DEEP FAKE OUT

Mihailis E. Diamantis,[1] Sean Sullivan,[2] and Eli Alshanetsky[3]

Deepfakes are visual and audio media that use artificial intelligence to portray people saying things they never said, doing things they never did, and experiencing events that never happened. They can be trivial ("Tom Cruise knows magic tricks?"), outlandish ("Why is Nancy Pelosi drunk on national television?"), or even dangerous ("Run, the Hollywood sign is burning!"). Because deepfakes can be both persuasive and pervasive, many commentators fear that humanity will soon take another step into the post-truth abyss.

This Article evaluates the threat deepfakes pose to truth by anticipating how they will impact the area of law most directly concerned with truth: the law of evidence. Deepfakes present an obvious challenge to the administration of justice in modern courtrooms, where audiovisual evidence plays an important role. Solutions offered in past legal scholarship—like relying on experts to identify deepfakes or criminalizing deepfake production—are superficial. They optimistically assume that deepfakes will always have a tell. To truly appreciate the threat deepfakes pose, the law must brace itself for the likely prospect of what this Article calls "deepest" fakes, which will be indistinguishable in every respect from authentic media.

Drawing on tools from philosophy, legal history, and technology studies, this Article demonstrates how evidence law can and likely will adapt to a world saturated with deepest fakes. Courts have long encountered the sort of philosophical skepticism that deepest fakes threaten today. In essence, the existence of deepest fakes reduces the truth value of digital media to the level of oral testimony, creative works, other easily falsifiable evidence. While deepest fakes do raise serious problems, trial procedure only needs to change if existing safeguards fail. This Article finds that deepest fakes present no different challenge for modern courts than oral testimony, paintings, and photographs presented for their early twentieth century counterparts. The safeguard then, as now, is a nuanced adversarial process that, refusing to take evidence at face value, probes each submission with contextual indicators of reliability. What emerges is an empowering picture in which human judgment, rather than blind trust in media of any sort, is the ultimate arbiter of truth.

[1] Ben V. Willie Professor of Excellence, University of Iowa College of Law.

[2] Bouma Faculty Fellow in Law, University of Iowa College of Law.

[3] Assistant Professor in Philosophy, Temple University.

Working paper for Mihailis E. Diamantis, Sean P. Sullivan, & Eli Alshanetsky, Deep Fake Out, 94 Geo. Wash. L. Rev. (forthcoming).

Don't believe everything you read on the internet.
– Abraham Lincoln (circa 1864)

I. Introduction

Deepfakes splashed into the 2024 election like never before.[4] Residents in New Hampshire received robocalls in which President Biden urged them not to turn out for the election.[5] Digital images showed Taylor Swift, dressed as Uncle Sam,

[4] Sarah Jeong, *The AI-Generated Hell of the 2024 Election*, THEVERGE (Sept. 15, 2024), https://www.theverge.com/policy/24098798/2024-election-ai-generated-disinformation. Perhaps surprisingly, we have it relatively easy here in the U.S. In the U.K., there are claims that politicians themselves are deepfakes. Mia Sato, *The UK Politician Accused of Being AI Is Actually a Real Person*, TheVerge (July 9, 2024), https://www.theverge.com/2024/7/9/24195005/reform-uk-candidate-election-ai-bot-mark-matlock.

[5] Lauren Feiner, *Telecom Will Pay $1 Million over Deepfake Joe Biden Robocall*, THEVERGE Aug. 21, 2024), https://www.theverge.com/2024/8/21/24225435/lingo-telecom-biden-deepfake-robocall-fcc-fine.

endorsing Donald Trump.[6] A photo on X showed an old photo of Trump groping a minor with sex-offender Jeffrey Epstein.[7] Biden in military fatigues.[8] Trump embraced by Black voters.[9]

Of course, none of it was real. The consultant behind the robocalls faced criminal charges and a hefty fine.[10] Swift later endorsed Harris.[11] The other images were all debunked. But even in their absence, deepfakes still drove news, as when Trump falsely claimed that photos of Vice President Harris' huge rally crowds were fake.[12]

Deepfakes are audiovisual media that use deep learning to seamlessly stitch together faces, voices, and other elements into highly realistic representations.[13] They are "fake" at two levels: their content (they portray events that never happened) and their presentation (they deceptively appear to be traditional recordings captured by mechanical devices like cameras). Deepfakes first gained attention on Reddit in 2017 as pornographic videos that swapped celebrities in for the true actors.[14] Shortly after, Lyrebird debuted, giving people "a way to recreate anyone's voice and get it to say almost anything."[15] A visual media program called FakeApp launched the same year

[6] Shannon Bond, *How AI-Generated Memes are Changing the 2024 Election*, NPR (Aug. 30, 2024), https://www.npr.org/2024/08/30/nx-s1-5087913/donald-trump-artificial-intelligence-memes-deepfakes-taylor-swift.

[7] Aleskandra Wrona, *Does Pic Shoe Trump and Epstein with Minor Girl*, Sopes (Jan. 9, 2024), https://www.snopes.com/fact-check/epstein-trump-young-girl-photo/.

[8] Bill McCarthy, *Image of Biden Planning Military Action in Fatigues Is Fake*, AFP Fact Check (Apr. 29, 2024), https://factcheck.afp.com/doc.afp.com.34H74GF.

[9] Marianna Spring, *Trump Supporters Target Black Voters with Faked AI Images*, BBC (Mar. 3. 2024), https://www.bbc.com/news/world-us-canada-68440150.

[10] Shannon Bond, *A Political Consultant Faces Charged and Fines for Biden Deepfake Robocalls*, NPR (May 23, 2024), https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative; *In the Matter of Lingo Telecom, LLC*, Order, File No.: EB-TCD-24-00036425 (Fed. Comm. Comm'n Aug 21, 2024).

[11] Chloe Veltman, *Taylor Swift Has Endorsed Kamala Harris for President—Will It Matter?*, NPR (Sept. 11, 2024), https://www.npr.org/2024/09/11/nx-s1-5108695/taylor-swift-endorsement-kamala-harris.

[12] Jude Jofe-Block, *Why False Claims That a Picture of Kamala Harris Rally Was AI-Generated Matter*, NPR (Aug. 14, 2024), https://www.npr.org/2024/08/14/nx-s1-5072687/trump-harris-walz-election-rally-ai-fakes.

[13] Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99, 99–100 (2019); 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 (2019).

[14] 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, 299 (2023).

[15] *New Software Can Mimic Anyone's Voice*, NPR (May 5, 2017), http://www.npr.org/2017/05/05/527013820/.

with the explicit goal of "mak[ing] deepfake[] technology available to people without a technical background or programming experience."[16] Since then, the beneficial and nefarious uses of this technology have been limited only by users' imaginations.

Deepfakes leave many people feeling rattled. Of course, there are obvious harmful applications for deepfakes, like stealing people's identities, creating unauthorized pornography, and engaging in fraudulent transactions. But some commentators foretell of a broader, more structural threat. According to them, deepfakes could sow social discord: "A well-timed and thoughtfully scripted deep fake could . . . tip an election, spark violence in a city primed for civil unrest, . . . or exacerbate political divisions in a society."[17] More worryingly, deepfakes could "tear the very fabric of democracy":[18] "The informational anarchy and paranoia [that deepfakes cause] might [challenge] individual decision making or collective self-rule."[19] Finally and most extreme, deepfakes could wage "war on what is real":[20] "They raise existential questions about reality on a profound and metaphysical level."[21] For some commentators, the "war" is not a metaphor; they suggest that certain deepfakes could warrant "a military response."[22]

Our starting observation is this: Statements like "deepfakes raise existential questions about reality"[23] or "people [can] no longer believe *anything* is real [if they cannot trust digital media]"[24] conflate reality with audiovisual representations of it. Of course, we can only know what is real if we have evidence of it, and digital photos and videos are one source of evidence. In a world where the average adult spends

[16] *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

[17] Robert Chesney & Danielle Citron, *Disinformation on Steroids: The Threat of Deep Fakes*, COUNCIL ON FOREIGN RELATIONS (Oct. 16, 2018), https://www.cfr.org/report/deep-fake-disinformation-steroids.

[18] Bobby Chesney and Danielle Citron, *Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?,* LAWFARE (Feb. 21, 2018), https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy.

[19] Marc Jonathan Blitz, *Lies, Line Drawing, and (Deep) Fake News*, 71 OKLA. L. REV. 59, 109 (2018).

[20] Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 8 (2020).

[21] Delfino, *supra* note 14, at 345.

[22] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1809 (2019)

[23] Rebecca A. Delfino, *The Deepfake Defense—Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers*, 84 OHIO ST. L.J. 1067, 1081 (2024)

[24] Agnieszka McPeak, *The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood*, 23 VAND. J. ENT. & TECH. L. 433, 439 (2021)

most of their waking life looking at a screen,[25] it may be easy to get reality and the projection of it mixed up. We agree that deepfakes present "a fundamental challenge to the [existing] information environment."[26] But, it is important to recognize that digital media are only one source of evidence about what's out there, and they have not even been around for that long. While it may be hard to remember a time before everyone had Snapchat in their pocket, digital cameras were not commercially available until 1990[27] (having been first invented fifteen years prior).[28]

We (a law and tech scholar, an evidence scholar, and an epistemologist) defend both a thesis and an anti-thesis in this Article. The thesis is that deepfakes pose even greater problems than commentators envision. The anti-thesis is that existing social, legal, and metaphysical institutions for responding to other creative forms of lying are more up to the task than commentators realize. Epistemology and the law of evidence help us to recall the various ways that humans guard against the possibility of deception and the various ways that reality reveals itself to skeptical audiences, even when once-reliable sources of information cease to be trustworthy.[29] Indeed, we predict that deepfakes will ultimately empower people for the very reasons that others see them as a threat. By undermining the reliability of digital media, deepfakes reaffirm the authority of, and our reliance on, human judgment.

To make our case, we turn to the law's most truth-focused institution: the courtroom. "The courtroom is a microcosm of society in general."[30] By assessing the impact deepfakes will have on courts' truth finding mission, we glean lessons for society more generally. Deepfakes are already appearing as evidence in trial,[31] and scholars are starting to propose various measures to preserve courts' integrity (Part

[25] Brian Stelter, *8 Hours a Day Spent on Screens, Study Finds*, N.Y. TIMES (Mar. 26, 2009); Jacqueline Howard, *Americans Devote More Than 10 Hours a Day to Screen Time, and Growing*, CNN (July 29, 2016), cnn.com/2016/06/30/health/americans-screen-time-nielsen/index.html,
People Staff, *Average U.S. Adult Will Spend Equivalent of 44 Years of Their Life Staring at Screens: Poll*, PEOPLE MAG. (June 3, 2020), people.com/human-interest/average-us-adult-screens-study/.

[26] Mark Corcoran & Matt Henry, *The Tom Cruise Deepfake that Set Off 'Terror' in the Heart of Washington DC*, ABC NEWS AUSTL. (June 27, 2021, 7:16 PM), https://www.abc.net.au/news/2021-06-24/tom-cruise-deepfake-chris-ume-security-washington-dc/100234772 (quoting former CIA agent Matt Ferraro).

[27] Lauren Cabral, *The History of Digital Cameras*, BACK THEN HISTORY (July 26, 2023), https://www.backthenhistory.com/articles/the-history-of-digital-cameras.

[28] Joanna Goodrich, *The First Digital Camera Was the Size of a Toaster*, IEEE SPECTRUM (Apr. 6, 2022), https://spectrum.ieee.org/first-digital-camera-history.

[29] As the Federal Rules of Evidence say, their goal is to help courts with "the end of ascertaining the truth" from available information. FED. R. EVID. 102.

[30] Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 PUB. INTEREST L.J. 245, 257 (2020).

