury determination that the evidence he has proffered is authentic."[159]

Like social media evidence, digital images must be authenticated to be admitted.[160] However, the Federal Rules of Evidence do not provide clear guidance for authenticating digital images;[161] thus, their admissibility has been determined using the same tests for authenticating traditional photographs.[162]

The common-law approach to authenticating digital images placed a high burden on the proponent of the evidence.[163] For example, in *Kaps Transport v. Henry*, the authenticity of a digital photograph was challenged because of a concern that it had been altered.[164] The court held that the photograph could be admitted if the proponent could reconcile the imperfections of the photograph.[165] However, the ruling of this case did not set forth a standard for the admissibility of digital images.[166]

One of the most impactful cases regarding the admissibility of social media evidence is *State v. Swinton*,[167] in which the defendant challenged the admissibility of photographs of bitemark evidence, some of which were software-enhanced, and some of which were created with photoshop software.[168] The Connecticut Supreme Court ultimately decided to adopt the following six factors for the authentication of evidence generated or enhanced by a computer:

[156] *Id.*
[157] Carlson, *supra* note 142, at 1046–47.
[158] *Id.* at 1047.
[159] Tienda v. State, 358 S.W.3d 633, 638 (Tex. Crim. App. 2012).
[160] Witkowski, *supra* note 114, at 273–74.
[161] *Id.* at 274.
[162] Parry, *supra* note 101, at 187–88.
[163] Witkowski, *supra* note 114, at 281.
[164] 572 P.2d 72, 75 (Alaska 1977).
[165] *Id.* at 76.
[166] Parry, *supra* note 101, at 193.
[167] 847 A.2d 921 (Conn. 2004); Guthrie & Mitchell, *supra* note 120, at 681.
[168] *Swinton*, 847 A.2d at 932.

> (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.[169]

4. Common-Law Standards for Admission of Scientific Evidence

Finally, given the complex and technical nature of the technology used to create deepfakes, the determination of whether deepfake evidence is authentic and thus admissible under the Federal Rules may also be informed by the rules courts have applied for admitting scientific, technical, or other specialized information beyond the understanding of lay jurors and many judges.

For almost fifty years, federal courts applied the standard established in *Frye v. United States*[170] to determine the admissibility of novel scientific evidence requiring an expert's scientific testimony.[171] The *Frye* standard, known as the "general acceptance" standard, held that the admissibility of novel scientific evidence is determined by whether the technique has been "sufficiently established to have gained general acceptance in the particular field in which it belongs."[172] It "placed a 'gatekeeping' responsibility upon judges to ensure that scientific evidence presented before a jury enjoyed 'general acceptance in the particular field to which it belongs.'"[173]

In 1993, the Supreme Court abandoned the *Frye* test in *Daubert v. Merrell Dow Pharmaceuticals*.[174] In *Daubert*, the Court held that the *Frye* standard of admissibility was incompatible with the Federal Rules of Evidence, which approached novel scientific evidence more broadly.[175] Ultimately, the *Daubert* court changed the standard of admissibility for new scientific evidence from "general acceptance" to "reliability."[176] In 1999, the Court in *Kumho Tire Co. v. Carmichael*[177] further expanded the reach of the reliability test in *Daubert* by applying it to nonscientific expert testimony.[178]

[169] *Id.* at 942.

[170] 293 F. 1013 (D.C. Cir. 1923).

[171] Kaushal B. Majmudar, Daubert v. Merrell Dow*: A Flexible Approach to the Admissibility of Novel Scientific Evidence*, 7 HARV. J.L. & TECH. 187, 199 (1993).

[172] 293 F. at 1014.

[173] Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from* Jennings *to* Llera Plaza *and Back Again*, 41 AM. CRIM. L. REV. 1189, 1221 (2004).

[174] 509 U.S. 579, 589 (1993).

[175] *Id.* at 588–89; Majmudar, *supra* note 171, at 199–200.

[176] Cole, *supra* note 173, at 1221–22 (stating that Federal Rule of Evidence 702 superseded the *Frye* standard and allows the introduction of new scientific evidence if such evidence will assist the trier of fact in understanding the evidence or determining a fact at issue); George Bundy Smith & Janet A. Gordon, *The Admission of DNA Evidence in State and Federal Courts*, 65 FORDHAM L. REV. 2465, 2480 (1997).

[177] 526 U.S. 137 (1999).

[178] *Id.* at 147–49; Paul C. Giannelli, *Daubert Challenges to Fingerprints*, 42 CRIM. L. BULL. 624, 624–25 (2006).

The *Daubert* factors were added to the Rules in 2000. 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 particular case.[179] Therefore, the factors discussed in the *Daubert* decision regarding the reliability of scientific or technical evidence are informative when determining whether Rule 702's reliability requirement has been met. 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.[180]

The evolution of the admissibility standards for scientific evidence and expert testimony from *Frye* to *Daubert* is exemplified in the treatment of DNA evidence, which has become increasingly important in the courtroom as a vital tool linking a defendant to the scene of the crime.[181] The first time DNA evidence was used to find a defendant guilty in the United States was in *Andrews v. State* in 1988.[182] In *Andrews*, the trial court conducted an evidentiary hearing, applied the *Frye* standard, and admitted the DNA evidence, concluding that it was generally accepted by experts in the field.[183] After a hung jury and a retrial where the DNA evidence was admitted again, the jury found the defendant guilty, and his conviction was affirmed on appeal.[184]

Later, the Eighth Circuit in *United States v. Martinez* analyzed the impact of the *Daubert* standard on the admissibility of DNA evidence,[185] holding that the *Daubert* standard required courts to engage in an inquiry of reliability "through preliminary hearing to determine if the expert properly performed the scientific procedure."[186] Thus, under *Daubert*, DNA evidence is now universally admitted into evidence "so long as proper procedures are undertaken in the lab."[187] The manner in which new scientific evidence such as DNA has been

[179] *See* FED. R. EVID. 702 (b)–(d); *see also In re* Paoli R.R. Yard PCB Litig., 35 F.3d 717, 742 (3d Cir. 1994) (discussing the importance of the reliability factor in the *Daubert* analysis, and the obligation of the trial judge to "take into account" all of the factors listed in *Daubert* relevant to determining the reliability of the scientific or technical evidence at issue).

[180] *See* FED. R. EVID. 702 advisory committee's note to 2000 amendment.

[181] Smith & Gordon, *supra* note 176, at 2465.

[182] AARON DANIEL BOBER & TYLER A. LONGMIRE, DNA FINGERPRINTING 28 (2004), https://core.ac.uk/download/pdf/212990088.pdf.

[183] *Id.*

[184] *Id.* at 28–29.

[185] 3 F.3d 1191 (8th Cir. 1993); R. Stephen Kramer, *Admissibility of DNA Statistical Data: A Proliferation of Misconception*, 30 CAL. W. L. REV. 145, 159 (1993).

[186] Kramer, *supra* note 185, at 160.

[187] Parry, *supra* note 101, at 191.

treated is a cautionary tale. Like DNA, deepfakes raise complex issues of reliability and authenticity. Detecting a deepfake, like comparing DNA samples, requires a level of expertise beyond the knowledge of most lay people, lawyers, and judges.

This history of the various approaches courts have taken in the admissibility of novel or scientific evidence shows that courts initially proceed with caution and suspicion—initially imposing strict and high bars to admissibility in demonstrating the reliability and authenticity of evidence, and assigning the preliminary fact determination to the judge. Courts' historical treatment of visual, audio, and digital scientific evidence charts the approach they will take with deepfakes. This history informs the current landscape of the evidentiary rules and theories of admissibility.

B. Current Evidence Rules and Theories

The purpose of the Federal Rules of Evidence is "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."[188] The Rules reflect the belief that an adversarial justice system is ideal for reaching the truth through litigation. Thus, the proponent of evidence bears the burden of establishing relevance and authenticity.[189] Other limitations also apply to hearsay evidence and balancing the evidence's probative value with unfair prejudice.[190] But even disputed evidence may be admissible, as opposing parties can present competing evidence, cross-examine witnesses, and otherwise seek out the truth throughout the litigation process. And after considering all the evidence presented, a trier of fact decides whether the proffered evidence is authentic or not.[191] This Subpart provides an overview of the Federal Rules of Evidence, focusing on the Rules on authentication.

1. The Federal Rules of Evidence

The Federal Rules of Evidence dictate that only relevant evidence is admissible.[192] Rule 104(b) provides that the preliminary admissibility standard for relevance depends on a fact, and states that "[w]hen the relevance of evidence depends on whether a fact exists, proof must be introduced sufficient to support

[188] Fed. R. Evid. 102.

[189] Fed. R. Evid. 901(a) (requiring that the proponent of the evidence show that the evidence is what it purports to be); Fed. R. Evid. 401 (requiring that evidence must have the tendency to make some fact that is of consequence to the litigation more or less probable).

[190] Fed. R. Evid. 403 (setting forth the balancing test that allows otherwise admissible evidence to be excluded on the basis of unfair prejudice outweighing the evidence's probative value); Fed. R. Evid. 801 (defining hearsay); Fed. R. Evid. 802 (hearsay exceptions).

[191] *See* Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 540 (D. Md. 2007) (determining that the question for the court under Rule 901 was "whether the proponent of evidence . . . 'offered a foundation from which the jury could reasonably find that the evidence [wa]s what the proponent sa[id] it [wa]s'").

[192] Fed. R. Evid. 401 (requiring that evidence must have the tendency to make some fact that is of consequence to the litigation more or less probable).

a finding that the fact does exist."[193] According to the Advisory Committee, "[a]uthentication and identification represent a special aspect of relevancy," as evidence must be authentic for it to be relevant.[194] 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)."[195] Thus, under the Federal Rules of Evidence, evidence can be deemed relevant and admissible only if it is authentic.[196]

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."[197] The authenticity of evidence is ultimately a factual determination for the trier of fact—traditionally a jury—to evaluate.[198] 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."[199] 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;[200] arguments concerning the unreliability of the evidence go to the weight of the evidence and not admissibility.[201] 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.[202]

[193] FED. R. EVID. 104(b); *see also* Edward J. Imwinkelried, *"Where There's Smoke There's Fire": Should the Judge or the Jury Decide the Question of Whether the Accused Committed an Alleged Uncharged Crime Proffered Under Federal Rule of Evidence 404*, 42 ST. LOUIS U. L.J. 813, 824 (explaining the development of what has been come to be known as the "conditional relevance rule" in applying Rule 104(b)).

[194] FED. R. EVID. 901(a) advisory committee's note.

[195] *Id.*

[196] FED. R. EVID. 901; *see also* United States v. Vayner, 769 F.3d 125, 129 (2d Cir. 2014) ("The requirement of authentication is . . . a condition precedent to admitting evidence." (quoting United States v. Sliker, 751 F.2d 477, 497 (2d Cir. 1984))).

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

[198] FED. R. EVID. 104 advisory committee's note ("If the evidence is not such as to allow a finding [that a jury could reasonably conclude authenticity], the judge withdraws the matter from their consideration.").

[199] *Id. See* IMWINKELRIED, *supra* note 114, § 4.01[1] (outlining the procedure for authentication under Rule 901).

[200] United States v. Branch, 970 F.2d 1368, 1371 (4th Cir. 1992).

