

ARTIFICIAL INTELLIGENCE SYSTEMS AND CRIMINAL PROCEDURE: AI, CHATGPT AND DEEPFAKES

*Michael Martin Losavio**

INTRODUCTION.....	1005
I. TRUE CRIME, TRUE EVIDENCE & AI DEEPFAKES	1009
A. <i>AI, Deepfakes, and Evidence</i>	1010
B. <i>Inverse Impact</i>	1013
C. <i>Deepfakes and Evidence Law</i>	1014
D. <i>Character Evidence; Other Crimes, Wrongs, or Acts – Federal Rule of Evidence 404</i>	1015
E. <i>Character Evidence</i>	1016
F. <i>Evidence of Relevance Beyond Simple “Other Facts”: Habit, Routine Practices, Other Sexual Behavior, and Other Sexual Misconduct</i>	1017
G. <i>True Reliability in the Face of AI Deepfakes</i>	1020
H. <i>Threats Abound</i>	1025
CONCLUSION	1028

INTRODUCTION

In a world of hyper-powerful machine systems, Artificial Intelligence (AI), Generative Adversarial Networks (GAN), and Generative Pre-trained Transformers (GPT), we must prepare for novel attacks on truth and justice via systems far beyond our experience. These technologies raise challenges to the entire discipline of cybersecurity, itself a foundation for reliability of the

* Associate Professor, Department of Criminal Justice and Department of Computer Science and Engineering, University of Louisville.

information we use everywhere.¹ And these systems may possibly be beyond the full understanding of anyone where machine learning builds itself using semi-supervised and unsupervised learning where the machine builds itself beyond programmer action. Legal and judicial practitioners must prepare for this computational world. This novel moment in information technology may shift responsibility for the generation and evaluation of evidence to an algorithm that can be manipulated and mutated to do bad things or open avenues for bad people to act freely without fear of accountability.

Further, the impact on political and personal life may be enormous. Conflict over the use of generative pre-trained transformer (“GPT”) manipulation of political speech threatens the life of the republic over “fake news” of great power.² The use of such manipulation in personal life via deepfake imagery threatens the destruction of personal autonomy and intimate family life where false images create false narratives of personal conduct.³

The justice system cannot let the bad guys win. Not one of those who committed a crime, failed to remedy injuries to others, or denied someone their civil rights should be allowed to evade punishment on account of these novel computational systems. There must be a review of evidentiary practices to assure this in the world of AI and its evidentiary artifacts. Lawyers and judges must be cognizant of the particular aspects of these computational systems that may not even be clearly understood by those who designed them. And issues relating to the manipulation of politics

¹ Bibhu Dash & Pawankumar Sharma, *Are ChatGPT and Deepfake Algorithms Endangering the Cybersecurity Industry? A Review*, 10(1) INT’L J. OF ENG’G & APPLIED SCI. 1 (Jan. 16, 2023), https://www.ijeas.org/download_data/IJEAS1001001.pdf [<https://perma.cc/7VPT-BUZ5>] (discussing how technology utilizes machine learning to manipulate images and videos jeopardizes the ability to differentiate between real and fake images).

² See, e.g., Robert McMillan et al., *New AI Deepfakes Complicate 2024 Votes Elections*, WALL ST. J., February 16, 2024, (discussing how AI could have a detrimental effect on voter turnout with the spread of deepfakes).

³ See, e.g., Ashley Belanger, *Teen Boys use AI to Make Fake Nudes of Classmates, Sparking Police Probe*, ARS TECHNICA (Nov. 2, 2023), <https://arstechnica.com/tech-policy/2023/11/deepfake-nudes-of-high-schoolers-spark-police-probe-in-nj/> [<https://perma.cc/PM46-YPPM>] (Last visited Feb. 18, 2024) (discussing how teenage boys had been using AI image generators to create and share fake nude photos of a female classmate).

through the use of the systems only now demonstrate need for powerful systems to prevent such manipulation.

Our adjudicative systems for truth and justice, as imperfect that effort may seem in the United States of America and many other nations, is driven in part by rules of procedure and evidence. The rules guide the process in order “to provide for the just determination of every criminal proceeding []”⁴ and to promote “ascertaining the truth and securing a just determination.”⁵ Other adjudicative forums for civil and administrative disputes similarly strive truth and justice. Yet these noble aspirations, although drafted with knowledge of human fallibility in our reasoning perceptions, now must face the possibility of error and falsification from technical systems beyond that we have ever experienced.⁶

The evidence continuum in American law sets out a range of weights of evidence that, with greater and greater weight, justify increasingly intrusive state action where that evidence passes at least basic standards of reliability. Those weights of evidence are measured against standards by a human adjudicator and human reason in all their glory and fallibility. The new challenge is that those standards of reliability may be spoofed by AI technology such that adjudication of AI-mediated facts becomes corrupted by those very machine logics.

One subset of possibilities for an overlap between criminal procedure and AI would be the following three areas:

- 1) The reliability of evidence,

⁴ FED. R. CRIM. P. 2. (“These rules are to be interpreted to provide for the just determination of every criminal proceeding, to secure simplicity in procedure and fairness in administration, and to eliminate unjustifiable expense and delay.”).

⁵ FED. R. EVID. 102. (“These rules should be construed so as to administer every proceeding fairly, eliminate unjustifiable expense and delay, and promote the development of evidence law, to the end of ascertaining the truth and securing a just determination.”).