[31] Matt Reynolds, *Courts and Lawyers Struggle with Growing Prevalence of Deepfakes*, AM. BAR ASS'N J. (June 9, 2020, 10:29 AM), https://www.abajournal.com/web/arti- cle/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes.

II). Existing recommendations assume that there will always be some sophisticated way to tell deepfakes and genuine media apart. But they would still leave courts vulnerable to what this Article calls "deepest fakes"—deepfake media that technologists predict will be costless to make and will perfectly mimic genuine media.[32] The prospect of deepest fakes places the threat of deepfakes, both to courtrooms and to broader institutions, in starkest terms. Any approach that solves the problem of deepest fakes in courtrooms could, *a fortiori*, solve the lesser epistemic problem of deepfakes in the wild.

For epistemologists, skeptical challenges to our grip on reality are as old as the discipline itself (Part III). Epistemology offers conceptual tools for evaluating sources of evidence and overcoming skepticism. Many of these tools have analogues in evidence law (Part IV). The constant possibility that any photo, video, or audio recording could be faked recalls the historic struggles of courts to separate truthful from dishonest testimony when accounts diverged, and any witness could be lying. We anticipate that courts will respond as they have in the past (Part V). Minimal procedural adjustments may help, but the key safeguard against both deepfakes and deepest fakes has to be a robust adversarial process that provides jurors not only with digital media evidence, but with the context factors that bear on its veracity. The most important change will not occur within courts, but within jurors as they become more astute judges of media evidence. The implications for courts and broader society are empowering (Part VI). As we collectively learn that we can no longer reflexively trust digital media, our own evolving epistemic practices will grow to take on a more central truth-finding role. We predict that this development will only strengthen humans' relationship with reality.

II. The Real Problem of Deepfakes

Before building out our anti-thesis, this Part starts by presenting the threat deepfakes pose in its starkest terms. We think that threat is greater than many realize. Legislative attempts to ban deepfakes in certain contexts (Section II.A) will not keep them from infiltrating the law's most truth-oriented institution: the courtroom (Section II.B). While a handful of evidence scholars have taken note (Section II.C), their solutions all rest on the misplaced optimism that deepfakes will always have some hidden tell that reveals them for what they are. Technologists are not so sanguine (Section II.D). Any framework for responding to deepfakes must be resilient enough to withstand the eventuality of undetectable deepest fakes.

[32] Marie-Helen Maras & Alex Alexandrou, *Determining Authenticity of Video Evidence in the Age of Artificial Intelligence in the Wake of Deepfake Videos*, 23 INT'L J. EVID. & PROOF 255, 257 (2018) ("Put simply, when AI technology is used in the future, it may be impossible to determine that the video is fake.").

A. The Law of Deepfakes

The First Amendment presents a significant barrier to legislation that prohibits deepfakes. "Deep fakes . . . are generally video or audio creations, and such creations have typically been considered a form of expression."[33] Deepfake legislation, by necessity, targets digital media that is fake. But, as the Supreme Court opined fifty years ago, "Under the First Amendment, there is no such thing as a false idea."[34] If deepfakes were an unmitigated social blight—like blackmail and fraud—the Constitution might not present a barrier to prohibiting them. Part of the problem from a First Amendment perspective is that the same technology used to produce deepfakes can be put to honest and valuable uses, like portraying deceased actors in movie sequels [35] or enabling brain and throat cancer survivors to continue communicating in their own voice.[36]

If a deepfake involves a matter of public concern, then the government could only proscribe it if it falls into a historically unprotected category of speech or expression.[37] There have been sporadic legislative efforts to control deepfakes in some particularly problematic contexts. The only federal legislation[38] on point is the National Defense Authorization Act of 2019, which requires annual assessments of foreign efforts to weaponize deepfakes or to use them for election interference.[39] Some states have also passed limited criminal statutes,[40] like Virginia's law[41] against

[33] Blitz, *supra* note 19, at 62.

[34] Gertz v. Robert Welch, Inc., 418 U.S. 323, 339 (1974).

[35] Peter Suciu, *Deepfake Star Wars Videos Portent Ways the Technology Could be Employed for Good and Bad*, Forbes (Dec. 11, 2020), https://www.forbes.com/sites/petersuciu/2020/12/11/deepfake-star-wars-videos-portent-ways-the-technology-could-be-employed-for-good-and-bad/.

[36] Brooke Steinberg, *I lost My Voice Because of a Tumor—but an AI Clone Gave It and My Confidence Back to Me*, N.Y. POST (May 13, 2024), https://nypost.com/2024/05/13/lifestyle/i-lost-my-voice-because-of-a-tumor-an-ai-clone-gave-it-back-to-me/.

[37] United States v. Alvarez, 567 U.S. 709 (2012).

[38] Other federal legislation has been proposed. For example, in 2018, Senator Ben Sasse introduced a bill criminalize certain harmful deepfakes, Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (2018), and in 2019, Representative Yvette Clark introduced the DEEPFAKES Accountability Act, which would have required deepfakes to carry watermark, Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, 116th Cong. (2019). Both bills died in committee.

[39] National Defense Authorization Act for Fiscal Year 2020, S. 1790, 116th Cong. (2019)

[40] International efforts are underway too. For example, Germany has made it a crime to create a deepfake that violates personal rights. §201a of the German Criminal Code.

[41] VA. CODE. ANN. § 18.2-386.2 (2020); *see also* A.B. 602, 2018-2019 Leg. Sess. (Ca. 2019) (providing a private right of action to victims of deepfake pornography)

deepfake revenge porn or Texas'[42] and California's[43] laws against using deepfakes to influence elections. Whether these state statutes will survive First Amendment scrutiny remains uncertain. One commentator has argued that statutes prohibiting deepfake pornography are unconstitutional.[44] Challenges to Texas' and Virginia's election statutes will surely arise in the litigious wake of the recent election cycle.

Even if laws that specifically target deepfakes are on constitutionally shaky ground, several existing speech-neutral statutes already offer civil and criminal remedies for objectionable acts that could (but needn't) involve deepfakes. As others have noted, laws in most states cover a range of harmful deepfake use cases, e.g. to defame, to intentionally inflict emotional distress, to impersonate another, to cyberstalk, etc.[45] In a similar vein, Virginia's prior revenge porn statute applies to the unauthorized and malicious distribution of pornographic content "created *by any means whatsoever* that depicts another person."[46] Its broad language would seemingly include deepfake revenge porn, even without the later clarifying amendment: "'another person' includes a person whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person."[47] New Jersey is considering,[48] but has yet to pass, its own deepfake statute. As with Virginia, New Jersey's existing revenge porn statute is arguably broad enough to encompass deepfakes and neutral enough to satisfy the First Amendment. It applies if someone "reproduces *in any manner* the image of another person whose intimate parts are exposed . . . without that person's consent.[49]

B. Deepfakes in the Courtroom

Commentators are rightly skeptical of the effectiveness of legislation aimed at proscribing deepfakes. Statutory bans may provide some recourse for victims, but they will not stem the creation of deepfakes themselves. "We may safely assume that

[42] TEX. ELEC. CODE § 255.004(e) (prohibiting use of "a deceptive video with intent to influence the outcome of an election").

[43] CAL. ELEC. CODE § 20010.

[44] Bradley Waldstreicher, *Deeply Fake, Deeply Disturbing, Deeply Constitutional: Why the First Amendment Likely Protects the Creation of Pornographic Deepfakes*, 24 CARDOZO L. REV. 729 (2021).

[45] Project Veritas v. Schmidt, 72 F.4th 1043, 1062 n.15 (9th Cir. 2023), *reh'g en banc granted, opinion vacated*, 95 F.4th 1152 (9th Cir. 2024) ("[V]ictims of [deepfake] fabrications can vindicate their rights through tort actions."); Chesney & Citron, *supra* note 22, at 1792-1804.

[46] VA. CODE. ANN. § 18.2-386.2(A) (2020) (emphasis added)

[47] *Id.*

[48] State of N.J., Senate no. 976, 221st Legis. (2024), https://pub.njleg.state.nj.us/Bills/2024/S1000/976_I1.PDF.

[49] N.J. STAT. ANN. § 2C:14-9(b)(1) (emphasis added).

the ready availability of deepfake tools, and antisocial uses thereof, will continue irrespective of how the law may attempt to contain, regulate, and punish them."[50]

One emerging antisocial use of deepfakes is to manipulate evidentiary records at trial. "[O]ur legal system is as vulnerable to content manipulation as any other area of civic life"[51] Indeed, while "[i]t is often illegal to make false statements where government needs honest answers to questions,"[52] courtrooms are already feeling the influence of deepfakes.

Deepfakes present two challenges to courts' factfinding mission. The first is obvious: parties may seek to introduce deepfakes as evidence that supports their case. "A video of the crime scene could be manipulated by the perpetrators to change their appearance; an audio recording could be manipulated to depict somebody as violent; a criminal could swap their face with somebody else's to create a perfect alibi; an innocent could be framed for revenge."[53] For example, in one U.K. case, a mother used deepfake audio recordings at a custody hearing to give a false impression that her ex-husband abused their children.[54] Of course, U.S. courts can regulate false content at trial without raising First Amendment concerns.[55] To date, no reported case in the United States has found that evidence entered into the record was a deepfake.[56] One dissenting judge did predict that police could use deepfakes in extra-judicial proceedings to dupe suspects into waiving procedural rights.[57] But the slim official record of deepfakes in evidence does not necessarily mean that deepfakes have yet to infiltrate courtrooms. It only means that they are rarely caught: "[I]t would never occur to most judges that deepfake material could be submitted as evidence."[58] It is only a matter of time.

[50] Pfefferkorn, *supra* note 30, at 253.

[51] Delfino, *supra* note 23, at 1076.

[52] Blitz, *supra* note 19, at 66-67.

[53] Francesca Palmiotto, *Detecting Deep Fake Evidence with Artificial Intelligence: A Critical Look from a Criminal Law Perspective* (March 10, 2023), *available at* https://ssrn.com/abstract=4384122.

[54] Matt Reynolds, *Courts and Lawyers Struggle with Growing Prevalence of Deepfakes*, AM. BAR ASS'N J. (June 9, 2020), https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes

[55] United States v. Alvarez, 567 U.S. 709, 718-22 (2012)

[56] A Westlaw search for ("deepfake" "deep fake") on all federal and state cases presently (September 26, 2024) returns only twenty cases.

[57] State v. Garrett, No. 124,329, 2024 WL 4245190, at *15 (Kan. Sept. 20, 2024) ("I fear it will not be long before law enforcement tests the limits of creating fabricated images of a detainee at the scene of the crime or artificially create other evidence in order to convince a suspect to forego their right to remain silent or cooperate with an investigation.").

[58] Patrick Ryan, *'Deepfake' Audio Evidence Used in UK Court to Discredit Dubai Dad*, NAT'L NEWS: UAE (Feb. 8, 2020), https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764 (quoting Byron James, a partner in the London-based law firm Expatriate Law).

Even once judges become aware of the possibility of deepfake evidence, deepfakes would begin to pose a second kind of challenge for courts: the existence of deepfakes can be used at trial to undermine the credibility of legitimate evidence. "Th[is] 'deepfake defense' is built around the premise that the audiovisual material introduced as evidence against the defendant is claimed to be fake."[59] The deepfake defense has appeared in several reported cases in the United States, though judges presently seem to view it with skepticism.[60] Commentators agree "the very existence of deepfakes will [inevitably] complicate the task of authenticating real evidence."[61]

The apocalyptic concerns voiced by some deepfake scholars extend into the courtroom. "Deepfakes pose dangers and risks to our society and democratic institutions, including our judicial system."[62] Given the lurking threat of deepfakes, "video evidence may ultimately lose its persuasive power and, if taken far enough, degrade public trust in the very institution of the courts."[63] The fear is that if jurors cannot trust digital media, they may lose their grip on truth itself; and, if there is no truth, what are courts for? "If juries cease believing that the truth exists and that it can be found out, then they will have little cause to keep believing in the courts."[64]

C. Academic Proposals and "Deepfake Detectors"

Even though deepfakes are a relatively recent phenomenon, there is no shortage of proposals for what to do about them. Some have expressed a reserved confidence that, at least for the time being, "[w]hen deepfakes result in harm, there are a variety of [existing] laws that may apply to punish and provide restitution."[65] Existing laws could be particularly effective if they could be applied to the platforms

[59] Delfino, *supra* note 23, at 1070. The more general phenomenon of undermining media by claiming it could be fake has been dubbed the "liar's dividend." Chesney & Citron, *supra* note 22, at 1758.