[201] *See* United States v. Capers, 61 F.3d 1100, 1106 (4th Cir. 1995) (determining that arguments on the reliability of the witness's identification of the voices on a tape recording went to the weight, and not admissibility, of the evidence).

[202] *See, e.g.*, United States v. Reilly, 33 F.3d 1396, 1404 (3d Cir. 1994) ("[T]he burden of proof for authentication is slight.").

a. The Evolution of the Allocation of Fact Determinations Between the Judge and Jury: A Brief History of Federal Rule of Evidence 104

This split of factfinding duties between the judge and the jury, embodied in Federal Rule of Evidence 104(a) and (b) and mirrored in the authenticity determinations under Rule 901, is an offspring of twentieth-century American law. Before that, the prevailing view, based on the English common-law approach, was that "the judge had plenary authority to decide all questions of fact[,] conditioning the admissibility of testimony."[203]

> English judges considered the testimony on both sides of the foundational testimony and resolved any incidental questions of authenticity and credibility. Even after the American Revolution, American courts tended to follow the British practice. . . . In short, until the modern era, there was virtually universal agreement that whenever the application of an evidentiary rule to an item of proffered testimony necessitated the resolution of a factual question, the judge—and the judge alone—decided the question.[204]

However, the traditional approach was eventually challenged by Jacksonian Democrats.[205] The Jacksonians were afraid of a powerful aristocratic judiciary.[206] They openly distrusted judges, worrying that judges would dictate the results in cases and undermine the jury's role. Thus, they favored the popular election of judges[207] and advocated a shift of preliminary factfinding power to the jury.[208] Based on this critique of the traditional English view, "a few isolated American opinions deviated from the English practice."[209]

> [M]ajor inroads in the English practice did not occur until the 1920s. . . . The new theory shifted to the jury all the preliminary factfinding power which the commentators believed lay jurors could be trusted with. In particular, the theory posited that the jury should be able to decide whether a lay witness has personal knowledge of the event [they] . . . testif[ied] about. . . .[210]

[203] 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* 45 AM. JUR. TRIALS §§ 6–9 (2019) (tracing the history of Federal Rule of Evidence 104).

[204] Imwinkelried, *supra* note 203, at 8–9.

[205] *See id.*; Edmund M. Morgan, *Functions of Judge and Jury in the Determination of Preliminary Questions of Fact*, 43 HARV. L. REV. 165, 191 (1929).

[206] Donald T. Weckstein, *Round Table Discussion of the Proposed Code of Judicial Conduct*, 9 SAN DIEGO L. REV. 785, 802 (1972).

[207] *See* 21A CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE AND PROCEDURE § 5052 (3d ed. 2012).

[208] *See* JOHN MACARTHUR MAGUIRE, EVIDENCE: COMMON SENSE AND COMMON LAW 218 (1947); Charles V. Laughlin, *Preliminary Questions of Fact: A New Theory*, 31 WASH. & LEE L. REV. 285, 302 (1974); John MacArthur Maguire & Charles S.S. Epstein, *Preliminary Questions of Fact in Determining the Admissibility of Evidence*, 40 HARV. L. REV. 392, 397 (1927).

[209] Imwinkelried, *supra* note 203, at 10; *see, e.g.*, Patton v. Bank of La Fayette, 53 S.E. 664 (Ga. 1906); Winslow v. Bailey, 16 Me. 319 (1839).

[210] Imwinkelried, *supra* note 203, at 10–11.

Fundamentally, allocating the power to make the fact determination to the jury was a fair tradeoff, because it was unlikely to "imperil the integrity of their later deliberations in the case."[211] The assumption underlying this shift to the jury was the faith that even if the jury decided that the witness did not observe the accident, common sense would "lead the jury to disregard the witness's testimony about the accident during the balance of their deliberations."[212]

> Thus, even if the jury was exposed to the foundational testimony and the foundational fact turned out to be false, the exposure w[ould] not distort the jury's deliberations about the merits of the case.
>
> Based on the same reasoning, the theory assigns the jury the power to decide whether an exhibit such as a letter is authentic. If the jury determines that the letter is a forgery, once again they should naturally disregard the letter's contents during their deliberations. If the plaintiff proffers the letter as an admission by the defendant but the jury finds that the defendant did not author the letter, it will be evident that they should attach no weight to the letter.
>
> These issues are usually designated "conditional relevance" questions. In an elementary sense, these facts condition the logical relevance of the evidence. If the witness is called to testify about an accident but lacks firsthand knowledge, the jury will naturally dismiss the witness's testimony as worthless. Similarly, if the prosecution claims that the defendant mailed a threatening letter to a witness but the jury concludes that the defendant did not write the letter, the jurors will probably put the letter aside during their deliberations.[213]

This theory underpins Federal Rule of Evidence 104(a) and (b) and its application. And it appears in Rule 901, which governs the admission of deepfake evidence.

b. Proving Authenticity Under Federal Rule of Evidence 901

Although Rule 901(a) is nonspecific in its mandate to prove that evidence is genuine, Rule 901(b) provides a variety of means through which a party can satisfy Rule 901(a). The text of Rule 901(b) provides a list of examples of proper authentication,[214] such as the testimony of a witness with knowledge[215] or the

[211] *Id.* at 11.

[212] *Id.*

[213] *Id. See* FED. R. EVID. 104(b) advisory committee's note; *see also* FED. R. EVID. 602 advisory committee's note; FED. R. EVID. 901 advisory committee's note.

[214] *See* FED. R. EVID. 901(b) (listing "examples only—not a complete list—of evidence that satisfies the requirement"); *see also* Tienda v. State, 358 S.W.3d 633, 640–41 (Tex. Crim. App. 2012) (discussing various modes of authentication used by courts, such as the creator admitting to authorship, witness testimony, business records, contextual or circumstantial information, and the reply doctrine).

[215] *See* FED. R. EVID. 901(b)(1) (providing that the testimony of a witness with knowledge is "[t]estimony that an item is what it is claimed to be"). A witness with knowledge could include someone who saw a document being signed or can provide testimony regarding the custody of an object from seizure to trial, commonly referred to as the "chain of custody." *See* FED. R. EVID. 901(b) advisory committee's note.

distinctive characteristics of the evidence, like "the appearance, contents, substance, internal patterns, or other distinctive characteristics of the item, taken together with all the circumstances,"[216] which show that the evidence is what the proponent claims. Rule 901(b) is non-exhaustive and intentionally broad;[217] it also offers examples of authenticating specific forms of evidence, including handwriting,[218] a voice,[219] and telephone communication.[220]

Under Rule 901(b)(9), digital evidence can be authenticated with evidence of a process or system that "produces an accurate result."[221] 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.[222] To authenticate voice audio, Rule 901(b)(5) requires "[a]n opinion identifying a person's voice—whether heard firsthand or through mechanical or electronic transmission or recording—based on hearing the voice

[216] *See* FED. R. EVID. 901(b)(4) (providing that distinctive characteristics include "appearance, contents, substance, internal patterns, or other distinctive characteristics of the item, taken together with all the circumstances"). The circumstantial evidence and distinctive features of an item provide "authentication techniques in great variety," including uniquely known facts to identify a speaker, contents of a letter that indicate it was a reply to an authenticated letter, or even language patterns. FED. R. EVID. 901(b) advisory committee's note.

[217] *See* FED. R. EVID. 901(b) advisory committee's note ("The examples are not intended as an exclusive enumeration of allowable methods but are meant to guide and suggest, leaving room for growth and development in this area of the law.").

[218] Handwriting can be authenticated through "[a] nonexpert's opinion that the handwriting is genuine," or through a comparison with an authenticated handwriting specimen by an expert witness or the factfinder. FED. R. EVID. 901(b)(2)–(3). The authentication of handwriting in subsection (2) requires a layperson's prelitigation familiarity, while subsection (3) requires that an authenticated sample or "exemplar" be available for expert comparison or comparison by a trier of fact. FED. R. EVID. 901(b) advisory committee's note.

[219] *See* FED. R. EVID. 901(b)(5) (providing for voice identification "based on hearing the voice at any time under circumstances that connect it with the alleged speaker").

[220] *See* FED. R. EVID. 901(b)(6) (providing that authenticating a telephone conversation may be made with "evidence that a call was made to the number assigned at the time," either to a certain person "if circumstances, including self-identification, show that the person answering was the one called," or to a certain business, "if the call was made to a business and the call related to business reasonably transacted over the telephone").

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

[222] As discussed, evidence may be authenticated by witness testimony, and proponents of proffered evidence may use an expert witness pursuant to Rule 702, which provides that "a witness qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if . . . the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue." FED. R. EVID. 702. An expert must provide a detailed description of the steps taken throughout the digital media forensics process, what was uncovered, and the conclusions reached. *See* Whole Woman's Health v. Hellerstedt, 579 U.S. 582, 620 (2016) ("[Rule 702 states that] an expert may testify in the 'form of an opinion' as long as that opinion rests upon 'sufficient facts or data' and 'reliable principles and methods.'" (quoting FED. R. EVID. 702)). In *Whole Woman's Health*, the doctor's opinion "rested upon his participation, along with other university researchers, in research that tracked the number of open facilities providing abortion care in the state." *Id.* at 620 (internal quotation marks omitted). The Court determined that "[t]he District Court acted within its legal authority in determining that [the doctor's] testimony was admissible." *Id.* at 621 (citing FED. R. EVID. 702); *see also* United States v. Espinal-Almeida, 699 F.3d 588, 610–14 (1st Cir. 2012) (affirming the trial court's authentication under Rule 901(b)(9) and admission of the computerized reproduction of the defendant's boat route at the time of the prosecuted drug transactions using the GPS device seized from the defendant's boat).

at any time under circumstances that connect it with the alleged speaker."[223] 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.[224] However, the testimony of someone who accessed content from the internet is generally insufficient to attribute content to a particular user without "personal knowledge of who maintains the website, who authored the documents, or the accuracy of their contents."[225]

Additionally, to address the challenges presented by the authentication of electronic evidence, 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 Rule 902 addressed electronically-stored information 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.[226] 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.[227] Rule 902(13) allows electronically-stored information to be authenticated without a witness at the stand to state what is supposedly obvious and unlikely to be challenged.[228] For instance, a party could establish how iPhone software captures the date, time, and GPS coordinates of each picture taken with an iPhone, permitting the court to quickly and conclusively determine that whoever took the picture did so at a particular time and from a particular place.[229] Such evidence is self-authenticating, or "allow[s] authentication of electronic information that would otherwise be established by a witness."[230] Rule 902(14) allows authentication of "[d]ata 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."[231] Under Rule 902(14), if proponents of electronically-stored information can extract a "hash value"—a unique numerical identifier that functions like a digital fingerprint—then the evidence

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

[224] *See* United States v. Salcido, 506 F.3d 729, 733 (9th Cir. 2007) (holding that the government properly authenticated child pornography taken from the defendant's computer by presenting detailed evidence as to the chain of custody, specifically how the images were retrieved from the defendant's computers).

[225] Wady v. Provident Life & Accident Ins. Co. of Am., 216 F. Supp. 2d 1060, 1064 (C.D. Cal. 2002).

[226] FED. R. EVID. 902(13)–(14).

[227] FED. R. EVID. 902(13).