⁶ As an example of a legal error, the Salem Witch Trials that happened under the auspices of “the law”, is considered a dark period on our country’s legal history. Twenty people were put to death when legal failings, mass paranoia, and puritan religious and societal rules converged against a backdrop of economic and political uncertainty in colonial Massachusetts. For further discussion, see *The True Legal Horror Story of the Salem Witch Trials*, NEW ENG. L. BOSTON, <https://www.nesl.edu/blog/detail/a-true-legal-horror-story-the-laws-leading-to-the-salem-witch-trials> [https://perma.cc/Q398-FRAX] (last visited May 9, 2024).

- 2) The identification of “persons of interest,” and
- 3) The adjudication of just outcomes as to issuance of search and seizure orders, adjudication of guilt, and determination of a just sentence.

Each of these contributes to the next as to just and fair determinations:

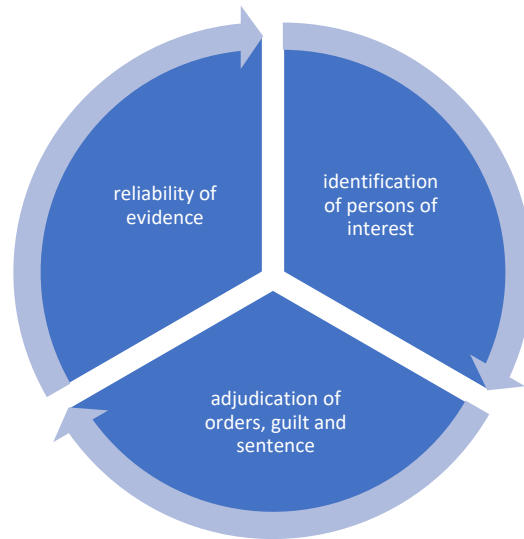


Figure 1: Sequencing Chart on Evidentiary Contributions.

AI outputs and their own dangers with embedded confirmation bias possibly driving an investigation and prosecution, regardless of exculpatory evidence, and reinforcing error.⁷ Bias in AI systems may be a product of biases in the training data for the system or its coding.⁸ AI biases are particularly problematic where they put

⁷ Michael D. Schlosser et al., *Confirmation bias: A barrier to community policing*, 6 J. CMTY. SAFETY & WELL-BEING 162, 162-63 (2021).

⁸ Zhisheng Chen, *Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices*, 10(567) HUMANITIES & SOC. SCI. COMM'N 1 (2023). <https://doi.org/10.1057/s41599-023-02079-x> [<https://perma.cc/8GEP-7EZQ>].

classes of people “at a systematic disadvantage.”⁹ Consider how each affects the system of criminal justice. Their interplay is inherent in how that system functions regardless of the inaccuracies and flaws that may impact the outcome. Thus, the questions and concerns then become: (1) how will machine systems improve the system of justice, and (2) how might they hurt it?

I. TRUE CRIME, TRUE EVIDENCE & AI DEEPFAKES

“Evidence” used in any aspect of criminal justice should have some indicia of reliability and weight, even when the finders-of-fact are desperate for something – anything – to guide them forward in resolving issues of criminal justice.

The Federal Rules of Evidence offers a helpful outline of the various ways in which evidence should be vetted for appropriate use in judicial forums. The foundation set by Rules 401 and 402 of the Federal Rules of Evidence provides that evidence that is relevant to a determination is generally admissible for use in a trial establishing liability, as well as other determinations.¹⁰ To be relevant, evidence must be proven that:

- (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and
- (b) the fact is of consequence in determining the action.¹¹

For example, images, audio and video that demonstrate that: a particular individual committed some act, was in some location, or possessed special agency related to misconduct, or which demonstrate some other facts that make a fact of consequence in a judicial determination, can all be direct evidence that can lead to an allocation of responsibility, including a finding of guilt and accompanying punishment. But if erroneous evidence leads to error in particular findings of fact, then error in judgment may occur. The ability to create “deep fake” images, audio and video raises the risk of the fabrication of false evidence leading to an erroneous finding. This is the risk and challenge that the evidence may be erroneous,

⁹ Xavier Ferrer, *Bias and Discrimination in AI: A Cross-Disciplinary Perspective*, IEEE TECH. & SOC'Y MAG., June 2021, at 72-80.

¹⁰ FED. R. EVID. 401; FED. R. EVID. 402.

¹¹ FED. R. EVID. 401 (alternation in original).

created by the new world of AI-generated deep fake artifacts as evidence.

Though a seemingly basic and clear threshold, the Rules of Evidence of the United States and other nations, and related jurisprudence, continue to define ways in which relevant evidence is appropriate and reliable. The fabrication powers of AI systems, easily generating evidentiary artifacts that, if admitted, would make facts of consequence more probable, increase the risks of erroneous outcomes.

A primary challenge is to the reliability of that evidence as to “ascertain the truth” and secure a just determination controversy at issue. Rule 901 of the Federal Rules of Evidence requires basic authentication of evidence that it is what it claims to be, supported by evidence sufficient to find that. This is especially important for images, audio and video, which historically were difficult for the unskilled to fabricate before the advent of Photoshop and other digital image manipulation tools. The evolution of even more sophisticated image tools via AI systems confronts the inherent treachery of images and the very liberties of our Republic where fabrication is easy and detection increasingly difficult.¹²

A. AI, Deepfakes, and Evidence

Deepfakes are manufactured objects, “an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said[.]”¹³ It may be more technically described as “A deepfake is an artificial image or video (a series of images) generated by a special kind of *machine learning* called “deep” learning (hence the name).”¹⁴ The origin of the term, per Merriam Webster Dictionary, is that while a fake is an artifact that is fake, or false, “[t]he *deep* is less self-explanatory: this half of the term

¹² See, e.g., René Magritte, *La Trahison des Images (Ceci n'est pas une pipe)* [*The Treachery of Images (This is not a pipe)*] (painting), in LOS ANGELES CNTY. MUSEUM OF ART, <https://collections.lacma.org/node/239578> [https://perma.cc/7UBS-Q7AZ] (last visited Jan. 28, 2024).