[60] *See, e.g.*, 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 social media posts into evidence that defendant alleged included deepfake photos); Matter of Gabriel H., 229 A.D.3d 1048, 1051 (2024) ("Respondent . . . contends that the videos should be given little to no weight because they could be 'deepfakes.' The court afforded the videos great weight based on clear evidence of their reliability."); Pittman v. Commonwealth, No. 0681-22-1, 2023 WL 3061782, at *6 (Va. Ct. App. Apr. 25, 2023) ("[T]here is no evidence of or contention that would call into question the veracity of the video or the possibility of a 'deep fake.'").

[61] Pfefferkorn, *supra* note 30, at 255.

[62] Delfino, *supra* note 14, at 296.

[63] *Id.* at 312-13.

[64] Pfefferkorn, *supra* note 30, at 276.

[65] Brown, *supra* note 20, at 37.

through which many users publish deepfakes.[66] "[B]anning the technology altogether" might provide a longer term solution,[67] but, as already discussed, "it is unlikely that a flat ban on deep fakes could withstand constitutional challenge."[68] Constitutional law scholars have responded, offering more limited bans, e.g., on deepfakes that misrepresent their "purported source or vehicle," that may have a better chance of passing pass constitutional muster.[69]

Turning specifically to the courtroom, scholars have proposed to beef up procedures and rules of evidence to neutralize deepfakes. One idea for curbing the misuse of the deepfake defense is to amend the rules of procedure to allow courts to sanction attorneys who, in bad faith, question the authenticity of digital media during oral argument.[70] Most of the attention, however, has been devoted to the challenge of excluding deepfakes from evidence. Here, the authentication standard of Federal Rules of Evidence 901 plays a prominent role.[71] One scholar would "expand the gatekeeping function of the court by assigning the responsibility of deciding authenticity issues [for digital media] solely to the judge."[72] Another would require "a person whose occupation or means of knowledge is in a specialized field . . . to testify about [digital] evidence" where there are questions of authenticity.[73] A final proposal would ask "[j]udges and attorneys . . . to find the originator of a video or photo" and warns that "it may no longer be prudent to admit video evidence when the origin of a video is indeterminable."[74] Wide-spread adoption of self-certifying technology on media capture devices could augment this approach.[75]

Some evidence scholars believe that technological developments will save the day without a need to amend current law. "[C]ourts are confident in the processes

[66] Chesney & Citron, *supra* note 22, at 1795 ("[T]he most efficient and effective way to mitigate harm may be to impose liability on platforms."). As Chesney and Citron note, Section 230 of the Communications Decency Act presently forecloses this possibility. 42 U.S.C.A. § 230.

[67] *Id.* at 32.

[68] Chesney & Citron, *supra* note 22, at 1790.

[69] Blitz, *supra* note 19, at 64-65 ("Fake news may lose protection, I suggest, when it is not only a falsity, but a forgery as well.").

[70] Delfino, *supra* note 23, at 1071. Currently, Federal Rule of Civil Procedure Rule 11 only applies to signed writings, *id.* at 1092, and only in civil trials, *id.* at 1095.

[71] McPeak, *supra* note 22, at 440 ("[P]roper use of the authentication rules in the Federal Rules of Evidence can alleviate both concerns [with deepfakes]."). *See* FED. R. EVID. 901.

[72] Delfino, *supra* note 14, at 341.

[73] Molly Mullen, *A New Reality: Deepfake Technology and the World Around Us*, 48 MITCHELL HAMLINE L. REV. 210, 229 (2022).

[74] John Channing Ruff, *The Federal Rules of Evidence Are Prepared for Deepfakes. Are You?*, 4 REV. LITIG. 103 (2021).

[75] Delfino, *supra* note 14, at 341 ("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.").

they already have in place for excluding manipulated evidence. [These scholars] share that confidence."[76] As they observe, there is "a long history of fakery" in evidence, even where digital images were concerned.[77] Since the release of Photoshop in 1990, users have been able to alter digital images. But "[n]o major regulation or legislation was needed to prevent the apocalyptic vision of Photoshop's future; society adapted on its own."[78] In the same way, courts may develop "strategies for keeping deepfake videos out of evidence,"[79] relying on experts where needed[80] or using their own "training in spotting outward signs of altered deepfake technology."[81] Because "deepfakes are still generally not very good," significantly altering the Federal Rules of Evidence at this stage would be a "gross overreaction."[82]

All of the scholars discussed in this Section seem to assume that there is—and always will be—some way to distinguish deepfakes from genuine media. Banning deepfake content or asking platforms to police it can only be effective if there is some reliable way to tell when media are fake. For confronting deepfakes at trial, these scholars note that "[i]dentifying potentially deepfake content is just the first of the necessary steps" in the solutions they envision.[83] Evidence scholars recommend "a 'go-slow-and-strict' . . . future to allow for the development of better technologies that can detect deepfakes."[84] In other words, "[d]eepfake detectors [will be] indispensable."[85] As discussed next, technologists are not so optimistic about the long term prospects for such detectors.

D. The Enduring Challenge of Deepest Fakes

There is an active research community focused on developing sophisticated techniques for detecting deepfakes. They have found several "tells." Some are digital artifacts, unintended products of the technology used to make deepfakes.[86] For

[76] Pfefferkorn, *supra* note 30, at 266.

[77] *Id.* at 256.

[78] Jeffrey Westling, *Deep Fakes: Let's Not Go Off the Deep End,* TECHDIRT (Jan. 30, 2019), https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-go-off-deep-end.shtml.

[79] Pfefferkorn, *supra* note 30, at 259.

[80] *Id.* at 263.

[81] Mullen, *supra* note 73, at 224.

[82] John Channing Ruff, *The Federal Rules of Evidence Are Prepared for Deepfakes. Are You?*, 4 REV. LITIG. 103, 125 (2021).

[83] Brown, *supra* note 20, at 58.

[84] Delfino, *supra* note 14, at 316.

[85] Palmiotto, *supra* note 53, at 225 ("[T]echnology involving [deep fake] detection will become indispensable in a courtroom scenario"). 53

[86] John Spacey, *7 Types of Data Artifact*, SIMPICABLE (Apr. 16, 2017), https://simplicable.com/new/data-artifact.

example, there may be a discrepancy between the expected file size of a video and its actual size[87] or there may be subtle clues left by "'digital manipulations such as scaling, rotation or splicing,' that are commonly employed in deepfakes." [88] Researchers at MIT and the Department of Defense have taken a different approach, examining subtle biometric markers that deepfakes sometimes botch, like distorting micro details of a person's iris or failing to match a person's blood pulse in all parts of their body.[89] Other biomarkers include rates of eye-blinking[90] or small distortions in facial regions.[91]

There is just one problem with this approach—every new deepfake detector ultimately helps refine deepfake generators. Deepfake generators are made using generative adversarial networks.[92] These "are two-part AI models consisting of a generator that creates samples [of video, images, or audio] and a discriminator that attempts to differentiate between the generated samples and real-world samples."[93] The successes of the discriminator feed back into the training of the generator, continually improving its ability to produce realistic outputs. For example, "[t]he same deep-learning technique that can spot face-swap videos can also be used to improve the quality of face swaps—and that could make them harder to detect."[94]

[87] Kaveh Waddel, *The Impending War over Deepfakes*, LOS ALAMOS NAT'L LAB. (July 22, 2018), https://www.lanl.gov/discover/features/top-media-stories/top-science-2018-22.php.

[88] John Villasenor, *Artificial Intelligence, Deepfakes, and the Uncertain Future of Truth*, BROOKINGS TECHTANK (Feb. 14, 2019), https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/.

[89] Satya Venneti, *Real-Time Extraction of Biometric Data from Video*, CARNEGIE MELLON SOFTWARE ENG'G INST. (Aug. 22, 2016), https://insights.sei.cmu.edu/blog/real-time-extraction-of-biometric-data-from-video/

[90] Yuezun Li, Ming-Ching Chang & Siwei Lyu, *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking*, UNIVERSITY OF ALBANY, SUNY (June 11, 2018), *available at* https://arxiv.org/pdf/1806.02877.pdf.

[91] Francesca Palmiotto, *Detecting Deep Fake Evidence with Artificial Intelligence: A Critical Look from a Criminal Law Perspective* (March 10, 2023), *available at* https://ssrn.com/abstract=4384122 (internal citations omitted); Andreas Rössler et al. (2018) *FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces* (Mar. 24, 2018), https://arxiv.org/pdf/1803.09179.pdf; Emerging Technology, *This Algorithm Automatically Spots 'Face Swaps' in Videos*, MIT TECH. REV. (Apr. 10, 2018), www.technologyreview.com/s/610784/ this-algorithm-automatically-spots-face-swaps-in-videos/ (

[92] Ian J. Goodfellow et al., *Generative Adversarial Nets* (June 10, 2014) (Neural Information Processing Systems conference paper), https://arxiv.org/abs/1406.2661.

[93] 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/

[94] Rössler et al., *supra* note 91.

Indeed, shortly after researchers discovered the eye-blinking test as a tell for detecting deepfakes, deepfake generators had figured out how to defeat it.[95]

It is little surprise, then, that "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."[96] It is a game of cat-and-mouse, and, as in the Sunday morning cartoons, the mouse is always wilier. Research on deepfake detection will continue, but each new advancement requires creative effort and innovation. Deepfake generators, by contrast, only need one trick to adapt—retrain using the new detector's data.[97] Lay people cannot tell today's deepfakes apart from genuine media,[98] and even experts are having trouble.[99] "A variety of detection mechanisms exist, and they are improving. But they still lag behind the sophistication of deepfakes, which continue to advance."[100] As technologists bluntly put it, "The adversary will always win."[101]

A more promising technical approach may be to mechanically certify media as unaltered rather than attempting to expose deepfakes. For example, location verification "is available already, thanks to the ubiquity of phones with location

[95] 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, 1956-57 (2020).

[96] 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/; Hilke Schellmann, *The Dangerous New Technology that will Make US Question Our Basic Idea of Reality*, QUARTZ (Dec. 5, 2017), https://qz.com/1145657/the-dangerous-new-technology-that-will-make-us-question-our-basic-idea-of-reality ("[F]orensic specialists predict that computers will be able to generate convincing, fabricated audio and video recordings at a rapid pace in the next few years.").

[97] Louise Matsakis, *Artificial Intelligence is Now Fighting Fake Porn*, WIRED (Feb. 14, 2018) https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes ("If you really want to fool the system you will start building into the deepfake ways to break the forensic system.").

[98] 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.

[99] Than Thi Nguyen et al., *Deep Learning for Deepfakes Creation and Detection: A Survey,* 223 COMP. VISION & IMAGE UNDERSTANDING (2023).

[100] Brown, *supra* note 20, at 23; Harris, *supra* note 13, at 102 ("[T]his type of production carries immense potential to be indistinguishable from real-life videos."); Dan Boneh, Andrew J. Grotto, et al., *How Relevant is the Turing Test in the Age of Sophisbots?* at *3, ARXIV (Aug. 30, 2019), https://arxiv.org/pdf/1909.00056.pdf ("[I]n the long-run [deepfake detection] is likely to be a losing battle or at best a stalemate.").