[228] John M. Facciola & Lindsey Barrett, *Law of the Foal: Careful Steps Towards Digital Competence in Proposed Rules 902(13) and 902(14)*, 1 GEO. L. TECH. REV. 6, 10–11 (2016).

[229] *Id.* at 10.

[230] COMM. ON RULES OF PRAC. & PROC., REPORT OF THE ADVISORY COMMITTEE ON EVIDENCE RULES (Comm. Print 2015), http://www.uscourts.gov/rules-policies/archives/agenda-books/committee-rules-practice-and-procedure-may-2015 (proposed amendments).

[231] FED. R. EVID. 902(14).

is self-authenticating and can be admitted without corroborating witness testimony.[232]

2. *Common-Law Legal Theories for Authenticating Audiovisual Evidence*

Concomitant with Rule 901(b)'s various means of authenticating evidence, courts typically admit photographic evidence under one of two theories: the "pictorial communication" theory and the "silent witness" theory.[233] Each theory utilizes a different subsection of Rule 901(b) to meet Rule 901(a)'s sufficient evidence standard for authentication.[234]

The logic behind distinct foundational standards for the pictorial communication and silent witness theories hinges on the intended purpose of substantive, rather than demonstrative or illustrative, evidence. Substantive evidence provides "independent probative value for proving a fact," such as a physical object recovered from a scene relevant to the case.[235] Demonstrative or illustrative evidence, on the other hand, accompanies witness testimony and is intended to "aid the trier [of fact] in understanding the witness's testimony."[236] The distinction is essential but problematic in the context of photographs and videos, because illustrative evidence often becomes substantive by showing the jury more than the witness can recollect or convey, thereby introducing independent substantive evidence for which there is no foundation.[237] Nonetheless, the pictorial communication and silent witness theories derive their separate standards from the supposition that illustrative evidence is limited to the perceptions and recollections of the witness's testimony.[238]

a. *Pictorial Communication Theory*

Rule 901(b)(1)—authentication through the testimony of a witness with knowledge—underpins the pictorial communication theory, also known as the pictorial testimony theory. Under this theory, audiovisual evidence is admissible only when a witness can testify before a jury that the evidence is a fair and accurate representation or depiction of what occurred.[239] The pictorial communication theory rests on the idea that "any photographic or video

[232] Facciola & Barrett, *supra* note 228, at 12.

[233] BROUN, *supra* note 135, § 215.

[234] *Id.*

[235] *Id.* § 212.

[236] *Id.*

[237] *Id.*

[238] *Id.* § 215.

[239] *See* Fisher v. State, 643 S.W.2d 571, 573 (Ark. Ct. App. 1982) (holding that videotape evidence could not be admitted without a witness under the pictorial evidence theory absent a witness to verify the events in the videotape and that the video was not tampered with before trial); *Ex parte* Fuller, 620 So. 2d 675, 679 (Ala. 1993) (providing an example of questioning that appropriately admits visual evidence under the pictorial evidence theory); *see also* Danielle C. Breen, *Silent No More: How Deepfakes Will Force Courts To Reconsider Video Admission Standards*, 21 J. HIGH TECH. L. 122, 126–27 (2021).

evidence is just a 'graphic portrayal of oral testimony,' and therefore must be verified as correct by a witness."[240] Thus, all this theory requires is that the witness possess personal knowledge of the subject matter to reliably confirm that the events presented in the evidence are "fair and accurate representations."[241] Because Rule 901(b)(1) does not specifically define "knowledge," courts may look to other rules, such as Rule 602, which requires witnesses to have personal knowledge of the matters about which they testify, based on their sensory perceptions.[242]

Under this theory, the witness does not need to be present when the evidence was created,[243] and there is also no requirement that the witness be an expert in photography or videography.[244] Instead, the witness only needs to have personal knowledge of the subject material to confirm that the presented events are authentic.[245] The classic example of the pictorial evidence theory is a medical examiner testifying before the jury during a murder trial about the nature of a victim's wound, as shown through autopsy photographs.[246]

The fair and accurate portrayal standard assumes that video is difficult to alter—the standard is rooted in an age of traditional film photography before the advent of digital photography and other media.[247] Traditional photography differs from digital media in several ways.[248] The most relevant difference is that digital media stores individual pixels as data in an electronic file; there is no

[240] Breen, *supra* note 239, at 126.

[241] *See* 16 AM. JUR. 3D § 5 (2019) (outlining the pictorial evidence theory).

[242] *See* 31 CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE AND PROCEDURE § 7106 (3d ed. 2012) ("The fact that Rule 901(b)(1) uses the word 'knowledge' without restrictions or modifiers suggests that authentication testimony may be based on knowledge of the sort described by either Rule 602 or Rule 702."). To meet Rule 602's personal knowledge elements in order to testify as to whether photographic evidence is a fair and accurate portrayal, the witness must base their fair and accurate portrayal judgment on the direct use of their own senses, must have comprehended what they perceived at the time as well as at the time of their testimony, and must have a recollection of that prior perception. FED. R. EVID. 602; *see also* 2 JOHN HENRY WIGMORE, EVIDENCE IN TRIALS AT COMMON LAW § 478 (James H. Chadbourn ed., 1979) (outlining observation or perception, recollection, and communication as requirements for testimonial assertions).

[243] *See* Benjamin V. Madison II, *Seeing Can Be Deceiving: Photographic Evidence in a Visual Age—How Much Weight Does It Deserve*?, 25 WM. & MARY L. REV. 705, 708 (1984) (requirements of the pictorial evidence theory). The pictorial evidence theory is used with fingerprint and other evidence that is not meaningful to an "untrained eye" without explanation. *Id.* The pictorial evidence theory is also commonly used to depict conditions described by a witness, such as how far away something was at the time of an accident. *Id.* at 710.

[244] *See* 16 AM. JUR. 3D § 5 (2019) (reiterating that the witness need only have sufficient personal knowledge under the pictorial evidence theory).

[245] *See id.* (maintaining that a witness may have any background so long as they have personal knowledge of the events).

[246] *See* Madison, *supra* note 243, at 709–10 (observing how medical examiners are frequently required to provide context in murder cases because images of wounds by themselves cannot be fully understood as accurate). "Photographic displays allow an examiner to illustrate wounds that are difficult to conceptualize, such as numerous stab wounds, multiple bruises, or extensive damage resulting from a gunshot wound." *Id.* at 710.

[247] Witkowski, *supra* note 114, at 282 n.65.

[248] *See id.* at 269–71 (outlining the digital image creation process in scientific detail, specifically image compression and physical characteristics).

traditional original image such as with, for example, older thirty-five-millimeter film cameras.[249]

Additionally, because early digital photography featured lower initial image quality than film photography, its proponents commonly needed to enhance digital photographs to aid the trier of fact.[250] Thus, cases have addressed the issue of non-insidious modifications of video, such as editing, enhancing, taping over, or curating certain portions of a longer video or recording.[251] In these types of cases, courts envision having the "original" recording to reference against;[252] but courts rarely consider the possibility of outright forgery when considering authentication standards for admitting photographic evidence.[253] The rare cases when courts reject photographic evidence are where there is no authenticating witness, or the witness expressly rejects that the photograph is an accurate depiction.[254] Such was the situation in *United States v. Lawson*,[255] where the court concluded that the photographs the defendant offered were properly excluded from evidence because the only witness at trial testified that the photographs "did not accurately reflect what he saw."[256]

The fair and accurate portrayal standard is not a difficult hurdle to clear. The standard to establish the foundation is so minimal that issues concerning whether the witness's fair and accurate testimony is "limited" or "defective," or whether the witness is "otherwise unsure of his perceptions," are matters for the

[249] *Id.* at 272–73. Traditional film cameras capture light data as imprinted onto physical film, which can then be protected through a secure chain of custody. *Id.* at 268 n.3, 272. Digital photography, however, as a "finite set of ones and zeroes," makes determining whether a digital photograph is an original or a copy nearly impossible. *Id.* at 272. *But see* Facciola & Barrett, *supra* note 228, at 11–12 (explaining how iPhone software captures the date, time, and GPS coordinates of pictures as metadata while acknowledging the possibility that such metadata can be altered); CHING-YUNG LIN & SHIH-FU CHANG, GENERATING ROBUST DIGITAL SIGNATURE FOR IMAGE/VIDEO AUTHENTICATION (1998) (suggesting the possibility of "digital signatures" to ensure image security).

[250] Witkowski, *supra* note 114, at 269 n.6, 271 n.16. "In general, both traditional photographs and digital images often need to be enhanced. Enhancing an image involves adjusting the contrast so that the picture is clearer." *Id.* at 271 n.17.

[251] *See, e.g.*, United States v. Seifert, 445 F.3d 1043, 1045–46 (8th Cir. 2006) (admitting a digitally enhanced surveillance tape after an expert video analyst's testimony about each step of the digital enhancement process); United States v. Mills, 194 F.3d 1108, 1111–12 (10th Cir. 1999) (admitting an incomplete videotape where an officer responsible for filming testified as to the authenticity of the tape and confirmed that, "except for the deleted portion, it accurately depicted the entire episode"). In these commonplace instances, courts have required no more than satisfaction of the fair and accurate portrayal standard—or the "evidence as a process or system" standard if admitted under the silent witness theory—to admit the recording. Witkowski, *supra* note 114, at 279.

[252] Witkowski, *supra* note 114, at 272.

[253] *Id.* at 285–86 (considering various reasons for the "infrequency of challenges to digital images," including a general lack of awareness and a focus on editing, not forgery).

[254] *See, e.g.*, United States v. Lawson, 494 F.3d 1046, 1052 (D.C. Cir. 2007) (determining that the trial court properly excluded photographs from evidence because they were not authenticated by the only witness familiar with the scene).

[255] *Id.*

[256] *Id.*; *see also* United States v. 320.0 Acres of Land, More or Less, 605 F.2d 762, 826 (5th Cir. 1979) (holding that the district court erred in admitting evidence of a movie under the pictorial evidence theory because a witness testified contrary to the plaintiff's allegations that the land was "swamp, muck, and water").

jury to assign weight to evaluate the evidence's credibility—not matters of admissibility with which the proponent of the evidence must grapple.[257] Thus, the standard imposes only a "sufficient to support a finding" requirement on the proponent.[258]

For example, in *United States v. Ray*, the defendant argued that photographs admitted into evidence were prejudicial, because they were used to prove the defendant's "general atmosphere" of enticing minors into sexual activity, and not the elements of the offense itself.[259] However, the Sixth Circuit upheld the photographs' admission, because the corroborating testimony and pictorial evidence accompanying the photographs demonstrated how the defendant enticed minors into sexual activity.[260]

b. Silent Witness Theory

In contrast with the pictorial evidence theory, the silent witness theory rooted in Rule 901(b)(9)[261] admits visual evidence without a qualifying witness.[262] Instead, a judge deems whether there is a sufficient foundation to admit the evidence absent a witness testifying before the jury, based on the court's faith in the reliability of the process by which the evidence was recorded or created.[263] The theory was initially proposed to allow x-ray images and surveillance videos into evidence.[264] Under this theory, evidence is admissible at the trial court judge's discretion upon a showing that the video was created under reliable processes and untampered with between the time it was taken and presented to the court.[265] Judges admit visual evidence under the silent witness

[257] WRIGHT ET AL., *supra* note 242.