¹³ *Deepfake*, MERRIAM-WEBSTER, <https://www.merriam-webster.com/dictionary/deepfake> [https://perma.cc/SZ56-42HV] (last visited Aug. 8, 2023).

¹⁴ Univ. Va. Info. Sec., *What the Heck is a Deepfake?*, <https://security.virginia.edu/deepfakes> [https://perma.cc/X2G5-HF4D] (last visited Feb. 17, 2024).

is specifically influenced by *deep learning*—that is, machine learning using artificial neural networks with multiple layers of algorithms.”¹⁵ These algorithms can massage a digital image as to reduce or remove the parts of the image or audio or video that would show manipulation of that file.

Those file aspects showing manipulation include inconsistencies in the imagery, such as inconsistent lighting, audio or movement, and in the file encoding itself that demonstrate manipulation. By using facial recognition algorithms, a deep learning computer network, and a variational auto-encoder,¹⁶ itself a generative AI algorithm, images can be curated to capture movement and lighting and normalize them as to make detection of these inconsistency artifacts difficult.¹⁷

Generative adversarial networks (“GAN”) take this process further in order to hide any indications of manipulation. The GAN will use two interconnected neural networks to remove evidence of manipulation. The first machine builds the false image (the generator), and then the second machine evaluates it (the discriminator) for flaws and then feeds its findings back to the generator for those flaws to be corrected. This process may go through several cycles and can be done with images, video and audio. Combining a system like Respocher, a system for generating fake voices and CannyAI, a system to synchronize video of a speaker to make mouth and facial movements seem normal, Francesca Panetta and Halsey Burgund of MIT built a project around President Richard Nixon making a speech he never gave.¹⁸

As one U.S. Congressman noted:

As we enter 2020, the problem of disinformation, and how it can spread rapidly on social media, is a ***central and continuing national security concern***, and a real

¹⁵ *Id.*

¹⁶ “A variational autoencoder (VAE) is a generative AI algorithm that uses deep learning to generate new content, detect anomalies and remove noise.” George Lawton, *Variational Autoencoder (VAE)*, TECHTARGET, <https://www.techtargget.com/searchenterpriseriseai/definition/variational-autoencoder-VAE> [https://perma.cc/9WMU-HDSK] (Last visited May 1, 2024).

¹⁷ Meredith Somers, *Deepfakes, Explained*, MIT MGMT. SLOAN SCH. (July 21, 2020), <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained> [https://perma.cc/CBU3-A33E].

¹⁸ *Id.*

threat to the health of our democracy.

... the next wave of disinformation that could come in the form of ‘deepfake’ — AI-generated video, audio, and images that are difficult or impossible to distinguish from real thing. As experts testified in an open hearing in the Intelligence Committee last year, the technology to create deepfakes is advancing rapidly and widely available, to state and non-state actors, and has already been used to target private individuals, primarily women, for abuse and harassment.¹⁹

The danger of the use of such fabrication technology in juridical proceedings is even greater. Deepfakes increase the difficulty in authenticating, or challenging the authenticity, of purported evidence.²⁰ Admission of such evidence risks heightened use of fabricated evidence supporting guilt or innocence, liability and accountability, in all manner of criminal, civil, administrative and private proceedings.

A great deal of work is ongoing as to the image inconsistencies indicating fabrication of evidence. Some of those artifacts have been noted by writer Laura Temme as potentially applicable to the fake evidence:

¹⁹ Adam Schiff, *Rep. Schiff Statement on Facebook’s Deepfake Policy*, ADAM SCHIFF (Jan. 7, 2020), <https://schiff.house.gov/news/press-releases/rep-schiff-statement-on-facebooks-deepfake-policy> [<https://perma.cc/D6SR-X49K>] (emphasis added).

²⁰ Agnieszka McPeak, *The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood*, 23 VAND. J. ENT. & TECH. L. 433, 444, 450 (2021).

- 1) Inconsistent lighting across the images,
- 2) Inconsistent or odd eye and body movements,
- 3) Unnatural facial features,
- 4) Suspicious metadata that is inconsistent with claimed authenticity or provenance of the media files.²¹

Yet the danger in Deepfakes is that these inconsistencies can be algorithmically corrected by a GAN AI system, as demonstrated by researchers at MIT, though they noted that unnatural facial, audio and lighting artefacts still can indicate manipulation.²² Artifacts can be corrected by the algorithms themselves as to make detection difficult and inadequate to prevent falsehood from spreading. Engler notes that when techniques analyzing irregular blinking patterns that could indicate Deepfakes emerged, a correction system for this indicator appeared within a few months.²³ Some speculate that deepfake “supremacy,” when intrinsic detection becomes impossible, may come within a decade. This increases the need for extrinsic validation tools against the immense world of data collections to track source objects from which the fakes were generated.²⁴ Without these, digital artifacts may undermine reliability and trust in adjudicative decisions, whether admitting or excluding them as evidence.