[101] Dan Robitzski, *DARPA Spent $68 Million on Technology to Spot Deepfakes*, FUTURISM (Nov. 19, 2018), https://futurism.com/darpa-68-million-technology-deepfakes (quoting Dartmouth Professor Hany Farid).

tracking features as well as cell-site location records."[102] It could help demonstrate that the camera and the subject were in the same place at the same time. Additionally, some media capture devices now come equipped with cryptographic or similar certification processes, which could be used to verify that the media originated from the device in question.

Unfortunately, even sophisticated authentication is far from foolproof. Location verification is ineffective if GPS locations can be spoofed—a vulnerability that has existed for decades.[103] And device authentication provides little assurance when media can be altered on device. In 2018, researchers showed how to do this with police bodycam footage. [104] Even if there were methods to defeat location spoofing and on-device manipulation, there is a simple workaround. Adversaries could use their media capture devices in situ to record the output of a second device playing a deepfake.

Five years ago, Bobby Chesney and Danielle Citron warned of "a worst-case scenario . . . in which it is cheap and easy to [make deepfakes] with inadequate technology to quickly and reliably expose [them]."[105] We anticipate an even worse worst-case scenario in which cheap and easy deepfakes are immune not only to quick and reliable detection, but to any detection at all. Deepfake generators will eventually evolve to create what this Article calls "deepest fakes," costless deepfakes that no procedure can distinguish from authentic media. Deepest fakes should shake the confidence of scholars who think that expert testimony and detection techniques will rescues from the coming upheaval. Deepest fakes undermine all existing proposals for devising enhanced authentication requirements for digital media. In a world where deepest fakes are prevalent, the only distinguishing feature of inauthentic media would be that it portrays an event that never occurred.[106] Verifying authenticity would create an impossible circularity—demanding (independent) proof of the event that the media itself serves to prove.

[102] Chesney & Citron, *supra* note 22, at 1815.

[103] *See* Niles Ole Tippenhauer, *On the Requirements for Successful GPS Spoofing Attacks*, Proceedings of the 18th ACM Conference on Comp. & Comm. Sec. (Oct. 2011), https://dl.acm.org/doi/pdf/10.1145/2046707.2046719. Today, even children use spoofing to access new locations in augmented reality videogames. Spoofer Go, https://www.spoofer-go.com/ ("Supported by Pokémon Go, Spoofer Go has a powerful fake location function along with great movement functions that make exploring the Pokémon Go world easier and more exciting.").

[104] Lily Hay Newman, *Police Bodycams Can Be Hacked to Doctor Footage*, WIRED (Aug. 11, 2018), https://www.wired.com/story/police-body-camera-vulnerabilities/.

[105] Chesney & Citron, *supra* note 22, at 1814.

[106] Blitz, *supra* note 19, at 68 ("[F]alse factual statements are unlike religious ideas and political opinions in at least one respect: they can be exposed as fake.").

III. Epistemology and Deepest Fakes

The threat of deepest fakes may be recent, but the worry is hardly novel. In his First Meditation on Philosophy (1641), René Descartes imagines that everything he thought he knew is merely an illusion:

> I will suppose therefore that . . . some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colors, shapes, sounds and all external things are merely the delusions of dreams which he has devised to ensnare my judgment. I shall consider myself as not having hands or eyes, or flesh, or blood or senses, but as falsely believing that I have all these things. . . . [E]ven if it is not in my power to know any truth, I shall at least . . . resolutely guard against assenting to any falsehoods.[107]

Descartes then asks whether any kind of knowledge is possible in the face of such extreme doubt.

Deepest fake generators are a far cry from Descartes' demon, but they do similarly destabilize an information environment we formerly trusted. Descartes' thought experiment forces us to confront the possibility of skepticism about perception. Can we trust our eyes and ears to deliver true information about reality, or, as Edgar Allan Poe write, could it be that "all that we see or seem is but a dream within a dream"?[108] Even if our eyes and ears *could* deliver true information about reality, the very possibility that Descartes' demon *might* be manipulating what we see and hear undermines the extent to which we can trust perception. The eventuality of deepest fakes forces a parallel skeptical worry: In the age of deepest fakes, could we ever rely on digital media, or must we treat it all as the digital equivalent of hallucination? This is the question that motivates deepfake alarmism.

This Part provides philosophical tools that help to structure a response. It begins by introducing some basic concepts from epistemology, including skepticism (Part III.A). While anti-skeptical philosophers have many responses to Descartes' thought experiment, these responses are surprisingly ineffective at addressing deepfake alarmism. As we show, deepfake alarmism poses a philosophical challenge that is in some respects less tractable than even Carteisan skepticism. While there is a nascent philosophical literature trying to characterize and ameliorate the epistemic threat that deepfakes pose, their solutions misfire when confronted with deepest fakes or the particular challenges of the courtroom context (Part III.B). All is not lost,

[107] RENÉ DESCARTES, MEDITATIONS ON FIRST PHILOSOPHY (Donald A. Cress trans., Hackett Publishing Co. 3d ed. 1993) (1641).

[108] Edgar Allan Poe, *Dream Within a Dream*, THE FLAG OF OUR UNION (Mar. 1849).

though. There are elements of epistemologists' views (Part III.C) that, modified and extended as we propose below, motivate strategies for courts (Part V) and ordinary people (Part VI) to adapt to an increasingly unreliable digital landscape.

A. Digital Media Skepticism

Epistemology is the philosophical study of knowledge and belief. It offers a conceptual framework for understanding the ways that deepfakes undermine our ability to know about the world around us. More importantly, epistemologists elucidate alternate pathways to knowledge.

According to the classic definition,[109] "knowledge" is justified true belief.[110] (Most philosophers today add various other complicating requirements that need not detain us here.[111]) "Belief" is a mental state that represents reality as being a particular way: for example, "it is raining outside."[112] A belief is "true" if it represents reality accurately: for example, "it *really is* raining outside." The justification element in the definition of knowledge is much more contested.

Two dominant views on epistemic justification are evidentialism and reliabilism. Evidentialists maintain that a person is justified in believing a proposition if she possesses sufficient evidence that it is true.[113] Types of evidence include perception (you see the rain outside your window), introspection (you experience the pain in your knee that often precedes rain), memory (you remember seeing rain clouds rolling in this morning), intuition (you had a premonition of rain), and testimony (your local meteorologist tells you it's raining). Depending on the circumstance, some types of evidence will be stronger than others. Seeing that it's raining is usually enough to justify the belief that it's raining, but merely hearing a pitter-patter on the roof may need supplementing by other evidence to justify your belief. Evidentialists usually do not specify a bright line threshold for what counts as sufficient evidence, and the threshold may vary. For example, more evidence may be

[109] The discussion throughout this paper focuses only on *a posteriori* knowledge, i.e. knowledge about the outside world obtained through experience. Bruce Russel, A Priori *Justification and Knowledge*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta & Uri Nodelman eds., 2024), https://plato.stanford.edu/archives/sum2024/entries/apriori/. *A priori* knowledge, i.e. knowledge obtained through reason alone, is not relevant here. *Id.*

[110] PAUL K. MOSER & ARNOLD VANDER NAT, HUMAN KNOWLEDGE: CLASSICAL AND CONTEMPORARY APPROACHES 12-15 (1987). Note, this is different from how the legal system often characterizes knowledge, which merely as true belief. Mihailis E. Diamantis, *Functional Corporate Knowledge*, 61 WM. & MARY L. REV. 319, 334-35 (2019).

[111] *See* Edmund L. Gettier, *Is Justified True Belief Knowledge?*, 23 ANALYSIS 121 (1963).

[112] DAVID HUME, TREATISE OF HUMAN NATURE, (L.A. Selby-Bigge & P.H. Nidditch eds., Oxford University Press, 1978) (1740).

[113] Jaegwon Kim, *What is Naturalized Epistemology*, *in* 2 PHILOSOPHICAL PERSPECTIVES: EPISTEMOLOGY 381(James Tomberlin ed., 1988).

needed for forming justified beliefs about matters of consequence (is it safe to go sailing today?) than for relative trivialities (will I need to mow my grass again next week?).[114]

Reliabilism is the view that beliefs are justified if they are formed using a process that usually results in true beliefs.[115] For example, someone could form the belief that it will rain tomorrow either by checking the weather forecast or their daily horoscope. The process that involves the forecast is reliable and would result in a justified belief. Not so for the horoscope. Some processes (e.g., normal human vision in good lighting) are unconditionally reliable, delivering appropriate outputs in most situations, while others (e.g., deductive inference) are only conditionally reliable, depending on correct inputs. Like evidentialists, reliabilists usually do not specify a precise threshold for how reliable a belief-forming process must be—how likely to result in true beliefs—and acknowledge that the threshold may depend on context.[116]

Often, reliabilists and evidentialists give the same answers in ordinary cases.[117] This is to be expected since both aspire to reflect commonsense intuitions about when beliefs are justified (and hence candidates for being knowledge). For example, both views would hold that a person's belief that it's raining is justified if that's what the meteorologist told her—she both possesses sufficient evidence and she formed her belief using a reliable process. The views may give different results in some exotic cases involving wishful thinking,[118] alternate universes,[119] or clairvoyance.[120] We will not attempt to referee which view is superior. For present purposes, it suffices to note that evidentialism and reliabilism offer differ conceptions of epistemic justification, one focused on evidence and the other focused on process.

[114] Jeremy Fangl & Matthew McGrath, *Evidence, Pragmatics, and Justification*, 111 PHIL. REV. 67 (2002); JASON STANLEY, KNOWLEDGE AND PRACTICAL INTERESTS (2005).

[115] F.P. Ramsey, *Knowledge*, *in* FOUNDATIONS OF MATHEMATICS AND OTHER LOGICAL ESSAYS 126 (R.B. Braithwaite ed., 1931); ALVIN I. GOLDMAN, EPISTEMOLOGY AND COGNITION (1986). There are other variants of reliabilism, *see, e.g.* ERNEST SOSA, KNOWLEDGE IN PERSPECTIVE: SELECTED ESSAYS IN EPISTEMOLOGY (1991) (discussing virtue reliabilism); Peter Unger, *An Analysis of Factual Knowledge*, 65 J. PHIL. 157 (1968) (defending a version of reliabilism according to which someone is justified in holding a belief just in case "it is not at all accidental" that the belief is true), but the process-oriented view is the most common.

[116] Alvin Goldman & Bob Beddor, *Reliabilist Epistemology*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2021), https://plato.stanford.edu/entries/reliabilism/ ("Just how high a truth-ratio a process must have to confer justification is left vague.").

[117] Indeed, some epistemologists have argued for a unified approach. *See, e.g.*, Juan Comesaña, *Evidentialist Reliabilism*, 44 Noûs 571 (2010); Alvin I. Goldman, *Toward a Synthesis of Reliabilism and Evidentialism*, *in* EVIDENTIALISM AND ITS DISCONTENTS (Trent Dougherty ed., 2011).

[118] Alvin I. Goldman, *What is Justified Belief?*, *in* JUSTIFICATION AND KNOWLEDGE 1 (G.S. Pappas ed., 1979).

[119] Hilary Putnam, *The Meaning of 'Meaning'*, *in* II PHILOSOPHICAL PAPERS: MIND, LANGUAGE AND REALITY (1975); Tyler Burge, *Individualism and the Mental*, 4 MIDWEST STUDS. PHIL. 73 (1979).

[120] Lawrence Bonjour, *Externalist Theories of Empirical Knowledge*, 5 MIDWEST STUDS. PHIL. 53 (1980); Ralph Wedgwood, *the Aim of Belief*, 16 PHIL. PERSPECTIVES 267 (2002).