[258] *Id.*

[259] United States v. Ray, 189 F. App'x 436, 444 (6th Cir. 2006).

[260] *Id.* at 445. The pictorial evidence theory has also been upheld in criminal proceedings. In *United States v. Reichart*, the U.S. Army Court found that videotape evidence of security surveillance was properly admitted under the pictorial evidence theory when the store security employee testified that she was familiar with the store location, that she observed the defendant stealing through the computer monitor, and that the video was an accurate representation of what she observed through the computer monitor at the time it occurred. 31 M.J. 521, 524 (A.C.M.R. 1990); *see also* United States v. Richendollar, 22 M.J. 231, 232 (C.M.A. 1986) (finding admissible photographs of the defendant, accused of engaging in inappropriate conduct with a minor, with the victim's friend when a clerk testified that the photo depicted was a scene that she was familiar with and was an accurate representation of that scene); United States v. Slaughter, No. 3:18-cr-00027, 2020 WL 1685117, at *4 (S.D. Tex. Apr. 6, 2020) ("[C]hild pornography is nothing more than pictorial evidence of crimes against children." (quoting United States v. Davin, No. 12-10141, 2012 WL 2359419, at *3 (D. Kan. June 20, 2012))); United States v. Nickelson, No. 18-mj-102, 2018 WL 4964506, at *4 (D.C. Cir. Oct. 15, 2018).

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

[262] *See generally* Tracy Bateman Farrell, Annotation, *Construction and Application of Silent Witness Theory*, 116 A.L.R.5th 373 (2019).

[263] *Id.*

[264] *See* 16 AM. JUR. 3D § 5 (2019) (outlining the adoption of the silent witness theory).

[265] *Id.* §§ 5–6 (outlining policy considerations behind the silent witness theory). If the process behind the creation of the video is deemed inherently reliable by the judge, the evidence may "speak for itself." *Id.* § 5. *See* Madison, *supra* note 243, at 711 (discussing why courts apply the silent witness theory). Courts are generally reluctant to limit the use of photographic evidence, which is a large underlying policy reason behind the application of the silent witness theory. *Id.*

theory as a trusted substitute for a qualifying witness's account of what happened; in other words, the process through which the evidence was obtained renders the evidence sufficiently reliable for admission.[266]

This theory represents the inherent trust society has placed in video evidence, because it demonstrates the belief that videos are a non-biased account of events—that seeing is believing.[267] The fact that most jurisdictions apply the silent witness theory further underscores the value of video evidence today.[268] Automated camera evidence, such as security footage, is often subject to presumed authentication under the silent witness theory. For example, in *United States v. Taylor*, although no one could testify about the events shown in the video footage, the court admitted the automated video evidence that showed a bank robbery in progress after the robbers locked employees in a bank vault.[269] Instead, witnesses testified about "the manner in which the film was installed in the camera, how the camera was activated, the fact that the film was removed immediately after the robbery, the chain of its possession, and the fact that it was properly developed and contact prints were made from it."[270] Because there was no suggestion that the video had been altered, the court deemed it admissible.[271] State courts have admitted videotape footage on similar grounds.[272]

However, as modern photo and video editing technology become more advanced, the silent witness theory invites error, because not all judges or lawyers are familiar enough with this technology or the processes through which these altered forms of evidence are created to correctly evaluate their

[266] *See Ex parte* Fuller, 620 So. 2d 675, 678 (Ala. 1993) (explaining why the silent witness theory allows evidence to be admitted even absent a witness). "[T]he process or mechanism substitutes for the witness's senses, and because the process or mechanism is explained before the photograph, etc., is admitted, the trust placed in its truthfulness comes from the proposition that, had a witness been there, the witness would have sensed what the photograph, etc., records." *Id. See* Madison, *supra* note 243, at 710–11 (articulating the weight evidence is given after admission under the silent witness theory). "In practical terms, such photographic evidence assumes greater significance than photographic evidence authenticated by testimony. Instead of supplementing testimony on an issue, the photographic evidence forms an independent basis upon which the proponent may establish a fact or occurrence." *Id.* at 711.

[267] For example, prosecutors striking blind persons from the jury demonstrates the idea that one must be able to see in order to fully comprehend all of the evidence in a case. United States v. Watson, 483 F.3d 828, 828 (D.C. Cir. 2007); *see also* Granot et al., *supra* note 98 (arguing that visual evidence is so highly regarded in the justice system that it is seen as a way to mitigate juror bias toward other pieces of evidence in a case).

[268] *See* Farrell, *supra* note 262 (asserting that while most jurisdictions have not expressly adopted the silent witness theory, very few have explicitly rejected it).

[269] United States v. Taylor, 530 F.2d 639, 641–62 (5th Cir. 1976).

[270] *Id.* at 642.

[271] *See id.*; *see also* United States v. Harris, 55 M.J. 433, 440 (C.A.A.F. 2001) (finding that videotape footage was properly authenticated because the chain of custody posed no evidence that any alteration to the videotape was made); United States v. Marshall, 332 F.3d 254, 263 (4th Cir. 2003) (holding that videotape evidence of a defendant's conspiracy to sell and offer drugs was admissible because the government introduced sufficient evidence establishing the videotape footage's reliability, thereby properly authenticating it).

[272] *See, e.g.*, Dolan v. State, 743 So. 2d 544, 546 (Fla. Dist. Ct. App. 1999) (admitting a videotape under the silent witness theory where the government provided testimony establishing the location and operation of the videotaping mechanisms).

authenticity.[273] As the next Subpart argues, given the speed of technological advances in the field of AI and deepfakes and the disruption to the truth-finding process they pose, the incremental approach to the issues of admissibility found in the common law is not an effective way to deal with deepfakes.

C. The Rules of Evidence and Common-Law Evidence Theories Are Inadequate To Address Deepfakes.

Standing alone, none of the Federal Rules of Evidence or their companion common-law theories are sufficient to address the significant challenges that deepfakes present, discussed in Part I—namely, problems with authenticating evidence, responding to the deepfake defense, and addressing juror skepticism.

The rapid advances in deepfake technology and the problems in detecting fake audio and video, coupled with the suggestibility of human perception, highlight the shortcomings of the current tools and legal theories grounded in the Rules for dealing with deepfakes. Recognition of the inadequacy of current evidentiary standards is not new, however. For almost thirty years, scholars have expressed that evidentiary standards are inadequate to address advances in digital photography.[274] Except for the 2017 amendments to Rule 902, few changes have been made to the authentication standards for electronic and digital evidence.[275] As discussed in the last Subpart, the threshold for making a prima facie showing of authenticity is not high for audio, photographic, or video evidence, and the proponent's burden is slight.[276] Historically, any negative impact of such a low bar has been mitigated by courts' reliance on expert witnesses to assist with authenticity determinations,[277] and because it was difficult to create high-quality fake audiovisual images at the time.[278] However, the deepfakes era brings the deficiency of relying on these standards to the forefront. As this Subpart explains, the proliferation of deepfake technology renders obsolete the assumptions upon which the pictorial communication theory relies, as witnesses will struggle to meet the personal knowledge standard required to authenticate video evidence. The advances in technology also render the silent witness theory unworkable. And in those relatively few cases where deepfake evidence has appeared, courts have struggled with applying the current Rules. Finally, this Subpart explores the exceptional and unprecedented

[273] *See* Grimm et al., *supra* note 77, at 88–95 (discussing the expert evidence that must be presented so that lawyers and judges gain sufficient familiarity to make authenticity evaluations).

[274] *See* Witkowski, *supra* note 114, at 285–87 (arguing in 2002 that the standard for admitting digital images was insufficient); *see also* Sharon Panian, *Truth, Lies, and Videotape: Are Current Federal Rules of Evidence Adequate?*, 21 Sw. U. L. Rev. 1199, 1205–14 (1992) (highlighting common distortion problems with misleading computer graphics and edited videotapes).

[275] *See, e.g.*, Owens v. State, 363 Ark. 413, 421 (2005) (refusing to alter the standard for digital photographs).

[276] *See supra* Part II.A.2.

[277] *See* Fed. R. Evid. 901(b)(3) (recognizing that courts can rely on a comparison with an authenticated specimen by an expert witness).

[278] *See* Witkowski, *supra* note 114, at 285–87.

challenge that deepfake evidence presents for jurors, given its technical and legal complexity.

1. Common-Law Theories of Admissibility No Longer Work, and the Existing Rules of Evidence Are Ineffective To Address Deepfake Evidence.

First, the pictorial communication theory may no longer be viable in the era of deepfakes because witnesses now lack sufficient personal knowledge to authenticate an image as a fair and accurate depiction of events. A deepfake image or video may be so technically sophisticated, so close in likeness, and appear so accurate that no witness can perceive the alterations or fabrications. Moreover, a witness may no longer be able to determine whether the video's depiction is a fair and accurate portrayal of their memory. Deepfakes' lifelike appearance reduces the likelihood that authentication witnesses will reliably declare either that something looks different from the way they remember it, or that they do not recall the event at all; the visuals are too convincing and too likely to take advantage of the suggestibility flaw inherent in human memories.[279] Thus, deepfakes vastly increase the likelihood that authenticating witnesses will not identify material changes from the actual scene that the video depicts.[280]

Moreover, a fake video is more likely to corrupt an authenticating witness's memories, leading the witness to recall the falsehoods that the video depicts.[281] The authenticating witness's struggle to detect alterations from what they actually observed and the possibility of false memories may lead to a complete inability to attest that a video reflects a fair and accurate depiction of the events that transpired.[282] As a result, the court may present fraudulent substantive evidence to a jury that a witness without proper personal knowledge has authenticated. Thus, even though problems with the pictorial communication theory did not emerge with the invention of deepfakes, these new fake audiovisuals critically reduce the effectiveness of authentication witnesses.

Second, the standard for admitting photographic evidence under the silent witness theory—that is, without an accompanying witness—poses its own set of problems when it comes to deepfakes. As discussed in the previous Subpart, the silent witness theory is premised on the belief that video is a reliable and trustworthy representation of reality and presents a nonbiased account of

[279] *See* Mark W. Bennett, *Unspringing the Witness Memory and Demeanor Trap: What Every Judge and Juror Needs To Know About Cognitive Psychology and Witness Credibility*, 64 AM. U. L. REV. 1331, 1335–37, 1352 (2015) (examining a host of challenges to accurate witness testimony and proposing a "Model Plain English Witness Credibility Instruction"); Wade et al., *supra* note 102 (using fabricated video in a psychological study to demonstrate witness suggestibility).

[280] Wade et al., *supra* note 102.

[281] *Id*. (using fabricated video in a psychological study to demonstrate witness suggestibility); Bennett, *supra* note 279, at 1357; Leggett, *supra* note 102.

[282] Wade et al., *supra* note 102, at 904–05.

events.[283] Deepfakes—doctored images designed to trick the viewer into seeing and hearing something that appears to be real but is not—upends this faith underlying the silent witness theory. In the era of deepfakes, audiovisual images can no longer serve as a trusted substitute for a qualifying witness's account of what happened. In addition, the silent witness theory is premised on the idea that audiovisual images were created under reliable processes and untampered with between the time they were taken and presented to the court.[284] But that idea does not hold for deepfakes. Although jurors and judges may have a general awareness that deepfakes exist, understanding the processes by which digital audiovisual images, fake or real, are created is well beyond the knowledge of most judges, jurors, and lawyers.[285] The process of creating deepfake images and their underlying AI technology is sufficiently complicated such that only a handful of technology experts and digital forensics experts fully grasp its operations.[286] This complexity makes it impossible for judges to apply the silent witness theory to determine whether the evidence obtained is undoctored and sufficiently reliable for admission. At bottom, the silent witness theory is unworkable for deepfakes, because the judicial system cannot place confidence in the reliability of a photographic process that it cannot understand without the assistance of experts.