B. Inverse Impact

Challenges have been made that tendered evidentiary objects are themselves the product of Deepfake technology and should not be used for judicial proceedings. This is the “liar’s dividend” posited by Chesney and Citron—one that threatens personal autonomy,

²¹ Laura Temme, *Tips for Catching Deepfakes in Evidence*, FINDLAW: PRAC. L. BLOG (May 21, 2023), <https://www.findlaw.com/legalblogs/practice-of-law/tips-for-catching-deepfakes-in-evidence/> [https://perma.cc/N974-T6T9].

²² *Id.*

²³ Alex Engler, *Fighting Deepfakes when Detection Fails*, BROOKINGS (Nov. 14, 2019), <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/> [https://perma.cc/R847-HH3W]; see also James Vincent, *Deepfake Detection Algorithms will never be enough*, THE VERGE (Jun. 27, 2019, 11:22 AM CDT), <https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work> [https://perma.cc/W2SZ-CSD9].

²⁴ *Id.*

democracy, and national security- all which require reliable facts.²⁵ They note that “*the ability to distort reality has taken an exponential leap forward with ‘deep fake’ technology.*”²⁶

Efforts to assert deepfake possibilities have been used in courts to try and exclude video evidence, including of Elon Musk, but have been rejected by the courts; some speculate that juries may be impacted by such claims, as with the “CSI Effect” leading juries to demand more proof of reliability, and thus lead to increasing denial of facts in the world.²⁷

How finders of fact, whether judge or jury, respond to these challenges in reliable evidence may have a profound impact on criminal and civil cases everywhere.

C. Deepfakes and Evidence Law

Even evidence deemed relevant may have its limits as to admissibility and use, as listed within the U.S. federal rules themselves. Though relevant, these rules expressly exclude evidence where its probative value is outweighed by other considerations that may interfere with the truth finding process.²⁸ These include situations where the usefulness of the relevant evidence is “[S]ubstantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.”²⁹ Evidence that a defendant used AI in the making of otherwise legal if disturbing pornography in a trial for fraud perpetrated by deepfakes might be excluded for unfair prejudice though marginally relevant to show the ability to use such technology. It would be within the discretion of the judge to admit or exclude such evidence.

²⁵ Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1758 (2019).

²⁶ *Id.*

²⁷ Shannon Bond, *People Are Trying to Claim Real Videos are Deepfakes. The Courts are Not Amused*, NAT'L PUB. RADIO (May 8, 2023, 5:01 AM), <https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused> [<https://perma.cc/WM78-HTVJ>] (Last visited May 1, 2024).

²⁸ FED. R. EVID. 403.

²⁹ *Id.*

These limitations equally apply in the context of AI-generated artifacts, beginning with the fundamental threshold requirement of any evidence, that it be relevant, and then the limitations that may exclude even relevant evidence from use.³⁰

*D. Character Evidence; Other Crimes, Wrongs, or Acts –
Federal Rule of Evidence 404*

Rule 404 generally prohibits the use of evidence of a person's character or other crimes, wrongs or acts by a person to demonstrate their responsibility for facts in issue in a particular dispute.³¹ Such evidence would be of minimal relevance as to causation in the instant case where it would be highly prejudicial to that person when it is used by the finder of fact as an indicator of his or her propensity to commit the offense. But such other acts evidence may be used prove "motive, opportunity, intent, preparation, plan, knowledge, identity, absence of mistake, or lack of accident."³² Proof of these facts may serve to link a defendant as the person responsible for the primary issues and actions in the dispute due to motive, plan, or other relevant information about that defendant.

Deepfake image, video and audio artifacts can circumvent the protections of Rule 404 by manufacturing media evidence that meets the requirements for permissible relevance under Rule 404(b).³³ Examples of such damaging Deepfake evidence include:

- In a prosecution involving emotional or romantic elements, deepfake images of an accused and the alleged subject of emotional engagement *in flagrante delicto* as to demonstrate a relationship that may promote jealousy against other objects of the alleged subject's actions, e.g., the motive for the murder of a romantic partner of X by another lover of X, the romantic relationship of X and lover demonstrated by images of their embrace.

³⁰ See FED. R. EVID. 403.

³¹ FED. R. EVID. 404.

³² FED. R. EVID. 404(b).

³³ *Id.*

- The use of the deepfake audio discussions between an accused and others as to show hatred or mendacity towards a victim, intent, and preparation for wrongful acts against a victim, and intentional action against a victim, showing the absence of a mistake or motive for harm.
- Deep fake video image showing that an accused who has asserted his innocence to particular misconduct was engaged in the direct commission of the misconduct (under FRE 402) or was in some way proximate to the offense as to establish his identity-shown in the imagery of the video itself. Such deepfake media may also show related facts associated with the misconduct, such as knowledge of methods of implementation, location, and time, as to establish for the finder of fact they may correctly infer that the accused did indeed commit the act.

These are just some ways in which the system of justice may be subverted by these new technologies. Human creativity is endless, especially towards malevolent action.

E. Character Evidence

Where character evidence may be introduced, means of proving it required testimony as to reputation or opinion as to the character secular individual at issue; but crucially, such an assessment may be challenged in cross-examination by inquiring into “relevant specific instances of.”³⁴ And when individuals character is an essential element of the claims against him or a defense thereto, character may be shown by “... relevant specific instances of the person’s conduct.”³⁵ Thus:

- Testimony that a particular individual’s law-abiding character can be countered by AI image artifacts showing “relevant specific instances of” the individual at issue committing illegal acts.

³⁴ FED. R. EVID. 405.