Skepticism is the worry that we do not know some class of beliefs that we ordinarily take ourselves to be justified in holding.[121] Different varieties of skepticism target different classes of beliefs.[122] Other minds skepticism, for example, arises from the solipsistic worry that everyone else might by unconscious.[123] It aims to undermine the justification for our beliefs about other people's minds. Pyrrhonian skepticism, by contrast, calls all knowledge into doubt by questioning whether we are ever justified in believing anything.[124] Cartesian skepticism falls somewhere between the two. Recall that Descartes entertains the thought "that the sky, the air, the earth, colors, shapes, sounds and all external things are merely the delusions of dreams."[125] What he articulates is external world skepticism: the view that we cannot know that there is an external world or know anything about it. While Descartes' thought experiment involves us being manipulated by an evil demon, more modern variants raise the possibility that we might just be brains in vats[126] or figments of some vast computer simulation.[127]

Both evidentialism and reliabilism must confront Descrates' radical doubt. If such widespread demonic deception is possible, how can anyone be confident in their evidence or in the reliability of their belief-forming processes? We ordinarily take perception to provide sufficient evidence for beliefs about the external world.[128] That presumption assumes that there is no genuine grounds for doubting our perception.[129] If we suspect we've ingested a hallucinogenic drug or are being manipulated by an evil demon, then grounds for doubt start to creep in, and perception may lose its justificatory power.

Descartes' strategy for overcoming external world skepticism relied on an elaborate argument against the very possibility of the evil demon. He started by identifying at least one thing he could know with certainty, even in the grips of a demon's illusions:

[121] Peter Klein, *Skepticism*, *in* SANDFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2015), https://plato.stanford.edu/archives/sum2015/entries/skepticism/.

[122] VARIETIES OF SKEPTICISM: ESSAYS AFTER KANT, WITTGENSTEIN, AND CAVELL (James Conant & Andrea Kern eds., 2014).

[123] *See generally* ANITA AVRAMIDES, OTHER MINDS (2001).

[124] Juan Comesaña, *The Pyrrhonian Problematic*, *in* ENCYCLOPEDIA OF PHILOSOPHY (Donald M. Borchert ed., 2d ed. 2005).

[125] Descartes, *supra* note 107; Lex Newman, *Descartes on the Method of Analysis*, *in* THE OXFORD HANDBOOK OF DESCARTES AND CARTESIANSIM 65 (S. Nadler et al. eds., 2019).

[126] HILARY PUTNAM, REASON, TRUTH, AND HISTORY (1981); David J. Chalmers, *The Matrix as Metaphysics*, *in* PHILOSOPHERS EXPLORE THE MATRIX (Christopher Grau ed., 2005).

[127] Nick Bostrom, *Are We Living in a Computer Simulation*, 53 PHIL. Q. 243 (2003).

[128] Ali Hasan, *The Evidence in Perception*, *in* THE ROUTLEDGE HANDBOOK OF THE PHILOSOPHY OF EVIDENCE (Maria Lasonen-Aarnio & Clayton Littlejohns eds., 2024).

[129] Klein, *supra* note 121.

> [L]et [the demon] deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. So after considering everything very thoroughly, I must finally conclude that this proposition, I am, I exist, is necessarily true whenever it is put forward by me or conceived in my mind.[130]

This is the famous "cogito" argument: "I think, therefore I am." While the cogito has some undeniable plausibility, fewer people find the rest of Descartes' argument against skepticism very persuasive. He leverages the cogito to conclude that there is an all-powerful God, that God must be benevolent, and that such a God would not permit an evil demon to deceive us.[131]

Modern evidentialists take a different approach to external world skepticism, offering a multi-pronged response that avoids Descartes' reliance on metaphysical guarantees. According to one prominent evidentialist position, "dogmatism," our perceptual beliefs about the external world have prima facie justificatory power for belief about the circumstances they convey.[132] This justification stands so long as there is no specific evidence that undermines or contradicts the perceptual evidence, i.e. so long as there are no "defeaters."[133] Importantly, there is no evidence for skeptical hypotheses like the evil demon. The mere conceptual possibility of an evil demon does not defeat the prima facie power of perceptual justification.[134] (Of course, the evidentiary landscape would be very different if we perceived the evil demon itself.)

Evidentialists have a second type of response to Cartesian skepticism: inference to the best explanation. In short, the hypothesis that there is an external world offers a simpler and more coherent explanation of several facts about our perceptual experience than skeptical alternatives: perceptual experiences tend to be stable over time (stop signs look red on every day of the week), people tend to perceive things similarly (stop signs look red to (nearly) all of us), and perceptions generally facilitate practical success (people who stop when they see stop signs live longer).[135] If there were an evil demon, we'd need a rather convoluted story to explain why it cares to assure our perceptual experiences have these traits.

Unlike evidentialists, reliabilists respond to external world skepticism without appealing to any features of our evidence.[136] A common reliabilist response invokes

[130] Descartes, *supra* note 107.
[131] *Id.*
[132] James Pryor, *The Skeptic and the Dogmatist*, 110 NOÛS 517 (2000).
[133] Ali Hasan, *The Evidence in Perception*, *in* THE ROUTLEDGE HANDBOOK OF THE PHILOSOPHY OF EVIDENCE 225 (Maria Lasonen-Aarnio & Clayton Littlejohn eds., 2024).
[134] Fred Dretske*, Epistemic Operators,* 67 J. OF PHIL 1007 (1970); Gail Stine*, Skepticism, Relevant Alternatives, and Deductive Closure*, 29 PHIL. STUDS. 249 (1976).
[135] Jonathan Vogel*, Cartesian Skepticism and Inference to the Best Explanation,* 10 J. PHIL. 602 (1990).
[136] Richard Feldman, *Reliability and Justification*, 68 MONIST 159 (1985); Richard Foley, *What's Wrong with Reliabilism*, 2 MONIST 188 (1985).

the concept of epistemic safety.[137] A belief is considered safe just in case the person could not have easily come to hold it without it being true. "Easily" is doing a lot of work in this definition. There is a technical definition of "easily" using possible world semantics,[138] but the general idea turns on how radically different the world would have to be for someone to form a belief based on the same evidence without the belief being true. The safety condition enriches the reliabilist conception of reliability by adding the requirement of robustness across possible worlds: not only must the processes producing a belief be reliable in the actual world, but the belief must remain true across nearby changes in circumstance.

An example will help. Suppose Tina believes her favorite band is playing because that's what she thinks she sees and hears. She also believes that she's at a rave and that hallucinogens are plentiful at raves. Her belief that her favorite band is playing probably is not very safe. Her perceptions *might* be distorted by a drug. It isn't such a remote possibility that she accidentally ingested a hallucinogen or that she purposely took it and forgot. Under the effects of the hallucinogen, Tina might mistake a second-rate cover band for her favorite band. In other words, she might easily have the same belief, without the belief being true. Applying the safety condition, reliabilists could conclude that Tina's belief does not amount to knowledge, even if her favorite band actually is playing.

Contrast Tina's belief about her favorite band with our belief in an external world. We move about in what we take to be a real external world filled with coffee shops, honking cars, and street musicians. We believe all these things are real because that's what we think we see and hear. Of course, as Descartes observed, it is *possible* that we are in the grips of an extended illusion orchestrated by an evil demon. But that possible scenario is very remote from the world we believe we inhabit—many things would have to be very different. It follows that our belief in an external world is safe, and Descartes' hypothetical does not undermine our knowledge.

What has all of this got to do with deepfakes? One way to understand the position of deepfake alarmists is that they are positing a new variety of skepticism. We might call it "digital media skepticism." While we ordinarily take ourselves to be justified in believing something after we've seen a video recording of it, digital media skepticism claims that the possibility that the recording could be a deepfake undermines our justification. "[We] may doubt [even] unaltered content simply because [we] know realistic deepfakes are possible."[139] The skeptical worry only grows as deepfakes become more prevalent and more realistic, i.e. as deepest fakes come into the picture.

[137] Ernest Sosa, *Tracking, Competence, and Knowledge*, *in* THE OXFORD HANDBOOK TO EPISTEMOLOGY 264 (Paul Moser ed., 2002).

[138] Duncan Pritchard, *Anti-Luck Epistemology*, 158 SYNTHESE 277 (2007).

[139] Delfino, *supra* note 14, at 337.

Surprisingly, digital media skepticism appears to be even more philosophically intractable than external world skepticism. Under the conditions that concern this paper—adversarial trial in a nearby future where deepest fakes exist—the standard evidentialist and reliabilist solutions to external world skepticism do not work.

For evidentialists, deepest fakes make it much harder for digital media to provide justifying evidence. Dogmatism about the justificatory power of digital media fails because evidence that deepfakes exist defeats the prima facie credibility of all digital media. The risk of a deepest fake is highly pertinent, not remote. Indeed, there are digital media that tell us about the threat of deepfakes. So, paradoxically, trusting digital media requires us to distrust digital media. Inference to the best explanation also struggles to justify trust in digital media since easy explanations for convincing forgeries are readily available.

For reliabilists, digital media skepticism is more troubling than external world skepticism precisely because it isn't some far-fetched scenario. While ordinary beliefs about the external world based on direct perception may be safe, beliefs based on digital media would be decidedly unsafe.[140] Digital media could very easily convey false content because a) that's what deepest fakes do and b) the adversarial context of the courtroom makes it much more likely that a video would be a deepest fake. Forming beliefs on the basis of digital media becomes a much less reliable process, particularly when highly motivated courtroom adversaries provide them. As deepest fakes become increasingly prevalent, there will be no sensibly "normal" reference context for reassessing the reliability of the process of forming beliefs using digital media.

Digital media skepticism raises an additional unsettling worry for both evidentialists and reliabilists that standard presentations of external world skepticism do not. On typical formulations of the evil demon hypothetical, the demon induces perceptions that reflect the sorts of experiences we ordinarily have. This means it looks to us like there is an external world, and part of the explanation is that our perceptions are stable, composed, and consistent. Digital media skepticism, by contrast, envisions a world in which digital media simultaneously present directly conflicting representations of the same reality. If the conflicting videos happen to be deepest fakes, there will be no internal indication that either video is more reliable or provides better evidence. The raises to salience the fact that at least one of the videos misrepresents reality. And possibly both do!

B. Backstops, Signals, and Norms

[140] Taylor Matthews & Ian James Kidd, *The Ethics and Epistemology of Deepfakes*, *in* THE ROUTLEDGE HANDBOOK OF PHILOSOPHY AND MEDIA ETHICS (Carl Fox & Joe Saunders eds.)

There is a nascent philosophical literature about deepfakes. Ethicists focus on the moral harm that deepfakes can cause. Adrienne de Reuiter, for example, uses Kantian ethics to characterize nonconsensual deepfakes as a form of "digital persona plagiarism." [141] Epistemologists aim to characterize the "epistemic harm" of deepfakes.[142] Pessimists envision a future in which deepfakes induce widespread digital media skepticism, and digital media skepticism undermines other knowledge practices. Optimists think they've identified systems that could mitigate deepfakes' epistemic harms. Both groups tend to consider the impact that existing deepfake technology will have on people as they go about their ordinary lives. Deepest fakes in the courtroom raise a novel set of challenges that amplify pessimists' concerns (Part III.B.1) and neutralize optimists' solutions (Part III.B.2).

1. The Epistemic Harm of Deepfakes

Don Fallis and Regina Rini offer the two most influential philosophical accounts of the ways that deepfakes can distort or undermine our ability to form true and justified beliefs. Fallis focuses on beliefs we form using digital media. He argues that such beliefs have become integral to how we learn about the world. While direct perception may be the evidentiary gold standard, we "cannot always be at the right place at the right time, to see things for ourselves. In such cases, videos are often the next best thing."[143] It is one thing to read about the devastation in Ukraine, but quite another to see video footage of it.