Moreover, the recent addition of Rule 902(13) and (14) does not solve the authenticity problem with deepfakes. These 2017 amendments to Federal Rule of Evidence 902 "were designed to simplify the legal process and reduce the costs associated with using electronically-stored information as evidence"[287] by

[283] *See Ex parte* Fuller, 620 So. 2d 675, 678 (Ala. 1993) (explaining why the silent witness theory allows evidence to be admitted even absent a witness); *see also* Madison, *supra* note 243, at 710–11.

[284] *See* 16 AM. JUR. 3D § 5 (2019) (outlining policy considerations for using the silent witness theory); *see also* Madison, *supra* note 243, at 711 (discussing why courts apply the silent witness theory).

[285] *See* Melissa Whitney, *How To Improve Technical Expertise for Judges in AI-Related Litigation*, BROOKINGS INST. (Nov. 7, 2019), https://www.brookings.edu/research/how-to-improve-technical-expertise-for-judges-in-ai-related-litigation/ (suggesting a need to have technical advisers educate judges on technology and AI-related issues in litigation to ensure that they properly consider the evidence); Herbert B. Dixon Jr., *Deepfakes: More Frightening Than Photoshop on Steroids*, AM. BAR ASS'N (Aug. 12, 2019), https://www.americanbar.org/groups/judicial/publications/judges_journal/2019/summer/deepfakes-more-frightening-photoshop-steroids/ (cautioning the challenges deepfakes will bring to the courtroom when parties present conflicting testimony about the authenticity of a video); Debra Cassens Weiss, *Should There Be a Duty of Tech Competence for Judges? Survey Raises Questions*, AM. BAR ASS'N J. (May 10, 2019, 4:13 PM), https://www.abajournal.com/news/article/should-there-be-a-duty-of-tech-competence-for-judges-survey-raises-questions (quoting a 2019 survey where two-thirds of judges stated that they need more e-discovery training); Riana Pfefferkorn, '*Deepfakes': A New Challenge for Trial Courts*, NWSIDEBAR (Mar. 13, 2019), https://nwsidebar.wsba.org/2019/03/13/deepfakes-a-new-challenge-for-trial-courts/ (warning how trial courts will need to become apt at confronting deepfakes).

[286] *See* Brown, *supra* note 78, at 25–26 (describing the efforts of a small universe of digital forensics experts and researchers working in the federal government and private sector on deepfake detection technologies); *see also* Deborah G. Johnson & Nicholas Diakopoulos, *Computing Ethics: What To Do About Deepfakes*, COMM'NS ACM, Mar. 2021, at 33–35 (discussing the role of expertise in addressing deepfakes).

[287] WITNESS MEDIA LAB, TICKS OR IT DIDN'T HAPPEN: CONFRONTING KEY DILEMMAS IN AUTHENTICITY INFRASTRUCTURE FOR MULTIMEDIA 30 (2019), https://lab.witness.org/ticks-or-it-didnt-happen/ ("The idea is

providing for the self-authentication of evidence generated by "an electronic process or system that produces an accurate result,"[288] or evidence copied from an electronic device if "authenticated by a process of digital identification."[289] These amendments anticipated that video and audio technologies would be developed to embed or upload authenticating data at the time of capture on the audiovisual content.[290] By attaching additional metadata the moment the video is taken, such "verified media capture technology" helps "ensure that the evidence [users] are recording . . . is trusted and admissible to courts of law."[291] Although self-authenticating technology may soon become standardized for body cameras, surveillance cameras, and video cameras used for recording depositions and interrogations, the technology is not yet widely used.[292] And until it is, newly amended Rule 902 is of limited assistance in sorting the current generation of deepfakes from genuine audiovisual evidence.

Neither courts nor litigants can reliably depend on one theory of evidence or Rule to demonstrate the authenticity of audiovisual evidence. Under the current regime, parties must seek out corroborating witnesses to back up what is depicted in the evidence, in addition to attaining digital forensics experts who can vouch for (or assail) the authenticity of the footage. Additionally, because the field is new, technical questions emerge as to whether there are experts available to testify on the matter.[293] Authenticating videos will run up costs through heightened due diligence, additional motion practice, increased expenditures on lay and expert witnesses, and longer hours in the courtroom.[294]

that if you cannot detect deepfakes, you can, instead, authenticate images, videos and audio recordings at their moment of capture.").

[288] FED. R. EVID. 902(13).

[289] FED. R. EVID. 902(14).

[290] *See* Tara Vassefi, *A Law You've Never Heard of Could Help Protect Us from Deceptive Photos and Videos*, MEDIUM: HUM. RTS. CTR. (Nov. 30, 2018), https://medium.com/humanrightscenter/a-law-youve-never-heard-of-could-help-protect-us-from-fake-photos-and-videos-df07119aaeec. This technology relies on generating "hashes," or cryptographic representations of data, on a public or private database called a blockchain. *Id.*; *see* WITNESS MEDIA LAB, *supra* note 287.

[291] WITNESS MEDIA LAB, *supra* note 287, at 22. For example, an app called eyeWitness "allows photos and videos to be captured with information that can firstly verify when and where the footage was taken, and secondly can confirm that the footage was not altered," all while the company's "transmission protocols and secure server system . . . create[] a chain of custody that allows this information to be presented in court." *Id.* at 27. "That information, paired with the app maker's willingness to provide a certification to the court or send a witness to testify if needed, could satisfy a court that the video is admissible, even if the videographer is unavailable." *Id.* at 29.

[292] *See* Pfefferkorn, *supra* note 14, at 268–69.

[293] *See id.* at 265.

[294] The costs of hiring a digital forensics expert can range from several thousand dollars to well over $100,000, with the typical analyses being somewhere in the $5,000 to $15,000 range, based on the factors involved. Betsy Mikalacki, *How Much Does Digital Forensic Services Cost?*, VESTIGE, https://www.vestigeltd.com/thought-leadership/digital-forensic-services-cost-guide-vestige-digital-investigations/#:~:text=In%20regard%20to%20digital%20fore (last visited Jan. 28. 2023); Edward J. Imwinkelried, *Impoverishing the Trier of Fact: Excluding the Proponent's Expert Testimony Due to the Opponent's Inability To Afford Rebuttal Evidence* 1–6 (U.C. Davis Legal Stud. Rsch., Working Paper No. 104, 2007), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=982209 (describing the costs of expert computer-generated evidence in the tens of

The mere existence of deepfake technology allows any party in an action to question whether audiovisual evidence is real or fake, further illuminating the Federal Rules' inadequacies in dealing with digital audiovisual evidence.[295]

2. The Current Allocation of Factfinding Responsibility Exacerbates the Problem.

The balance of decision-making authority—the split between the court and the jury—for determining the authenticity of a piece of evidence that is baked into the Rules and the common law amplifies the danger that deepfakes pose. As discussed earlier in this Part, the historical practice of resolving authenticity issues assigns the trial court the preliminary determination of the foundation fact of whether the item of evidence is authentic—that is, whether a reasonable jury could find that the item is what the proponent claims.[296] If that prima facie showing is made, the evidence is admitted, and the jury determines authenticity.[297] The allocation of decision-making on the question of authenticity under Rule 901 is derived from an interpretation of Rule 104(b),[298] which governs "preliminary questions" on the admissibility of evidence.[299]

This allocation of responsibility between the court and the jury on the issue of authenticity is troublesome in the era of deepfakes. Assigning the ultimate determination of the authenticity of audiovisual images to the jury presents two main challenges.

thousands of dollars and reflecting on the fact that it will be cost-prohibitive for some to pay for the evidence they need to litigate their claims).

[295] Imwinkelried, *supra* note 294.

[296] *See supra* Part II.A.4; *see also* 3 JOHN H. WIGMORE, EVIDENCE § 693 (3d ed. 1940); 7 JOHN H. WIGMORE, EVIDENCE § 2135 (3d ed. 1940); CHARLES T. MCCORMICK, EVIDENCE § 189 (1954). *See, e.g.*, United States v. Natale, 526 F.2d 1160, 1173 (2d Cir. 1975) (reflecting this common practice); United States v. King, 472 F.2d 1, 7 (9th Cir. 1972) (same); United States v. Addonizio, 451 F.2d 49, 69 (3d Cir. 1971) (same); United States v. Tellier, 255 F.2d 441, 448 (2d Cir. 1958) (same).

[297] *See supra* Part II.A.4.

[298] FED. R. EVID. 104 ("(a) In General. 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. (b) Relevance That Depends on a Fact. 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.").

[299] Zenith Radio Corp. v. Matsushita Elec. Indus. Co., 505 F. Supp. 1190, 1218–19 (E.D. Pa. 1980) ("Under the aegis of Rule 104(b), the judge makes a preliminary determination whether the foundation evidence is sufficient to permit a factfinder to conclude that the condition in question has been fulfilled. If so, according to the Advisory Committee Note: . . . the item is admitted. If after all of the evidence on the issue is in, pro and con, the jury could reasonably conclude that fulfillment of the condition is not established the issue is for them. . . . Once a prima facie showing has been made to the court that a document is what its proponent claims, it should be admitted. At that point the burden of going forward with respect to authentication shifts to the opponent to rebut the prima facie showing by presenting evidence to the trier of fact which would raise questions as to the genuineness of the document. The required prima facie showing of authentication need not consist of a preponderance of the evidence. Rather, all that is required is substantial evidence from which the trier of fact might conclude that a document is authentic." (internal citations and quotation marks omitted)), *aff'd in part, rev'd in part on other grounds*, 723 F.2d 238 (3d Cir. 1983), *rev'd on other grounds*, 475 U.S. 574 (1986).

First, jurors may struggle to determine whether audiovisual images are authentic because deepfakes may be convincing, compelling, and difficult to detect; some deepfakes go beyond a layperson's ability to visually ascertain whether an image or video accurately portrays what it purports to. Jurors may struggle to parse the real from the fake. Although few empirical studies on the human ability to detect deepfakes currently exist, one behavioral experiment study published in late 2021, conducted by researchers from the Center for Humans and Machines at the Max Planck Institute for Human Development and the University of Amsterdam School of Economics, found that the study participants could not reliably detect deepfakes even after they were taught about them.[300] The study demonstrates that people are biased toward mistaking deepfakes for authentic videos and overestimate their abilities to detect deepfakes.[301] Furthermore, researchers observed that participants adopted a "seeing is believing" heuristic for deepfake detection.[302] This combination renders people particularly susceptible to being influenced by deepfake content.[303] The study concludes that "detection of deepfakes is not a matter of lacking motivation but inability."[304]

Second, the ability to decide questions of authenticity may be infected by juror skepticism and bias. Some jurors may doubt unaltered content simply because they know realistic deepfakes are possible.[305] The flood of online information—real, fake, and in between—has made it more difficult for individuals to ascertain truth.[306] The proliferation of disinformation makes people question their ability to trust anything.[307] Social media disinformation has made it harder for people to distinguish truth from fiction online.[308] Thus, in addition to the risk of deepfakes being perceived as real, the knowledge that deepfakes exist undermines belief in the authenticity of undoctored images.[309]

[300] Köbis et al., *supra* note 91. In the experiment, more than 200 participants watched a series of deepfake and authentic videos and guessed which ones were deepfakes. *Id.* The researchers included experimental manipulations to increase people's motivation to detect deepfakes by providing a short prompt emphasizing the danger of deepfakes (Awareness Treatment) or by financially incentivizing accuracy (Financial Incentive Treatment). *Id.* They compared the interventions to a control treatment without motivational interventions (Control). *Id.* They further assessed participants' confidence in their guesses. *Id.*

[301] *Id.* at 8.