³⁵ *Id.*

- Testimony that a particular individual is honest can be countered by AI audio artifacts discussing dishonest conduct, including false testimony in the instant forum.
- Testimony that a particular individual is faithful can be countered by AI video artifacts demonstrating marital misconduct with third parties.

The use of deep fake artifacts to attack one's character can extend both to an accused in a criminal matter but also as to every witness testifying against the defendant. By undermining the credibility of witnesses, their testimony, though relevant, may then be discounted by the finder of fact even where true.

One example of this would be where the defendant produces an audio recording of a witness offering to give testimony in favor of the defendant in exchange for money or property or some other consideration. That it is claimed this was from the witness to the defendant avoids the question of how the defendant acquired the recording and possible challenges to two party consent for recording that conversation.³⁶

F. Evidence of Relevance Beyond Simple "Other Facts": Habit, Routine Practices, Other Sexual Behavior, and Other Sexual Misconduct

Evidence of a person's habit or an organization's routine practices may be admitted to show that they acted in accordance with those habits or routine practices as relevant to facts in issue in a dispute, and can be very powerful evidence for a finder of fact.³⁷ The Federal Rules of Evidence also permit the use of evidence that a defendant accused of sexual assault or child molestation committed other acts of sexual assault or child molestation.³⁸

³⁶ Author's Note: Where this occurs in a state that requires two party consent to a telephone recording then another exception would need to apply. Or the defendant may feel the punishment for the recording violation is worth the risk to avoid the more serious punishment of the instant prosecution.

³⁷ FED. R. EVID. 406. ("Habit; Routine Practice- Evidence of a person's habit or an organization's routine practice may be admitted to prove that on a particular occasion the person or organization acted in accordance with the habit or routine practice. The court may admit this evidence regardless of whether it is corroborated or whether there was an eyewitness.")

³⁸ FED. R. EVID. 413; FED. R. EVID. 414.

And in limited circumstances evidence of a victim's sexual behavior or disposition, or which may have a powerful tendency to devalue a victim in the eyes of the finder of fact, may be admitted. Although such attacks on the victim are limited by the Rules of Evidence, they may be used by an accused to show that some other person was responsible for physical evidence of the assault, that the sexual activity was consensual and where "... Exclusion would violate the defendant's constitutional rights."³⁹

Deep fake artifacts might be used to establish that one had particular habits or an organization had a routine practice that is relevant to establishing facts in issue in the forum. Establishing such strictly from the artifacts may require timeseries evidence, such as a set of images with time stamps of the purported recorded activity. It may be more effectively used to bolster the testimony of other witnesses as to those habits or routine practices. But conversely, these may be used controvert the testimony of habit or routine practice. Examples might be:

- Testimony that a particular person would always go for the morning coffee in a particular place far from the scene of an alleged activity might be controverted by images or video showing deviation from this habit,
- Testimony that a particular organization always processed particular activity in a particular way might be controverted images, audio or video showing this to be untrue through a deviation from that particular practice,
- Testimony used to support claims of a team practice to be controverted by audio relevant parties discussing variations in routine practice efforts to falsify evidence relating to it.

Use of deep fake artifacts relating to sex offense prosecutions are of even greater concern. There has been a focus on generation of sexual artifacts for improper purposes.⁴⁰ The use of such artifacts can be to both support a prosecution and support a defense thereto beyond the harassment, torment and extortion of women.

³⁹ FED. R. EVID. 412.

⁴⁰ See Karen Hao, *Deepfake porn is ruining women's lives. Now the law may finally ban it*, MIT TECH. REV. (Feb. 12, 2021), <https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/> [<https://perma.cc/8DUX-9ZY9>].

Given the latitude that is given in sex offense prosecutions to show evidence of other acts relating to sexual offenses and child molestation, the deepfake AI generation of artifacts may be particularly dangerous:

- In a prosecution for child molestation, deepfake images and video of a defendant engaging in sexual molestation with other children would be damning evidence in support of the instant prosecution.
- In a prosecution for sexual assault, deepfake images and video of a defendant engaging in sexual activity with particular idiosyncrasies of conduct similar to those used against a purported victim, e.g., bitemarks or bondage.
- In a prosecution for sexual assault, deepfake images and video of a defendant engaging in seemingly consensual sexual activity with the purported victim as to support the defense of consent by the victim.

The mischief deepfake technology can do with these preliminary concerns of relevancy and reliability of evidence demonstrates the risks to truth and justice. The challenge, again, is how to confront these technical powers.

There are preliminary protections for the use of such evidentiary artefacts within the rules. Before evidence relating to a victim's sexual conduct, a hearing must be held on the admissibility of such evidence and notice must be given 14 days prior to the trial, absent good cause shown for waving that requirement.⁴¹ Before evidence of a defendant's prior sexual assaults or child molestations can be admitted, it must be disclosed to the defendant, including witness statements or a summary of expected testimony, 15 days prior to trial, absent good cause shown for waiving that requirement.⁴²

Similarly, for character and other acts evidence under FRE 404 in a criminal case, the prosecutor must:

⁴¹ FED. R. EVID. 412.

⁴² FED. R. EVID. 413; FED. R. EVID. 414.