Fallis persuasively describes how deepfakes make it harder to form justified beliefs by "reduc[ing] the *amount of information* that videos carry to viewers."[144] "Information" is a technical term in epistemology that refers how much some piece evidence tells a given viewer about the world,[145] and how reliably.[146] A video will typically carry less information if it is low resolution or shot from a bad angle. It also carries less information if it is less likely to portray something true. More formally speaking:

> R [some piece of evidence] carries the information that S [some signal about a state of the world] when the likelihood of R being sent when S is true is greater than the likelihood of R being sent when S is

[141] Adrienne de Reuiter, *The Distinct Wrong of Deepfakes*, PHIL. & TECH. 14, 16 (June 10, 2021) (""Non-consensual deepfakes wrong the persons they portray because they manipulate the process through which people's identity is socially constituted.").

[142] Don Fallis, *The Epistemic Threat of Deepfakes*, 34 PHIL. & TECH. 623, 624 (2021).

[143] *Id.* at 624.

[144] *Id.*

[145] FRED DRETSKE, KNOWLEDGE AND THE FLOW OF INFORMATION (1981); Jonathan Cohen & Aaron Meskin, *On the Epistemic Value of Photographs*, 62 J. AESTHETICS & ART CRIT. 197 (2004).

[146] BRIAN SKYRMS, SIGNALS: EVOLUTION, LEARNING, AND INFORMATION (2010).

> false. . . . [T]he more likely it is for a signal R to be sent in the state where S is true than it is for R to be sent in the state where S is false, the more information that R carries about S.[147]

Deepfakes make it more likely that a video conveying some content about the world would exist even if that content were false. "Deepfake technology increases the probability of a false positive. . . . As a result, videos carry less information than they once did."[148] The more prevalent deepfakes become, the less information all videos (even truthful ones) will carry. This generates epistemic harm: "We cannot learn as much about the world if less information is carried by videos"[149]

Videos will carry *even less* information in courtrooms once deepest fakes are possible. The eventuality of deepest fakes will dampen information carry in two predictable ways on Fallis framework. First, because deepest fakes will be indistinguishable from authentic media, it will become easier for them to send an undetectably false signal.[150] Second, because deepest fakes will be very easy to make, there will simply be more of them. As the ratio of deepest fakes to authentic videos (i.e. the ratio of noise to signal) increases, videos will become more likely to contain false content.[151] The courtroom context makes matters worse. Highly motivated, adversarial parties have stronger incentives to create and/or introduce fake content. We should expect that deepest fakes would be even more highly concentrated in courtrooms than in the wild.

Regina Rini argues that the epistemic harms of deepfakes reach far beyond digital media. Rather than start with the important epistemic role of videos, Rini begins with testimony, i.e. evidence we receive from others' telling it to us. "Our collective epistemic practices are highly reliant on testimony."[152] Indeed, most of our higher order learning comes from testimony. For those of us who do not perform basic science, unearth ancient artifacts, or visit foreign countries, testimony is often the only evidence we have. I know water is made of hydrogen and oxygen, that I have two kidneys, and that iPhones are made in China only because others have told me.

Of course, we can't believe everything we're told. We're only justified in believing something someone tells us if we are antecedently justified in believing that person is trustworthy.[153] We might believe someone is trustworthy because we have an extended relationship with them, in the course of which they have displayed their

[147] Fallis, *supra* note 142, at 629.
[148] *Id.* at 632.
[149] *Id.* at 633.
[150] *Id.* at 632 ("The probability of a false positive depends on the viewer's ability to distinguish between genuine videos and fake videos.").
[151] *Id.* at 626 ("[D]eepfake technology threatens to drastically increase the number of realistic fake videos in circulation.").
[152] Regina Rini, *Deepfakes and the Epistemic Backstop*, 20 PHIL. IMPRINT 1 (2020)
[153] Don Fallis, *Lying and Omissions*, *in* OXFORD HANDBOOK OF LYING 183 (Jorg Meibauer ed., 2018).

commitment and capacity to say true things to us. For everyone else—from teachers, to book authors, to random people we ask for directions—we usually start by trusting what they say of the social norms that govern testimony. "When a person attempts to provide testimony, she is taken to be implying that she is both sincere and competent."[154] One reason we can rely on others to follow these norms is that there are reputational consequences for misstating the truth.[155] A scholar whose articles contain falsehoods won't have many readers after the word gets out.

On Rini's account, videos play a critical role enforcing testimonial norms.[156] Her idea is that in public spaces, there is an ever-present possibility of being recorded, whether by a security CCTV, a doorbell camera, or in the background of someone's TikTok montage.[157] Testimony is more trustworthy because people know there's a decent chance that what they say is being recorded and that any falsehoods could have reputational consequences. For this reason, Rini says, "[v]ideo and audio recordings function as an epistemic backstop."[158]

Rini agrees with Fallis that deepfakes make videos overall less reliable because they carry less information.[159] That itself is an epistemic harm. But the more worrisome effect is that "[v]ideo and audio recordings may lose their status as acute correctors of the testimonial record."[160] Rini envisions a world in which we not only trust videos less, but also each other. Deepfakes undermine the epistemic backstop of video, making the entire network of testimonial knowledge vulnerable to collapse. "[T]he gravest danger of deepfakes [is that] [w]ithin a few years, we may have little reason to trust the testimony of strangers."[161]

It is easy to see how deepest fakes and the courtroom context would amplify Rini's concerns. As argued above, when deepest fakes are possible, videos overall will carry less information, even more so in the courtroom than in other contexts. If Rini is right that videos serve as a critical epistemic backstop for testimony, this could be devastating in the courtroom. Trials are almost always about past events that happened out of the courtroom. This makes factfinders especially dependent on

[154] Rini, *supra* note 152, at 2.

[155] *Id.*

[156] *Id.* ("The availability of recordings undergirds the norms of testimonial practice, increasing the incentive for testifiers to speak with sincerity and competence.").

[157] *Id.* at 3 ("When we are in public urban spaces, we know that we're much more likely than not covered by CCTV cameras or traipsing through the background of any number of strangers' selfie-directed phones.").

[158] *Id.* at 2.

[159] *Id.* at 7 ("The obvious worry about deepfakes is that they will be used to propagate vivid disinformation. . . . But I think that the most important risk is that . . . increasingly savvy information consumers will come to reflexively distrust *all* recordings.").

[160] *Id.* at 8.

[161] *Id.*

witness testimony. If witnesses become unreliable because videos can no longer credibly impeach them, it is hard to see how courts could continue to function.

2. Deepfake Optimism

There are philosophers who are more optimistic about the resilience of our epistemic practices. Some place their confidence in technological interventions like those discussed above, e.g. investing in deepfake detectors or blockchain video authentication.[162] We have already shown why those approaches are unlikely to help. Two epistemologists—Joshua Habgood-Coote and Keith Raymond Harris—offer more sophisticated reasons for optimism. Unfortunately, neither is up to the challenges of deepest fakes and courtrooms.

Unlike Rini, Habgood-Coote doesn't think video has a unique epistemic role to play vis-à-vis other sorts of media.[163] As he argues through a detailed history of manipulation in photographs, persuasive media fakery is nothing new.[164] Yet, somehow, our epistemic practices adapted so that we do sometimes rely on photographs for forming justified beliefs.[165] The reason, Habgood-Coote says, is that norms developed to govern photography,[166] much like the epistemic norms that Rini says govern testimony. When we trust what we see in a photograph, we are not only trusting an individual photographer, but a diffuse set of epistemic practices that binds photographers.[167] "The reason why we continue to trust photography is that, in epistemic photographic practices, photo-manipulation is unprofessional, and is punished."[168]

[162] *See, e.g.*, Luciano Floridi, *Artificial Intelligence, Deepfakes and a Future of Ectypes*; Fallis, *supra* note 142, at 640 ("Another possible strategy for increasing the amount of information videos carry is for us to get better at identifying deepfakes."). Keith Raymond Harris offers additional persuasive reasons against relying on deepfake detectors, even when they are provably accurate. Keith Raymond Harris, *AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors*, 27 PHIL. & TECH. 3 (2024).

[163] Habgood-Coote, *Deepfakes and the Epistemic Apocalypse* 1.

[164] *Id.* 18 ("Forgetting the history of photographic manipulation both encourages us to think of deepfakes as a novel problem, and amplifies our perception of the seriousness of the problem."); *see also* Britt Paris & Joan Donovan, *Deepfakes and Cheapfakes: The Manipulation of Audio and Visual Evidence*, DATA & SOC. (Sept. 18, 2019), https://datasociety.net/library/deepfakes-and-cheap-fakes//

[165] Indeed, some philosophers have argued that seeing something in a photo is akin to directly perceiving it. Dan Cavedon-Taylor, *Photographically Based Knowledge*, 10 EPISTEME 283 (2013).

[166] Habgood-Coote, *supra* note 163, at 13 ("[T]he development of the professional identity of the documentary photographer did establish a practice of photography in which photographers were both trusted and trustworthy, within which manipulated photos counted as norm violations.")

[167] Sandy Goldberg, *The Division of Epistemic Labor*, 8 EPISTEME 112 (2011); Habgood-Coote, *Deepfakes and the Epistemic Apocalypse* 7 ("We rely on a set of information-dissemination practices.").

[168] DOMINIC MCIVER LOPES, FOUR ARTS OF PHOTOGRAPHY 110 (2016).

Habgood-Coote predicts that similar norms for video creation will develop, if they are not already in effect: "I take it as given that producing inaccurate deepfakes and disseminating them as real videos is a violation of the norms of producing and disseminating videos."[169] Of course, deepfakes do exist, but the reason they do is that there are "long-running problems of management of the norms of producing and disseminating recordings." [170] To the extent that these norms need a little encouragement from the outside, Habgood-Coote is confident that we have a good "sense of how to design better social practices."[171] He suggests removing the financial incentives for making deepfakes (particularly pornography), banning online forums where deepfakes are shared, and taking down widely used tools for making deepfakes.[172]

Setting aside the question of whether the interventions Habgood-Coote proposes are consistent with the First Amendment,[173] it is doubtful that social regulation would work for deepest fakes or courtrooms. Any ban or restriction on deepfakes will be exceedingly hard to enforce against deepest fakes, which, by definition, are indistinguishable from authentic media. Habgood-Coote presupposes there would be no "catastrophic norm flouting," but that is exactly what deepest fakes enable.[174] Even if he is right that epistemic norms around video creation will develop, like those that govern professional photographers, there is no reason to think those norms would extend to the courtroom. Most plaintiffs and defendants are not professional videographers, so they would have no reason to know or follow the relevant norms. Even if they did, the adversarial context can provide very strong incentives for flouting norms when there is little chance of detection.

Keith Raymond Harris has different reasons for concluding that "[c]oncerns that deepfakes will bring about epistemic catastrophe are overblown."[175] Harris' important insight is that "the evidential power of video derives not solely from its content, but also from its source."[176] A video handed to us from a trusted source holds a different epistemic value than a video that comes from an unknown or untrusted source.[177] Ordinary people can avoid broader digital media skepticism by taking a

[169] Habgood-Coote, *supra* note 163, at 8

[170] *Id.* at 19 ("Whether [deepfake] videos are created is a matter of the social context in which they are deployed.").

[171] *Id.* at 9

[172] *Id.*

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

[174] Habgood-Coote, *supra* note 163, at 8.

[175] Keith Raymond Harris, *Video on Demand: What Deepfakes Do and How They Harm*, 199 SYNTHESE 13373, 13374 (2021).

[176] *Id.* at 13374.

[177] Fallis makes a related point, though he doesn't expanding much on it. Fallis, *supra* note 142, at 640 ("Even without laws against deepfakes, the evening news is subject to normative constraints. Thus, we can try to identify those videos that still carry a lot of information.").