[302] *Id.* at 9.

[303] *Id.*

[304] *Id.*

[305] *See* Brown, *supra* note 78, at 25–26 (asserting that even if a realistic deepfake is publicly identified as fake, it is unclear that the public would believe it).

[306] *See generally* Julie A. Seaman, *Black Boxes*, 58 EMORY L.J. 427 (2008) (noting how the U.S. system clings to the jury's role as the judge of credibility despite technological advances); John Villasenor, *Artificial Intelligence, Deepfakes, and the Uncertain Future of Truth*, BROOKINGS (Feb. 14, 2019), https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/.

[307] *See* Villasenor, *supra* note 306 (noting the need for public awareness of deepfakes).

[308] *See* Janna Anderson & Lee Rainie, *The Future of Truth and Misinformation Online*, PEW RSCH. CTR. (Oct. 19, 2017), https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/ [https://perma.cc/VH2F-M23Q].

[309] *See* Brown, *supra* note 78, at 26.

In the age of deepfakes, jurors may expect the proponent of a video to use sophisticated technology to prove to their satisfaction that the video is not fake in every case.[310]

The challenge for jurors to make these authenticity assessments may be exacerbated when the opponent of an authentic video asserts the deepfake defense to try to exclude a genuine piece of evidence, or at least sow doubt in jurors' minds.[311] If juries, entrusted as the final arbiters of fact on the issue of authenticity, start to doubt whether it is possible to know what is real, their skepticism will undermine the justice system as a whole.

3. Courts Strain To Figure Out a Way Forward.

The judiciary may fare no better than juries, and may similarly struggle to determine the authenticity of audiovisual images in the age of audiovisual fakery. Because deepfake evidence is a new phenomenon, few published opinions address the struggle courts face in determining whether to admit such evidence under existing evidentiary frameworks. However, reviewing those cases indicates that courts acknowledge the problem they face.

For example, *People v. Smith* concerns the admissibility of Facebook postings that purportedly connected the defendant with gang activity.[312] The prosecutor used Facebook posts created by others to associate the defendant with his gang moniker. The defendant challenged the authenticity of the posts.[313] In assessing their admissibility under the state-law equivalent of Federal Rule of Evidence 901(a), the appellate court acknowledged that after the trial court makes the preliminary determination of the posts' authenticity for admissibility purposes, the jury is left to determine whether the evidence is authentic and the weight to assign to it.[314] In discussing the trial court's obligation under the preliminary stage of conditional authenticity, the court recognized that "while the showing at the first stage is not a particularly rigorous one, . . . in the age of fake social-media accounts, hacked accounts, and so-called deep fakes, a trial court faced with the question whether a social-media account is authentic must itself be mindful of these concerns."[315] The court further acknowledged the lack

[310] Commentators have called jurors' expectation of being presented with scientific evidence in every case the "*CSI* effect." *See e.g.*, Vikas Khanna & Scott Resnik, *The* CSI *Effect*, AM. BAR ASS'N LITIG. J., Jan. 7, 2021, https://www.americanbar.org/content/dam/aba/publications/litigation_journal/winter-2021/the-csi-effect.pdf; Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 349 (2012) (defining the *CSI* effect as the phenomenon "by which jurors in routine criminal cases expect prosecutors to introduce evidence collected using high-tech investigatory tools like those featured on popular television dramas such as *Law & Order* and *CSI*"); Tom R. Tyler, *Viewing* CSI *and the Threshold of Guilt: Managing Truth and Justice in Reality and Fiction*, 115 YALE L.J. 1050, 1052 (2006) (describing that the *CSI* effect occurs when "people who watch the series develop unrealistic expectations about the type of evidence typically available during trials, which, in turn, increases the likelihood that they will have a 'reasonable doubt' about a defendant's guilt").

[311] *See supra* Part I.D.2.b.

[312] 969 N.W.2d 548, 565 (Mich. Ct. App. 2021).

[313] *Id.* at 554.

[314] *Id.* at 565.

[315] *Id.*

of published opinions accessible to the trial court, which had to extend the reasoning from other cases involving social media posts directly authored by the defendants.[316] The *Smith* court's reference to the new "mindfulness" a trial court must acquire about the possibility of deepfakes and the lack of directly relevant case law underscores the present challenges deepfakes pose for courts, even at the preliminary stages of the admissibility determination.[317]

III. Proposed Solutions

The challenges deepfakes pose to the integrity of court proceedings are significant, but not insurmountable. Since mid-2018, the impact of deepfakes on the court system has received attention from scholars, law students, journalists, and practitioners.[318] Some are optimistic, believing that the Federal Rules of Evidence and practice norms are currently sufficient to address evidentiary issues concerning deepfakes.[319] These scholars recommend evaluating the authenticity of deepfake evidence under Rule 901, and have urged applying the *Daubert* factors in making those evaluations.[320] Others are less sanguine, recommending various amendments to the Rules, specifically Rule 901, to expressly address authenticating digital and audiovisual evidence[321] or require more circumstantial evidence to corroborate video evidence than courts currently require under Rule 901.[322] Additionally, some argue that courts should reexamine the common-law theories for authenticity and abandon the silent witness theory.[323] Still other commentators offer a middle-ground approach,

[316] *Id.*

[317] The same concerns operate in civil actions. In *Hatteberg v. Capital One Bank*, for instance, the court acknowledged that the existence of easily fabricated evidence such as deepfakes will make bringing cases under the heightened federal pleading standard more challenging, if not impossible. No. SA CV 19-1425, 2019 WL 8888087, at *1, *4 (C.D. Cal. Nov. 20, 2019).

[318] *See, e.g.*, Jonathan Mraunac, *The Future of Authenticating Audio and Video Evidence*, Law360 (July 26, 2018, 12:57 PM), https://www.law360.com/articles/1067033/the-future-of-authenticating-audio-and-video-evidence; Theodore F. Claypoole, *AI and Evidence: Let's Start To Worry*, The Nat'l L. Rev. (Nov. 14, 2019), https://www.natlawreview.com/article/ai-and-evidence-let-s-start-to-worry; Ashley Dean, *Deepfakes, Pose Detection, and the Death of "Seeing Is Believing,"* L. Tech. Today (Aug. 6, 2019), https://www.lawtechnologytoday.org/2020/08/deepfakes-pose-detection-and-the-death-of-seeing-is-believing/; Lehman et al., *supra* note 79; Maras & Alexandrou, *supra* note 93; Jason Tashea, *As Deepfakes Make It Harder To Discern Truth, Lawyers Can Be Gatekeepers*, Am. Bar. Ass'n J.: L. Scribbler (Feb. 26, 2019, 6:30 AM), http://www.abajournal.com/lawscribbler/article/as-deepfakes-make-it-harder-to-discern-truth-lawyers-can-be-gatekeepers; Pfefferkorn, *supra* note 14, at 246, 275–76; Carlson, *supra* note 142, at 1060–64; Breen, *supra* note 239, at 162; McPeak, *supra* note 75, at 448–50.

[319] Grimm et al., *supra* note 77, at 85; Pfefferkorn, *supra* note 14, at 248, 275–76 (arguing that existing rules and practices "are sufficient as-is to deal with deepfakes, and that raising the bar for authenticating video evidence would do more harm than good").

[320] *See* Grimm et al., *supra* note 77, at 95–97 (urging the use of the *Daubert* factors to determine the authenticity of AI evidence).

[321] Carlson, *supra* note 142, at 1060–64 (arguing for multiple amendments to Rule 901(b) to address diffuse categories of electronic and digital evidence).

[322] John P. LaMonaca, *A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes*, 69 Am. U. L. Rev. 1945, 1984–86 (2020).

[323] *See generally* Breen, *supra* note 239.

urging courts to redefine the quantity and quality of circumstantial evidence for a reasonable jury to determine the authenticity of audiovisual evidence under each common-law theory (pictorial evidence and silent witness).[324] Finally, some suggest providing juries with specific, detailed jury instructions on how to assess evidence to detect and authenticate deepfakes.[325]

The proposals and commentary recognize the significant challenges deepfake evidence poses for court proceedings, and are thus a good start to the conversation. The scholarship focuses exclusively on how the factfinder should evaluate whether deepfake evidence is authentic under the existing Rules, or theoretical approaches to authenticity underlying the Rules. However, there has been no commentary on the procedural facets of the problem. Overlooked in the discussion of deepfakes in legal proceedings are the questions of the procedure that should be used to determine the preliminary facts; that is, who, the judge or the jury, gets to decide whether a deepfake is authentic. As this Part explains, the challenges of deepfakes—challenges of proof, the deepfake defense, and juror skepticism—can be best addressed by amending the Rules for authenticating digital audiovisual evidence, instructing the jury on its use of that evidence, and limiting counsel's efforts to exploit the existence of deepfakes.

A. Solving the Challenge of Proof

As described in Part I, deepfakes present challenges in proving that digital audiovisual evidence is authentic and gathering the requisite expertise to make that showing. Proving that an image or audio is real (or not) may require multiple complicated, expensive proofs to corroborate the evidence at issue.[326] The expert analysis needed to test audiovisual evidence may also require more time to develop, and may be cost prohibitive for some litigants. The Rules must be amended to address these challenges.

During an evidentiary hearing considering the admissibility of such evidence, courts should use the existing, non-exhaustive means of authentication under Rule 901(b) to assess whether the deepfake evidence is authentic. As discussed in Part I, the history of the treatment of other scientific evidence and audiovisual evidence, including photographs, digital media, audio recordings, and social media, shows that the non-exhaustive means of evaluating

[324] McPeak, *supra* note 75, at 447–50.

[325] *See* Venema & Geradts, *supra* note 94. For videos, images, and audio evidence that need to be authenticated with circumstantial evidence, the court may opt to include a special jury instruction that explains some criteria that jurors can use to determine whether the evidence is what it purports to be. Aspects like lighting, blinking, and editing clues can be included as factors. Elizabeth Caldera, Comment, *"Reject the Evidence of Your Eyes and Ears"[1]: Deepfakes and the Law of Virtual Replicants*, 51 SETON HALL L. REV. 177, 189 (2019) (explaining some of the subtle ways deepfake content can be discerned, like watching for a lack of blinking by the subject, but noting that the technology will make it even harder to rely on visual clues in the future). Unfortunately, a jury instruction on ways to spot deepfakes may become obsolete as technology advances. But some sort of detailed guidance for the jury on gauging authenticity may be warranted, at least as a short-term solution.