(A) provide reasonable notice of any such evidence that the prosecutor intends to offer at trial, so that the defendant has a fair opportunity to meet it;

(B) articulate in the notice the permitted purpose for which the prosecutor intends to offer the evidence and the reasoning that supports the purpose; and

(C) do so in writing before trial — or in any form during trial if the court, for good cause, excuses lack of pretrial notice⁴³

These requirements of notice and hearing meet fundamental due process notions of fairness and the right to challenge such evidence with reduced risk of surprise. The challenge, again, with deepfake technology is that technical challenges may be difficult or impossible to mount, especially on a short timeline where expert skills will be needed to examine these evidence artefacts.

The rules of evidence relating to reliability and the ability to challenge the reliability of evidence become all the more important in the era of deepfakes.

G. True Reliability in the Face of AI Deepfakes.

Under most legal regimes a base requirement of authentication is needed, some external evidence that a particular evidence artifact is what it is claimed to be.⁴⁴ This may be by a witness with knowledge saying it is what it claims to be and evidence showing a system produces an accurate result, as well as other means of authentication.⁴⁵

Testimony by a witness with knowledge is a common authenticator. Examples of authentication relevant to deepfake evidence include:

⁴³ FED. R. EVID. 404(b)(3).

⁴⁴ *See, e.g.*, FED. R. EVID. 901.

⁴⁵ FED. R. EVID. 901(b)(1); FED. R. EVID. 901(b)(9) (“Evidence About a Process or System. Evidence describing a process or system and showing that it produces an accurate result”).

- (1) *Testimony of a Witness with Knowledge*. Testimony that an item is what it is claimed to be.
- (2) *Nonexpert Opinion About Handwriting*. A nonexpert's opinion that handwriting is genuine, based on a familiarity with it that was not acquired for the current litigation.
- (3) *Comparison by an Expert Witness or the Trier of Fact*. A comparison with an authenticated specimen by an expert witness or the trier of fact.
- (4) *Distinctive Characteristics and the Like*. The appearance, contents, substance, internal patterns, or other distinctive characteristics of the item, taken together with all the circumstances.
- (5) *Opinion About a Voice*. An opinion identifying a person's voice — whether heard firsthand or through mechanical or electronic transmission or recording — based on hearing the voice at any time under circumstances that connect it with the alleged speaker.⁴⁶

Yet if we are faced with deepfake fabrication, a risk is that the witness may similarly fabricate his or her testimony to authenticate the deepfake. It then falls upon the engine of cross-examination to attack the weight of that testimony as to challenge the probative value of the deepfake evidence.⁴⁷ That cross-examination- where an opposing party questions a witness as to aspects and foundations of their claims - may be bolstered by technical data, but the novelty and technical excellence of deepfake technology may weaken the ability to use such expert testimony.

Other principles of reliability similarly place base requirements of reliability on admission of evidence, in particular the rules regarding the use of hearsay evidence.⁴⁸ These engage the truth-testing requirements of confrontation as to test evidence

⁴⁶ FED. R. EVID. 901(b)(1)-(5).

⁴⁷ See Rebecca A. Delfino, *Deepfakes on Trial: A Call To Expand the Trial Judge's Gatekeeping Role To Protect Legal Proceedings from Technological Fakery*, 74 HASTINGS L.J. 293 (2023); see also McPeak, *supra* note 20 (discussing how a more rigorous approach to admitting and authenticating evidence should be used in regards to Deepfakes).

⁴⁸ FED. R. EVID. 801.

mandated by the Sixth Amendment to the United States Constitution.

Digital evidence, with its electronic provenance, ease of manipulation and fabrication raises the stakes for confrontation. Though courts have excluded digital evidence⁴⁹, especially from online social media, they have also permitted its use.

The Third Circuit, excluding such evidence, in *United States v. Browne* found authentication failed where the government never established the defendant was the one who authored the Facebook communications nor otherwise authenticated the contents beyond the Facebook custodian's certification.⁵⁰ The Seventh Circuit in *Griffin v. Bell* upheld exclusion of a video and pictures for failure to authenticate those pictures.⁵¹ In *United States v. Vayner*, the Second Circuit found the mere fact that a social media page existed with defendant's name and photograph was insufficient to authenticate that printout under Rule 901(a).⁵² In *United States v. Jackson*, the Seventh Circuit barred evidence of website postings because the party attempting to oust the postings failed to show that the sponsoring organization of the website has actually posted the statements itself, as opposed to an unrelated third party.⁵³

Indeed, the Federal Committee on Rules of Practice and Procedure noted "Committee members unanimously agreed that Rule 803(16) was problematic, as it was based on the false premise that authenticity of a document means that the assertions in the document are reliable — this is patently not the case."⁵⁴

Similar issues of reliability are found with hearsay, "a statement, other than one made by the declarant while testifying at the trial or hearing, offered in evidence to prove the truth of the matter asserted."⁵⁵ With a hearsay statement there are few or no

⁴⁹ See SEAN E. GOODISON ET AL., DIGITAL EVIDENCE AND THE U.S. CRIMINAL JUSTICE SYSTEM 23 (2015), <https://www.ojp.gov/pdffiles1/nij/grants/248770.pdf> [<https://perma.cc/WV88-8DYE>] ("Some courts are skeptical of digital evidence due to uncertainties about chain of custody and validity of information obtained from devices.").

⁵⁰ *United States v. Browne*, 834 F.3d 403 (3d Cir. 2016).

⁵¹ *Griffin v. Bell*, 694 F.3d 817 (7th Cir. 2012).

⁵² *United States v. Vayner*, 769 F.3d 125 (2d Cir. 2014).

⁵³ *United States v. Jackson*, 208 F.3d 633, 638 (7th Cir. 2000).