"skeptical attitude [only] toward video footage that does not come from trusted sources [while] continu[ing] to rely on video footage from trusted sources."[178]

While Harris' approach to deepfakes strikes us as a step in the right direction, it does not have internal resources for handling deepest fakes. It is clear from Harris' framework that the "source" of a video is the person or entity who provides or "present[s]" the video, not the person who records the video.[179] For example, if a news station plays footage provided from an informant, the source for the viewers is the news station (not the informant). As Harris anticipates, deepest fake technology could "generate fabricated video footage [falsely] depicting [that it comes from] a trusted [source]."[180] For example, a TikTok reel could falsely depict a CNN anchor introducing a fake video. Harris offers only a partial solution. Ordinary people may learn to access videos directly from the channels the videos purport to come from, e.g. by tuning into CNN rather than watching TikTok.[181] However, this "is of no utility to sources themselves."[182] While channels "can continue to rely on their own video footage,"[183] e.g. CNN can rely on footage its own reporters take, they must reckon with the threat of digital media skepticism for everything else.

It is also difficult to see how Harris' approach would translate to the courtroom. He says little of how we come to trust a source or channel. Presumably trust is the sort of thing that builds over time through repeated interaction. In the adversarial courtroom setting, trust is usually in short supply. There are no court-sanctioned media sources, let alone media channels. So, while Harris is certainly right that that a video's evidential value depends in part on the video's source, he does not have resources internal to his view for warding off digital media skepticism in the courtroom.

C. Philosophical Takeaways

By now, the full threat that deepfakes pose should be clear. Digital media have become an indispensable part of our epistemic practices. Because each of us can only directly learn about a very narrow slice of reality, we must learn everything else indirectly from other sources. Digital media are one such source because they can provide a durable, accurate record of events that took place at distant places and

[178] Harris, *supra* note 175, at 13380.

[179] *Id.*

[180] *Id.* at 13383; Fallis, *supra* note 142, at 640 ("Purveyors of deepfakes can try to make it difficult for people to determine whether a video comes from a source that is subject to normative constraints.").

[181] Harris, *supra* note 175, at 13384 ("The present concern draws attention to the oft-neglected significance of what we might call channels of information. . . . While purveyors of deepfakes might exploit patters of trust by using certain likenesses and logos, they cannot easily inject deepfakes into particular channels.").

[182] *Id.* at 13382.

[183] *Id.* at 13383.

times. Deepfakes destabilize this pathway to knowledge by pushing us toward *digital media skepticism*. Once we learn that digital media content can be persuasively faked, we may "come to reflexively distrust *all* recordings."[184] If we become digital media skeptics, we won't be able to know nearly as many things as we previously could.

To make matters worse, digital media skepticism might not be an overreaction. Prominent views in epistemology explain why, under the right conditions, digital media skepticism might become a rational response. For purposes of this Article, we have been assuming two such conditions hold. First, we suppose that deepest fakes (which are indistinguishable from genuine media) will eventually be possible, plentiful, and costless to make. Second, focus our discussion on the adversarial courtroom, where inter-personal trust is extremely low. Under these conditions, there are strong evidentialist and reliabilist arguments for digital media skepticism. *Evidentialists* believe that we come to know things by forming justified beliefs based on sufficient evidence. While digital media might once have been good evidence for believing the events they portray, the existence of deepest fakes *defeats* digital media's justificatory power. As deepest fakes become increasingly prevalent, the *information* digital media carry (or the *signal* they send) weakens because the information environment becomes polluted with indistinguishable *noise* (i.e. false signals, i.e. misleading content). Eventually, the truth signal digital media sends will become so low that the rational response will be to distrust just about all digital media. *Reliabilists* argue that we come to know things by forming justified beliefs using processes that tend to generate true beliefs. When deepest fakes are possible and interpersonal trust is minimal, forming beliefs by viewing digital media becomes a highly unreliable process. In such contexts, reliabilism would also recommend digital media skepticism.

This marks the turning point of the Article. Digital media skepticism seems all but inevitable from both philosophical and legal perspectives. Purported solutions melt away in the face of deepest fakes and the adversarial context of the courtroom. Yet, in what follows, we argue that evidence law has long had resources for staving off digital media skepticism, and that evidence law holds lessons for responsible media consumption in ordinary life.

We depart from prior philosophical work on deepfakes in a critical respect: we center testimony. *Testimony* is the type of evidence we gain from other people when they tell us (verbally or in writing) that some state of the world obtains. As mentioned above, both evidentialists and reliabilists think testimony is important. "So much of what we know about the world, e.g., history, science, politics, one another, etc., comes from the testimony of others."[185] When your science teacher told you that

[184] Rini, *supra* note 152, at 7.

[185] Nick Leonard, *Epistemological Problems of Testimony*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta & Uri Nodelman, eds., 2003), https://plato.stanford.edu/archives/spr2023/entries/testimony-episprob/.

the earth is round, you probably formed a justified belief about the shape of the planet. You didn't have to see the horizon's curvature yourself or view photos taken from space. You acquired good evidence and employed reliable belief-forming processes.

Philosophers writing about deepfakes generally ignore testimony, or they bring it up only to diminish it. Recall that Rini believes our everyday epistemic practice of relying on others' testimony only works because video recordings can help us detect people who lie.[186] She envisions a world of cascading skepticism, where deepfakes lead to digital media skepticism, and digital media skepticism leads to testimony skepticism: "[R]ecordings will be demoted . . . to sources of mere testimonial evidence. And if they are simply just another source of testimony, they cannot be relied upon to correct or regulate testimonial practice."[187] We find Rini's conclusions overblown because, as most epistemologists agree, "mere" testimony is generally a good (and indispensable) source of evidence. Testimonial evidence existed before digital media, and it will survive the advent of deepest fakes.

Indeed, we will argue that tethering digital media closer to testimony is the best way to avoid digital media skepticism. The reliability of human testimony can save deepfakes, rather than vice versa. This flips Rini's justificatory picture on its head. The misstep we see in the philosophy of deepfakes is that epistemologists often compare deepfake technology to Photoshop.[188] They puzzle over how humanity managed to avoid photo skepticism in the face of highly persuasive photo-manipulation techniques. We frame our discussion in terms of a much more ancient form of deceit, humanity's original fakery: the simple fib. Ever since humans could represent states of the world through language, they have also *mis*represented states of the world through language. There are diverse types of misrepresentation, and a speaker's intention plays an important role distinguishing between them:[189] from lies that are designed to deceive, to creative expressions that are designed to entertain or educate, to accidental inaccuracies that arise because of speaker incompetence. Humans have developed epistemic tools for sorting good from bad testimony. In

[186] Rini, *supra* note 152, at 8 ("Within a few years, we may have little reason to trust the testimony of strangers, as the norms securing their anticipated cooperation come gradually undone.").

[187] Rini, *supra* note 152, at 10; Habgood-Coote, *supra* note 163, at 6 (""Once we become aware of the possibility of deepfakes, when we form beliefs based on videos, we must either extend our trust to the videographer, making the videographic knowledge akin to knowledge form testimony, or rely on background beliefs about the likelihood of faking, making it a kind of inferential knowledge. Either way, videographic knowledge loses its distinctive character as non-interpersonal knowledge." (interpreting Rini)).

[188] Rini, *supra* note 152; Habgood-Coote, *supra* note 163.

[189] *See also* Adrienne de Reuiter, *The Distinct Wrong of Deepfakes*, PHIL. & TECH. 3 (June 10, 2021) ("The moral evaluation of specific deepfakes depends on . . . the intent with which the deepfake was created.").

what follows, we contend that these are the key for assessing the impact deepfakes will have.

IV. Evidence Law of Fakeries and Other Creative Content

As the guiding framework for nearly every inquiry into accuracy, authenticity, relevance, and reliability in the structured court environment, the rules and history of evidence hold clues for how people can adapt to a real-world media environment saturated with deepfakes. Optimistic evidence scholars seem to think that if they can just find the right rules, a judge mechanically applying them will filter out deepfakes and expose jurors only to media that provide a direct window into truth. Pessimists agree with the goal but despair of ever finding such rules. They worry that the looming possibility of deepfakes will drive jurors toward digital media skepticism and that courts' fact-finding mission will lose all credibility or meaning.

To be blunt, we think the optimists are naïve and the pessimists are short-sighted. Both sides overlook the essential and powerful intermediating role of human judgment. Optimists' quest for rule-based guarantees will fail because fact-finding cannot be a mechanical process and there are no direct windows to truth. Pessimists' predictions will fail because human judgment stands as a bulwark between distrust and skepticism.

The developments in evidentiary practice that we envision are contemporary retellings of a familiar story. Since the beginning of modern evidence practice—marked by a shift from relying on divine omniscience to relying on human judgment—the law has grappled with the challenge of handling untrue and deceptive evidence. Fibs and forgeries of all sorts are a potent concern when litigants have life, limb, and purse are on the line. But excluding every category of newly manipulable evidence would leave courts with precious little to consider. Instead, the law has continuously striven to include all but the riskiest evidence in an adversarial process that culminates in jurors' common-sense assessments of credibility. The emergence and refinement of evidence law's answers to the threat of deepfakes can be seen in the history of this law's response to two key evidentiary challenges: the problem of lying witnesses (Part IV.A) and the problem of mechanically recorded evidence (Part IV.B).

A. The History of Lies and Deception

Deception through deepfakes is the novel expression of an exceedingly old legal challenge. For as long as disputes have been resolved through litigation, courts have needed a method for distinguishing truthful testimony from lies. The precursor to most trials is, after all, conflicting claims about who did what. The plaintiff says that the defendant stole a horse; the defendant denies it. If both sides swear on their oath that their respective statement of the facts is true, then the court is handed the unenviable task of deciding which of two earnestly pledged factual statements is true—or, to put it another way, to decide which side is lying.[190]

Everyone who is capable of providing testimony is capable of lying, and the trial context can supply strong motivation to do so. Who would not claim innocence if it bought them a slim chance of escaping capital punishment? In a civil context, what private party does not feel at least the urge to edit and exaggerate their story to the benefit of their litigation posture? Even a third-party witness may feel compelled to shade or embellish her testimony in order to protect a party—or herself—from its consequences. These propositions sound prosaic to the modern ear, but they reflect a surprisingly fundamental and enduring challenge for our legal system.

If every witness could be lying, and if many have strong motivations to do so, then what justification can there be for favoring one witness's story over another's? Grave consequences turn upon the answer. Yet few of us are born with any innate ability to separate truth from fiction when handed equally compelling but conflicting accounts about what has happened.[191] The intertwined histories of trials and the law of evidence are substantially a history of trying to reach just results despite the limits of human lie detection.

In the eleventh century precursors to modern trials, for example, the need for mortal judgement between conflicting statements was often relieved by performances tinged with interpretation as divine judgment. Let us suppose that a member of a medieval community had been credibly accused of stealing another's horse. Perhaps motivated by a desire to escape punishment—which might be as severe as death or mutilation—the accused emphatically swears her innocence. To avoid the stalemate that would result from opposing sworn statements by the accused and the accuser, courts would task the accused with establishing her innocence through something like trial by the ordeal of hot iron.[192]

In this ordeal, a judge or priest would order the accused to pick up and carry a searing iron weight for a length of nine paces. After this performance, the accused's hands would be bandaged for three days. Once the waiting period was up, the court would reassemble to unwrap and inspect the hands. Infected, pustulant burns were proof of a guilty soul and thus established that the accuse had falsely claimed her innocence. Healed—or better yet, undamaged—hands proved the purity of the accused's soul, and thus her innocence.

[190] This is particularly apparent in the context of a criminal defendant who vigorously claims their innocence. As others have noted, there is no logical separation between the conclusion that the defendant is guilty as charged and that the defendant is guilty of perjury.