[326] Caldera, *supra* note 325, at 90–91.

authenticity under Rule 901(b) are effective, efficient, and have stood the test of time. Thus, trial courts retain the flexibility to rely on multiple means to corroborate authenticity, looking to the totality of the evidence and a combination of sources, including percipient witness authentication and digital forensics evidence and expertise. As self-authenticating software becomes available on more devices, a court may be able to look to Rule 902(13) and (14) to make the required determination of authenticity. However, as some commentators have noted,[327] the silent witness theory will not be helpful when handling deepfakes, because the technology is too sophisticated to warrant the trust required to authenticate evidence under this theory without an authenticating witness.

B. Solving the Problem of Juror Skepticism and the Deepfake Defense: Rules of Evidence and Instructions

Countering juror skepticism and doubt over the authenticity of audiovisual images in the era of fake news and deepfakes calls for reallocating the factfinding authority to determine the authenticity of audiovisual evidence. As discussed in Part II, the trial court currently makes the initial determination on authenticity under Rule 104(a), and after that prima facie showing has been made, the evidence is presented to the jury for the ultimate determination as to its authenticity.[328] In light of the challenges posed by deepfakes, this process is not sufficient to dispel juror doubts over what is real. Thus, Rule 901 should be amended to add a new subdivision (c), which would provide:

> 901(c). Notwithstanding subdivision (a), to satisfy the requirement of authenticating or identifying an item of audiovisual evidence, the proponent must produce evidence that the item is what the proponent claims it is in accordance with subdivision (b). The court must decide any question about whether the evidence is admissible.

This new subdivision would relocate the authenticity of digital audiovisual evidence from Rule 104(b) to the category of relevancy in Rule 104(a).[329] Proposed Rule 901(c) would therefore expand the gatekeeping function of the court by assigning the responsibility of deciding authenticity issues solely to the judge.

The proposed rule would operate as follows. After the pretrial hearing to determine the authenticity of the evidence, if the court determines that the evidence is authentic under Rule 901(b), the court admits the evidence. The court would address the issue through jury instructions informing the jury that it must

[327] *See generally* Breen, *supra* note 239.

[328] *See supra* Part II.B.1.

[329] *Compare* 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."), *with* 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.").

accept as authentic the evidence that the court has determined is genuine. The court would also admonish the jury to weigh that evidence, but not question its authenticity. Thus, under Rule 901(c), the jury would no longer be the final arbiter of the authenticity of digital audiovisual evidence. This new rule would take the jury out of the business of determining authenticity, thereby avoiding the problems invited by juror distrust and doubt. Finally, the court would address the threat of counsel exploiting juror doubts over the authenticity of evidence using the deepfake defense by ordering counsel not to make such arguments.[330]

Although this expansion of the judge's fact-determining function is unusual, it is not without precedent. As discussed in Part II, before the 1930s, courts in the United States applied the traditional English view that the judge had plenary authority to decide all questions of fact conditioning the admissibility of testimony.[331] And even the proponents of the fact-determination approach embodied in Rule 104(a) and (b) acknowledge that it is sometimes still necessary to apply the English view to many foundational facts.[332]

The judge's role over admissibility is expanded, for example—and relevant to the discussion of deepfake evidence—where the evidence is so prejudicial, or so complex that committing the factual determination to the jury under Rule 104(b) would taint its deliberation process. Admissibility under these circumstances is treated as a preliminary foundational fact decided by the judge under Rule 104(a). The facts conditioning the application of attorney-client privilege are illustrative of the type of evidence that causes prejudice, and thus warrants special treatment. One of the foundational facts is that the client's communication with their attorney occurred in physical privacy.[333] Take the example of an eavesdropper who hears an accused charged with child abuse confess the crime to their attorney. The prosecutor may seek admission of the confession, arguing that "the conversation was not private because the client knew that a third party was standing within easy earshot," while the accused may deny that "the third person was present at the time."[334]

> In this situation, realistically the lay jurors cannot be trusted to administer the privilege doctrine. Even if they found that the attorney-client communication was confidential and technically privileged, during any later deliberations at the subconscious level they might nevertheless be influenced by their exposure to the testimony about the confession. For that matter, at a conscious level some jurors might be tempted to consider the inadmissible confession; it

[330] The use of the deepfake defense raises ethical issues for lawyers, which may require the imposition of sanctions and professional discipline. Those concerns and the solutions aimed at deterring the deepfake defense warrant scholarly consideration, but are beyond the scope of this Article.

[331] *See supra* Part II.B.1.a.

[332] *See generally* Morgan, *supra* note 205.

[333] EDWARD IMWINKELRIED & TIMOTHY HALLAHAN, CALIFORNIA EVIDENCE CODE ANNOTATED 33, 36 (2013).

[334] Imwinkelried, *supra* note 203, at 12. Although early common law permitted eavesdroppers to testify to otherwise privileged communications, the modern view runs contrary to that approach. 1 CHARLES T. MCCORMICK, EVIDENCE § 74 (7th ed. 2013).

> might strike them that applying the "technical" evidentiary rule would frustrate substantive justice and set a guilty person free.[335]

"No one doubts the jury's competence to decide whether a person was physically present when the attorney and client conversed; the issue of a person's presence at a particular time and place is a simple, straightforward question that jurors can easily resolve."[336] Nevertheless, this preliminary fact determination has been assigned to the judge.[337] This approach reflects the policy underlying the privilege and its relative importance to the justice system: encouraging and facilitating free communication within certain protected relationships.[338] It also reflects the fear that the jury, having heard the foundational fact and "contents of the proffered privileged communication, will have difficulty complying with an instruction to disregard the communication [even] if they decide that it was privileged."[339]

Another example of prejudicial evidence is the identity of the person who commits an act of misconduct under Federal Rule of Evidence 404(b).[340] There is some disagreement as to the proper treatment of facts showing the identity of the person who commits an act of misconduct under the Rule. Suppose, for example, that in a personal injury case, the plaintiff contends they were injured when the defendant punched them in the face during a bar fight. It would be permissible for the plaintiff to introduce evidence that shortly before trial, the defendant attempted to bribe the bartender to induce the witness to give perjurious testimony that the defendant was not in the bar the night the fight occurred. The judge would admit the evidence under Rule 404(b) to show the defendant's consciousness of liability or guilt;[341] the defendant's misconduct is logically relevant on a noncharacter theory as circumstantial evidence of the defendant's identity as the person who assaulted the plaintiff. Of course, before the proponent may introduce this evidence, the proponent must prove that the

[335] Imwinkelried, *supra* note 203, at 12–13; *see also* Stephen A. Saltzburg, *Standards of Proof and Preliminary Questions of Fact*, 27 STAN. L. REV. 271, 271 n.2 (1975) ("[T]repidations as to the ability of jurors fairly to evaluate certain kinds of evidence give rise to exclusionary rules."); United States v. James, 576 F.2d 1121, 1127–32 (5th Cir. 1978) (explaining that the determination of the voluntariness of a confession should be allocated to the trial judge because after exposure to a confession, jurors may be "influenced by [a] belief that the confession, even though coerced was true," reasoning that it is "unrealistic" to think that lay jurors would not be "swayed" by the foundation testimony).

[336] Edward J. Imwinkelried, *Judge Versus Jury: Who Should Decide Questions of Preliminary Facts Conditioning the Admissibility of Scientific Evidence?*, 25 WM. & MARY L. REV. 577, 615 (1984) (citations omitted).

[337] *Id.*

[338] FED. R. EVID. 501 advisory committee's note (1974 enactment).

[339] Imwinkelried, *supra* note 336.

[340] FED. R. EVID. 404(b); *see also* 29 C.F.R. § 18.404 (2022) ("Evidence of other crimes, wrongs, or acts is not admissible to prove the character of a person in order to show action in conformity therewith. It may, however, be admissible for other purposes, such as proof of motive, opportunity, intent, preparation, plan, knowledge, identity, or absence of mistake or accident."). Rule 404(b) allows a proponent to present evidence of uncharged misconduct when the misconduct has special noncharacter relevance to the facts of the case. *See* FED. R. EVID. 404(b).

[341] FED. R. EVID. 404(b).

defendant is the person who committed the uncharged act, namely, the attempted bribery. Before the adoption of the Federal Rules of Evidence, the prevailing view was that the person's identity as the perpetrator of the uncharged act is a fact conditioning the competence of the evidence.[342] Courts believed that this species of evidence was particularly prejudicial. Courts feared that jurors might reason that "where there's smoke, there's fire";[343] in other words, even if the evidence of the uncharged act was weak, courts feared the testimony might impugn the defendant's character and make jurors more willing to find liability or guilt.[344] Based on this reasoning, some courts continue to allocate this foundational fact to the judge, classifying the actor's identity as a Rule 104(a) competence question.[345] Courts that continue to follow this approach do so because of the prejudicial nature of the evidence; they take the view that it is unrealistic to expect lay jurors to completely disregard testimony about a party's uncharged misdeeds even when the testimony falls short of supporting a permissive inference of the commission of the misdeeds.[346]

This same rationale also applies to the second type of complex evidence. Under the prevailing view, the trial judge determines the foundational fact of the methodological validity of a scientific theory as a Rule 104(a) issue.[347] There is a plausible policy justification for the prevailing view—that jurors lack the ability to disregard scientific evidence they have been exposed to, even if they believe that the evidence is technically inadmissible.[348] Indeed, the foundations for scientific evidence tend to be longer than typical evidentiary foundations. In several minutes, a proponent can lay a foundation for firsthand knowledge under Rule 602. By contrast, laying a foundation for DNA or other scientific evidence may consume hours or even days of courtroom time for jurors to sit through.[349] Commonsense and psychological literature suggest that even if the jury makes a conscious decision that the evidence is inadmissible, it may be difficult, if not impossible, for jurors to put the evidence out of their minds during deliberations, even if they determine that the preliminary fact does not exist.[350] Thus, the overwhelming majority view is that the trial judge should classify this issue as a competence question, listen to the foundational testimony on both sides, and rule

[342] *See* Imwinkelried, *supra* note 193, at 820–25.

[343] *Id.* at 840.

[344] *Id.* at 837–44.

[345] *See* People v. Garner, 806 P.2d 366, 374 (Colo. 1991); Phillips v. State, 591 So. 2d 987, 989 (Fla. App. Dist. Ct. App. 1991).

[346] *See Garner*, 806 P.2d at 374; *Phillips*, 591 So. 2d at 989.

[347] *See e.g.*, People v. King, 72 Cal. Rptr. 478, 482 (Cal. Ct. App. 1968).

[348] *See* Imwinkelried, *supra* note 336, at 597–98.

[349] "[I]n one case involving a challenge to radar speedometer evidence, the out-of-court admissibility hearing generated over 2000 pages of testimony and arguments." Imwinkelried, *supra* note 203, at 14–15 (internal quotation marks omitted). And in another case involving DNA, the admissibility hearing "took place over a twelve week period producing a transcript of approximately five thousand pages." *Id.* (quoting People v. Castro, 545 N.Y.S.2d 985, 986 (N.Y. 1989)).

[350] *See* Imwinkelried *supra* note 336, at 605–06.

on the merits of the validity of the scientific theory or technique.[351] As in the case of privilege foundations, if the jury were exposed to the foundational testimony on the scientific evidence and found the evidence technically inadmissible, there would be a grave risk that the foundational testimony would nevertheless distort their subsequent deliberations.