⁵⁴ William K. Sessions, *Report of the Advisory Committee on Evidence Rules* (Nov. 15, 2014), http://www.uscourts.gov/sites/default/files/fr_import/ST2015-01.pdf [<https://perma.cc/KPG8-JEED>].

⁵⁵ See FED. R. EVID. 801(c).

ways to test the reliability of the evidence nor may there be other indicia supporting its reliability.⁵⁶ For example, the Sixth Circuit in *United States v. Martinez*, held that a video was improperly admitted as evidence where it purported to demonstrate proper medical procedures as it was a “statement” offered for the “truth of the matter asserted” under Rule 801.⁵⁷

Yet despite these concerns with the authenticity issues and hearsay, electronic evidence may be admitted with little further support as to reliability. In *United States v. Farrad*, the Sixth Circuit found online Facebook social media photos of contraband properly authenticated from related identifying evidence in the photographs themselves.⁵⁸ It discussed this in light of *United States v. Thomas*, and the hearsay prohibition.⁵⁹

Although it did not discuss the business-records exception, the closest analog from our precedents is *United States v. Thomas*—a case that itself illustrates that the business-records exception was not necessary to admit the photos here.⁶⁰ The defendant in that case, Jabron Thomas, was charged with armed bank robbery, and he was identified at trial in part through photos obtained from Meta social media platforms Facebook and, Instagram.⁶¹

He argued that the photos “could not be authenticated and were thus inadmissible in part because [the investigator who obtained them] admitted that he did not know who created the Facebook page or whether the Facebook page itself was authentic.”⁶² We turned that argument aside, however, noting that there was simply enough evidence presented “to support a finding that the [photograph] [was] what the proponent claim[ed] it is.”⁶³ The photos “appeared to show Thomas with distinctive tattoos” and clothing, which was enough given that “the government was not seeking to authenticate Jabron Thomas’s [or any] Facebook page or Instagram page (nor any of the factual information contained therein, such as Thomas’s workplace)[,] and the government was

⁵⁶ *Id.*; see also [article].

⁵⁷ 588 F.3d 301, 311 (6th Cir. 2009).

⁵⁸ 895 F.3d 859, 878 (6th Cir. 2018).

⁵⁹ *Id.*

⁶⁰ 701 F. App’x 414 (6th Cir. 2017).

⁶¹ *Id.* at 418.

⁶² *Id.* at 419.

⁶³ *Id.* (quoting FED. R. EVID. 901(a)).

not even necessarily presenting the photographs as ‘pictures of Jabron Thomas’—the jury was free to consider the photographs as identifying Thomas or not.”⁶⁴ Given that there was sufficient evidence that the photos were what they were claimed to be, in short, we ruled that there was no abuse of discretion in admitting them.⁶⁵

The Sixth Circuit did note that in sustaining the admission of social media images that “[n]o specific evidence was shown to suggest that the photographs were not “accurate representation[s] of the scene depicted.”⁶⁶ But this seems to push back the burden of proof on a *defendant*, requiring a challenge to authenticity supported by evidence contravening authenticity. It did find, in accord with other federal circuit courts that the business records exception would not authenticate the Facebook records as there was no engagement by Facebook to assure the accuracy of the postings.⁶⁷

The growth in the use of online and other digital media creates new dangers as to the establishment of authenticity, dangers that grow with AI fabrication resources. Lynch suggests that the means to remedy the authenticity challenge is through:

- examination of object provenance for trust and reliability and completeness, such as with challenges in the chain of custody of evidentiary objects,
- examination using forensic and diplomatic techniques of the object to verify its content (data) and metadata support that it is what it claims to be,
- examination and forensic/diplomatic evaluation of the signatures and seals that assert authenticity and integrity, and

⁶⁴ *Id.*

⁶⁵ *Id.* at 419-20.

⁶⁶ *United States v. Farrad*, 895 F.3d 859, 878 (6th Cir. 2018) (citing *United States v. Hobbs*, 403 F.2d 977, 979 (6th Cir. 1968)).

⁶⁷ *Id.* at 878-80 (citing *United States v. Browne*, 834 F.3d 403, 409 (3d Cir. 2016), *cert. denied*, 137 S. Ct. 695 (2017); *United States v. Gurr*, 471 F.3d 144, 151-52 (D.C. Cir. 2006); *United States v. Jackson*, 208 F.3d 633, 637-38 (7th Cir. 2000); *United States v. Barnes*, 803 F.3d 209, 217 (5th Cir. 2015)).

- examination and comparison of a challenged object with other copies of the object which may have validated integrity and provenance.⁶⁸

The key to authentication are links between the digital evidence and its authenticity that may be intrinsic to the evidence itself; external testing may not be available. Each step in the Interim Measures for Generative AI proposed by the Cyberspace Administration of China requires technical validation through the watermarking of digital objects to link them to evidence of its provenance.⁶⁹ Yet while this may support admissibility, it is not exclusive and may be circumvented via other methods.

Authentication faces its gravest challenge via AI image/audio/video technology, especially through the growth and ease of creating deepfake imagery and manipulation, where a text string may create artifacts of events that never existed.

H. Threats Abound

The explosion of ChatGPT and like systems for analysis and creation shows how popular their use is. But the growing body mishaps and “hallucinations” demonstrate the risks to subjects of those systems. The injuries from AI may appear in many forms, ranging from errors in information to mistakes in control systems. AI-mediated facial recognition systems could wrongly identify an

⁶⁸ Clifford Lynch, *Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust*, COUNCIL ON LIBR. & INFO. RES., <https://www.clir.org/pubs/reports/pub92/lynch/> [<https://perma.cc/5RSV-QK69>] (last visited Apr. 10, 2024).