To be sure, these policy-based allocations of factfinding to the court in the application of privilege or the identity of the perpetrator of uncharged misconduct do not purport to improve the veridicality of legal factfinding. Instead, they are understood to sacrifice some relevant information to the jury in service of some other policy goal. Likewise, in the case of foundational testimony on scientific evidence, the jury is shielded from the evidence to prevent jurors from giving it improper weight in their deliberations. Professor James R. Steiner-Dillion, a critic of displacing the jury in favor of judicial factfinding, has characterized this idea of assigning fact determinations to a judge as "epistemic curation"; that is, the process by which the Rules limit what is presented to the jury due to certain policy choices.[352]

Deepfake evidence warrants epistemic curation. Deepfakes are technically complex and highly prejudicial to jury deliberations. Thus, they present the same dangers and warrant the same treatment as privilege, bad character evidence, and foundational scientific evidence. When technologies such as DNA; x-rays; computer technologies; and video, audio, and other digital recordings first presented authentication problems, the initial reaction of some academic commentators and a few courts was to call for a special authentication procedure.[353] But in all those situations, the system adapted to the conditional relevance procedures and the division of labor between the judge and the jury baked into Rule 104(a) and (b).[354] The science or technology was comparatively static and accessible; the trier of fact could be taught the science, and juries would trust a properly qualified expert to explain it. The jury's decision-making thus became more akin to the categorical and routine decision-making used in assessing other foundational questions under Rule 104(b), such as a lay witness's personal knowledge or a letter's authenticity.

Deepfakes are truly different; by design, they trick the viewer. They raise existential questions about reality on a profound and metaphysical level;

[351] *Id.* at 597–99. In addition, as Professor Imwinkelried points out, the probabilistic nature of foundation testimony for scientific evidence increases the chance that jurors will not be able to set aside the testimony during their deliberations, even if at a conscious level they have decided that the evidence is technically inadmissible. Unlike in the case of other foundational questions such as a lay witness's personal knowledge or a letter's authenticity, jurors may be inclined to conceive of the question as a categorical issue: either the witness viewed the accident, or did not. In contrast, the foundation testimony for a scientific theory or technique is often explicitly probabilistic, identifying the margin of error for the scientific hypothesis. Knowledge of this margin of error may taint the jury's deliberation process. *Id.* at 604–06.

[352] James R. Steiner-Dillion, *Is Truth Truth?*, 109 KY. L.J. 477, 490–92 (2021).

[353] *See supra* Part II.A.

[354] *See* Imwinkelried, *supra* note 193, at 822–33 (discussing the history of the allocation of fact determinations between judges and juries).

deepfakes, fake news, and conspiracy theories have invaded the cultural ethos and political pathos. Moreover, leading digital forensics experts worry that the fight to detect deepfakes is a losing battle—that deepfake technology is outstripping the ability of those trying to detect the deepfakes.[355] Deepfakes present an exceptional and unprecedented challenge to determining authenticity that warrants unique treatment under the Rules of Evidence. There is a genuine risk that jurors will harbor doubts about what they see and hear either based on a "seeing is believing" mindset—or the opposite view that everything is fake—and that those biases will taint their deliberative process. As the Max Planck Center for Humans and Machine's recent study revealed, people cannot reliably detect deepfakes even when made aware of them.[356] This study and other anecdotal reports show that the determination of whether a deepfake is genuine will, in some cases, exceed the jury's abilities.[357] Thus, this evidence presents the same risk as other highly prejudicial[358] or technical evidence,[359] 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.

However, the proposal to reallocate the factfinding duties between the jury and the court in a way that expands the judge's gatekeeping role under Rule 901 is not offered lightly. Critics of this proposal will point out that judges are human, too. They will argue that judges are subject to the same cognitive biases as jurors,[360] and that judges rely on the same heuristic responses as lay people in making decisions.[361] Scholars like Professor Steiner-Dillion complain about the

[355] *See* Brown, *supra* note 78, at 25 (discussing the belief among the community of computer science and digital forensics experts that detection methods cannot keep pace with the innovations aimed at evading detection).

[356] *See* Köbis et al., *supra* note 91.

[357] *Id.* at 11; *see also, e.g.*, Sophie J. Nightingale, Kimberley A. Wade & Derrick G. Watson, *Can People Identify Original and Manipulated Photos of Real-World Scenes?*, 2 COGNITIVE RSCH. 30, 30 (2017); Oscar Schwartz, *You Thought Fake News Was Bad? DeepFakes Are Where Truth Goes To Die*, THE GUARDIAN (Nov. 12, 2018, 5:00 AM), https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth; Camila Domonoske, *Students Have 'Dismaying' Inability To Tell Fake News from Real, Study Finds*, NPR (Nov. 23, 2016, 12:44 PM), https://www.npr.org/sections/thetwo-way/2016/11/23/503129818/study-finds-students-have-dismaying-inability-to-tell-fake-news-from-real.

[358] *See* Imwinkelried, *supra* note 193, at 817–18 (arguing that the judge rather than the jury should be allocated the responsibility to determine the preliminary evidentiary fact of identity prior to the admission of evidence of uncharged misconduct under Federal Rule of Evidence 404).

[359] *See* Markman v. Westview Instruments, 517 U.S. 376, 384 n.10 (1996) (holding that the Seventh Amendment, which guarantees a jury trial in patent infringement cases, does not commit to the jury the construction of claim terms). The court in *Markman* held that "judges, not juries, are the better suited to find the acquired meaning of patent terms." *Id.* at 388. This is because "[t]he judge, from his training and discipline, is more likely to give a proper interpretation to [patent claims] than a jury; and he is, therefore, more likely to be right, in performing such a duty, than a jury can be expected to be." *Id.* at 388–89 (quoting Parker v. Hulme, 18 F. Cas. 1138, 1140 (C.C.E.D. Pa. 1849)).

[360] James R. Steiner-Dillion, *Epistemic Exceptionalism*, 52 IND. L. REV. 207, 224–42 (2019).

[361] *See* Stephanie L. Damon-Moore, *Trial Judges and The Forensic Science Problem*, 92 N.Y.U. L. REV. 1532, 1559–60 (2017) ("Judges faced with complex scientific questions for which they have little or no training, often combined with an absence of adequate information from the defense, may fail to grasp the

tendency to supplant the jury in favor of judicial factfinding based on the unfounded conclusion that the task is too challenging for the jury to perform competently.[362] Relying on empirical literature on human cognition, Professor Steiner-Dillion asserts that "epistemic exceptionalism"—the claim that judges' cognitive processes can be trusted to operate with greater competence and objectivity than those of laypersons even in the absence of evidentiary and procedural constraints—is an exaggeration.[363] Professor Steiner-Dillion maintains that the cognitive differences between judges and laypersons are generally insignificant compared to the extent to which both groups show similar susceptibility to cognitive illusions, implicit bias, and motivated reasoning.[364]

Although judges are not immune from all the cognitive errors and biases that affect lay jurors, empirical studies also show that judges are not the same as lay persons in respects that are important for evaluating deepfake evidence. Judicial experience grants resistance to some cognitive fallacies.[365] Studies also show that properly trained judges are better at making certain determinations than lay jurors.[366] While the limited available research shows that lay people cannot reliably identify deepfakes even with exposure to them,[367] judges may be better suited for the task. Based on existing published research,[368] judges should be more adept at engaging in the mental discipline of setting aside skepticism and irrelevant matters in determining the authenticity of deepfake evidence, because they are professional evidence evaluators. Judges can also develop

shortcomings of forensic evidence because they are relying on heuristics, or mental shortcuts, to sidestep the substantive question altogether.").

[362] *See* Steiner-Dillion, *supra* note 360.

[363] *Id.*

[364] *Id.*

[365] *See* Valerie Hans, *Judges, Juries, and Scientific Evidence*, 16 J.L. & POL'Y 19, 36–37 (2007). In Hans's comparative study, only fifteen percent of judges accepted the fallacious argument presented by a hypothetical defense attorney that associative mitochondrial DNA evidence was irrelevant, as compared to forty-nine percent of mock jurors. *Id.*

[366] *See* Elizabeth Thornburg, *(Un)Conscious Judging*, 76 WASH. & LEE L. REV. 1567, 1620–23 (2019) (arguing that with sufficient relevant-subject-matter training judges can overcome any heuristic response that might otherwise infect their decision-making process). Some empirical evidence suggests that judges are not as susceptible to all cognitive biases as is the population as a whole. Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind*, 86 CORNELL L. REV. 777, 784, 816–17 (2001) (finding in a study of 167 federal magistrate judges that they were just as susceptible as lay people to anchoring, hindsight bias, and egocentric bias, but were less susceptible to framing and the representativeness heuristic).

[367] *See* Köbis et al., *supra* note 91. There is no indication in the study whether any of the test subjects were members of the judiciary, and there are no published research studies that specifically examine the acumen of judges to discern deepfakes, which counsels that there is more work that needs to be done in this space.

[368] *See* Andrew J. Wistrich, Chris Guthrie & Jeffrey J. Rachlinski, *Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding*, 153 U. PA. L. REV. 1251, 1318–22 (2005) (testing the ability of judges to disregard several categories of legally inadmissible evidence and finding that judges' decision-making process was not influenced by information deemed inadmissible in certain areas; specifically, judges were generally able to disregard information concerning the outcome of a police search when adjudicating issues of probable cause and showed some ability to disregard illegally obtained confessions); *see also* Jeffrey J. Rachlinski, Chris Guthrie & Andrew J. Wistrich, *Probable Cause, Probability, and Hindsight*, 8 J. EMPIRICAL LEGAL STUD. 72, 72 (2011) (conducting a study of 900 state and federal judges that found that judges' probable cause judgments are generally unaffected by knowledge of the outcome).

expertise in assessing deepfake evidence and digital forensics technology outside the context of any particular case, which should minimize any biases they may bring to any individual case.[369]

To be sure, treating the question of deepfake authenticity as one for the court to decide as part of its gatekeeping function under Rule 104(a) may not be the foolproof or final fix to the challenges deepfake evidence presents in legal proceedings. But given that we live in an era in which we can no longer trust our eyes because of the daily bombardment of fake news, conspiracy theories, and technological trickery, this proposal is the much-needed place to start.

Conclusion

One of the fundamental tenets of the American legal system is that the trier of fact—either the judge or the jury—is best equipped to find the truth based on the evidence presented. But individuals cannot consistently determine truth from lies as they confront deepfakes. Deepfake evidence will increasingly appear in legal proceedings as the technology to create them becomes more accessible, complicating evidentiary issues in court. It will require lawyers and courts to take additional measures and employ more resources to determine the authenticity of digital audiovisual evidence before presenting it to the jury. As lawyers, courts, and jurors traverse this new technological landscape, the way forward will require a robust response from the law of evidence, including a reimaging of the roles of the judge and jury in the fact-determining process. Deploying these new approaches and mechanisms will protect the truth-seeking function at the heart of our justice system.

[369] Thornburg, *supra* note 366, at 1620–23. *See* Steiner-Dillion, *supra* note 360, at 230, 241–42 (recognizing that training and expertise make judges more resistant to cognitive error than lay jurors, and that when judges specialize or receive special training they more effectively mitigate bias).