⁶⁹ The AI laws of the Peoples Republic of China address this through requirements of digital watermarking of images. See *Interim Measures for the Management of Generative Artificial Intelligence Services*, CYBERSPACE ADMIN. OF CHINA (Jul. 10, 2023), <https://www.chinalawtranslate.com/en/generative-ai-interim/> [<https://perma.cc/4BCY-9AK4>] (“Article 8: When manual tagging is conducted in the course of researching and developing generative AI technology, the providers shall formulate clear, specific, and feasible tagging rules that meet the requirements of these Measures; carry out assessments of the quality of data tagging, with spot checks to verify the accuracy of tagging content; and conduct necessary training for tagging personnel to increase their awareness of legal compliance and oversee and guide them to carry out tagging efforts in a standardized way.”....“Article 12: Providers shall label generated content such as images and video in accordance with the Provisions on the Administration of Deep Synthesis Internet Information Services.”).

innocent person as the perpetrator of a crime. An AI-mediated system for ruling on a defendant may find an innocent person has committed a crime, leading to criminal prosecution, detention, and seizure of assets. Analysis of the growing databases on everything in the lives of others can destroy notions of any real privacy. These can lead to injuries and harms to people, from inconvenience to homicide.

In *Project Veritas v. Schmidt*, the dissent argued that permitting such unannounced recordings increased the risk of deepfake fabrications, which the majority dismissed as such fabrication could be addressed through tort actions by the victims of the fabrication and, in any event, could be done using announced recordings.⁷⁰ The dissent responded that the majority "... [missed] the critical point: once a person has notice that her conversation will be recorded, she can choose not only what to say, but also whether to speak at all."⁷¹

Citing Chesney and Citron,⁷² the dissent explained these risks further, noting the scope of the threat from AI technologies:

The secret recording of speech is far more destructive to one's privacy than merely having oral communications heard and repeated. Recorded speech can be stored indefinitely, disseminated widely, and viewed repeatedly. In the age of the internet and generative artificial intelligence (AI), surreptitious recording of in-person conversations risks massive and ongoing invasions of privacy. Today, anyone can access and learn how to use AI powered generative adversarial networks to create convincing audio or video "deepfakes" that make people appear to say or do things they never actually did. With these tools, "the only practical constraint on one's ability to produce a deepfake [is] access to training materials—that is, audio and video of the person to be modeled." The importance of the right to have notice before one's oral communications are recorded cannot be overstated because technology now allows

⁷⁰ 72 F.4th 1043 (9th Cir. 2023) (Christen, J., dissenting).

⁷¹ *Id.*

⁷² Robert Chesney & Danielle Citron, *Deepfakes and the New Disinformation War*, Foreign Affairs (Dec. 11, 2018), <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakesand-new-disinformation-war> accessed 8/8/2023 [<https://perma.cc/6KR2-A5M6>].

recordings to be selectively edited, manipulated, and shared across the internet in a matter of seconds.⁷³

In a prophetic comment on the challenges in the authentication of deepfake artefacts, the dissent addressed the challenge of authentication of audio recordings. It noted “the self-authenticating character of audio recordings is rapidly eroding as modern technology renders ‘deepfakes’ ever more accessible and difficult to distinguish from actual recordings.”⁷⁴ Grimm raised similar concerns as to the adequacy of the Federal Rules of Evidence to vet AI generated evidence, including as to relevance, authenticity and the use of expert testimony.⁷⁵

Grimm suggested that lawyers and judges faced with AI evidence begin with questions to help them answer the admissibility issues:

1. What problem was the AI created to solve, as to help assess accuracy of output, reliability and if its use conforms to its purpose,
2. How was the AI developed, and by whom, as to begin evaluation of the competence, biases, and motivations of the developers that may lead to questions about reliability?
3. Was the validity and reliability of the AI sufficiently tested, and what were the testing protocols invoked?
4. Is the manner in which the AI operates “explainable” so that it can be understood by counsel, the court, and the jury, and be subject to validation under the rules of evidence?
5. What are the risks of harm if the AI evidence of uncertain trustworthiness is admitted as evidence?
6. Timing Issues – given the complexity of these issues, they should be resolved during pretrial if possible, with more

⁷³ *Project Veritas*, 72 F.4th at 1075 (alteration in original).

⁷⁴ *Id.* at 1083, n.16 (citing Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1755-68 (2019)).

⁷⁵ Paul W. Grimm et al., *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. 9 (2021) (Discussing the validity and reliability of AI evidence in deciding whether it should be admitted in civil or criminal cases).

detailed disclosure prior to trial than currently required under the rules of evidence.⁷⁶

Answering these questions as to test the validity of AI evidence will be essential for fair and just determinations of the issues at hand. The time to address these risks is before harm is done. AI systems will only expand into more and more aspects of the human sphere. This, in turn, brings more risks from them as they create more and more artifacts of evidence in litigation of all types.

CONCLUSION

It is crucial that the authenticity challenge presented by the ever-expanding availability of deepfake technology be addressed sooner rather than later. The use of these techniques to injure others either personally or in their reputations is terrible. But the further weaponization of the fake imagery can do greater and greater harm to all institutions, as well as punishing the innocent at the behest of the guilty. We must act now on a course of action to address the risks of deepfakes before its corruption of the judicial, political and personal undermines our faith in them all.

⁷⁶ *Id.* at 97-105.