[191] *See infra* notes 216–217 (discussing the dismal empirical record on human skill in lie detection).

[192] Specific details on when and what ordeals would be ordered seem to have varied over time and by jurisdiction. *See generally* ROBERT BARTLETT, TRIAL BY FIRE AND WATER: THE MEDIEVAL JUDICIAL ORDEAL (1986). This example is meant only as an illustration.

This ordeal parallels other fact-finding devices of the time. In the ordeal of cold water, a person's honesty was established when, bound and lowered into water, they did not float but were received by the water and sank.[193] In a trial by judicial combat, the party whose account was truthful would be propelled by their purity to triumph in battle against their opponent.[194] In every case, the conclusion of who was telling the truth was entrusted to divinity and to the spirit.[195] God revealed which side was telling the truth, obviating the need for mortal judgment. This rested the order to dispatch legal sanction upon the highest and most unquestionable authority.

Arresting as this period of trial practice was in legal history, it came to an end in 1215 when the church withdrew its endorsement of trial by ordeal.[196] The historic record is regrettably sparse on what exactly trial practice looked like during the next several hundred years, but a few details are reasonably certain. Without the aid of divine intervention, juries were forced to step forward into something analogous to their current role.[197] The job of finding legal truth thus became a human undertaking. Still, a variety of practices continued to relieve juries from needing to process directly conflicting testimonial narratives.

One such practice was an apparently strong preference for documentary evidence during this period.[198] Documents, especially those sealed and trustworthy by virtue of how they had been created, were preferred and emphasized over live witness testimony, especially in contract disputes.[199] Sir Geoffrey Gilbert's respected treatise

[193] *See* Margaret H. Kerr, Richard D. Forsyth & Michael J. Plyley, *Cold Water and Hot Iron: Trial by Ordeal in England*, 22 J. INTERDISC. HIST. 573, 582–83 (1992).

[194] *See* ROBERT BARTLETT, TRIAL BY FIRE AND WATER: THE MEDIEVAL JUDICIAL ORDEAL 101–126 (1986); *see generally* GEORGE NEILSON, TRIAL BY COMBAT (Williams & Norgate et al. eds., 1890).

[195] PAUL R. HYAMS, TRIAL BY ORDEAL: THE KEY TO PROOF IN THE EARLY COMMON LAW 101–126 (1981) ("Unilateral ordeals, oaths, and duels share one important factor. All three methods of proof purport to work by revealing God's judgment."); GEORGE NEILSON, TRIAL BY COMBAT 111 (Williams & Norgate et al. eds., 1890) ("In such circumstances the accused was bound to purge himself by the judgment of God, viz., by the hot iron if a free man, by water if a villein, according to the divers conditions of men.").

[196] Roger D. Groot, *The Early-Thirteenth-Century Criminal Jury*, in TWELVE GOOD MEN AND TRUE 3, 3 (J. S. Cockburn & Thomas A. Green, eds., 1988) ("The most important event in the history of the criminal jury was the abolition of the ordeal by the Catholic church in 1215."); *see also* ROBERT BARTLETT, TRIAL BY FIRE AND WATER: THE MEDIEVAL JUDICIAL ORDEAL (1986) (commenting that the ordeals were "everywhere vestigial" by 1300).

[197] *See* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575, 585–86 (1997) (explaining that "[the] occasion of [the] sudden birth of trial by jury was the sudden death of trial by ordeal").

[198] *See* John H. Langbein, *Historical Foundations of the Law of Evidence: A View from the Ryder Sources*, 96 COLUM. L. REV. 1168, 1181 (1996) ("The law of evidence in its infancy was concerned almost entirely with rules about the authenticity and the sufficiency of writings.").

[199] *Id.* at 1183 (1996) ("The preference for written evidence extended back to the Middle Ages, and was particularly apparent in contract and conveyancing. The judges determined by the fourteenth century that only contracts written and sealed would be actionable under the writ of covenant. ... The

on evidence, published posthumously in 1754, emphasized the identification, authentication, and epistemic ranking of documentary evidence in detail, while devoting comparatively little attention to the subject of witness testimony.[200] Legal historians report this treatment to be consistent with the surviving record of trial practice at this time.[201]

Even more important were a variety of rules that simply prohibited all testimony from witnesses deemed likely to lie under oath. "Incompetent" witnesses in this regime included criminal defendants,[202] both parties to civil disputes,[203] the spouses of parties,[204] others with personal interest in the outcome,[205] children,[206] convicted criminals,[207] and atheists.[208] Though the details of the rationale varied from one category to the next, the basic reasoning was always the same: because the oath of an incompetent witness could not be trusted to compel truthful testimony, the witness was prophylactically stricken from the stand. This was an act of generosity to the jurors (who would not be challenged with conflicting testimony)[209] as well as to the witness (who would not be placed in a position where she might succumb to temptation and taint her soul with perjury).[210]

legal system that endured into [the 1750s] had exhibited a centuries-long proclivity for suppressing resort to oral evidence at jury trial in civil matters.").

[200] T. P. Gallanis, *The Rise of Modern Evidence Law*, 84 IOWA L. REV. 499, 506–07 (1999) (noting the comparative emphasis on written over unwritten evidence in Gilbert's treatise).

[201] Langbein, *supra* note 198, at 1183 ("Ryder's trial practice 69 reflects the preoccupation with written evidence that we find in Gilbert and the other eighteenth-century writers."); T. P. Gallanis, *The Rise of Modern Evidence Law*, 84 IOWA L. REV. 499, 511 (1999) ("Evidentiary practice in civil trials focused principally on questions of written evidence.").

[202] *See generally* George Fisher, *The Jury's Rise as Lie Detector*, 107 YALE L.J. 575 (1997) (discussing this incompetency rule and its predecessors in detail).

[203] *See* Langbein, *supra* note 198, at 1184–86 (discussing this competency rule).

[204] *See* Fisher, *supra* note 197, at 624.

[205] G.S., *Competency of Witnesses*, 10 AM. L. REG. 257, 265 (1862) ("The rule is that a present interest in the event of a suit excludes the witness. But it must be a certain interest, and then no matter how small it is.").

[206] Gallanis, *supra* note 200, at 507 (paraphrasing Gilbert on an incompetence category compose of "those lacking in discernment, included the mentally retarded, the insane, and children under the "age of common knowledge").

[207] G.S., *supra* note 205, at 264 (summarizing the rule that "judgment against any person for treason, felony, or the *crimen falsi*, renders him incompetent to testify").

[208] Fisher, *supra* note 197, at 624.

[209] *See id.* at 625 (describing these rules as "declaring certain witnesses to be likely liars as a matter of law").

[210] Gilbert's treatise puts the matter in essentially these same terms. GEOFFREY GILBERT, JAMES SEDGWICK, & CAPEL LOFT, THE LAW OF EVIDENCE 106 (6th ed. 1801):

> Now where a man who is interested in the matter in question, would also prove it, it is rather a ground for distrust than any just cause of believe; for men are generally so short sighted, as to look at their own private benefit which is near to them, rather than to the good of the world, that is more remote; therefore, from the nature of

Finally, a variety of additional rules and instructions apparently stood ready to relieve jurors of the need to identify a lie in cases where conflicting sworn testimony did manage to come before the court. One example was an instruction sometimes given to jurors that they should attempt, where possible, to reconcile sworn testimony so that their interpretation did not require assuming that either witness was lying.[211] Another was an occasional suggestion that jurors should resolve conflicts in sworn testimony by counting the number of witnesses for an against a proposition rather than by trying to evaluate the individual credibility of each witness.[212]

Like trial by ordeal before it, this approach to trial process eventually came to an end. Starting in the 1840s, a series of legislative acts on both sides of the Atlantic toppled one competency rule after another.[213] Concerns about fairness, the need for information, and other less obvious considerations[214] motivated these retractions. But these concerns could not plausibly have escaped legal thinkers during the long tenure of the competency rules. A better explanation of the previous practice seems to be a kind of testimonial skepticism, i.e. the simple doubt that a jury of laypeople would be able to distinguish truth from well-presented fiction in the trial context. If the jury's handling of contradictory testimony was really nothing more than a random guess about who was telling the truth, then plausibly the accuracy of verdicts in the presence of that conflicting testimony would be no better than the quality of verdicts in its absence.

Contemporary trial practice evinces a willingness to entrust judges and juries with the testimony of every witness who appears. This reflects recent confidence in the power of human agents to sort truth from fiction. Richard Uviller once said, "At the heart of our adversarial mode of adjudication lies the *assumption* that trial jurors—a fair mix of ordinary, relatively openminded folk—can from across the jury rail distinguish liars from truthtellers."[215] Yet, decades of research shows fairly consistently

> human passions and actions, there is more reason to distrust such a biassed testimony than to believe it; it is also easy for persons who are prejudiced and prepossessed, to put false and unequal glosses for what they give in evidence, and therefore the law removes them from testimony, to prevent their sliding into perjury; and it can be no injury to truth, to remove those from the jury, whose testimony may hurt themselves, and can never induce any rational belief.

[211] *See* Fisher*, supra* note 197, at 626–33 (describing "the Rule of Bethel's Case").

[212] *Id.* at 653 ("Almost every major treatise suggested that whenever jurors faced the task of choosing between conflicting oaths, they should tend to give more credit to the side that produced the greater number of witnesses.").

[213] *See id.* at 658–59 (listing several of the relevant acts); G.S., *supra* note 205, at 257 (observing at the time of writing that "Practically now in the English courts all persons are competent witnesses, their credibility being left to the jury").

[214] *See id.* at 662–97 (describing how racial considerations interacted with rules of witness competency).

[215] H. Richard Uviller, *Credence, Character, and the Rules of Evidence: Seeing Through the Liar's Tale*, 42 DUKE L.J. 776,780 (emphasis added); *see also id.* at 776–77 ("Our faith in the adversary

that most people do only a little better than a coin flip in discerning truthful statements from lies.[216] This is true even when factoring in demeanor and other non-verbal evidence supposedly indicative of a witness's state of mind. Indeed, some research suggests that a focus on demeanor cues does more harm than good in helping jurors identify the truth.[217] It seems jurors are no better at detecting lies than they are at detecting sophisticated deepfakes.

How, then, could we possibly justify courts' enduring reliance on witness testimony? The epistemological frameworks discussed in Part III can help diagnose the problem. To perform their task well, jurors need to form justified beliefs about the content of witness testimony. While a witness' words and demeanor may be important evidence, studies show that they are insufficient. The fact that some witnesses are skilled liars defeats the evidentiary value of witness words and demeanor by reducing the signal that they can send. On a systems level, this puts court verdicts at risk of being epistemically unjustified because they result from an unreliable process reliant on jurors assessing witness testimony.

The solution was to make courts' process more reliable by putting jurors in a better position to form justified beliefs about witness testimony. Ultimately, that means providing jurors more and better evidence on what witnesses say. One possibility (analogous to some proposals regarding deepfakes), would be to recruit expert lie detectors. Fortunately, courts did not go that route—experts don't perform much better than laypeople.[218] Rather, courts realized that there is evidence beyond a

system ... depends in large measure on our confidence that, assisted by courtroom procedure, our jurors will usually return a verdict consistent with the historical fact.").

[216] *See* David M. Markowitz, *Self and Other-Perceived Deception Detection Abilities Are Highly Correlated but Unassociated with Objective Detection Ability: Examining the Detection Consensus Effect*, 14 SCI. REP. 1, 2* (2024) ("Overwhelming evidence in the deception detection literature suggests that on average, people are often slightly greater than chance at lie-truth judgments. Deception detection accuracy tends to hover around 54%, with truths being evaluated more accurately than lies because people are truth-biased."); *see generally* Charles F. Bond & Bella M. DePaulo *Accuracy of Deception Judgments*, 10 PERS. & SOC. PSYCHOL. REV. 214, 217 (2006) (describing a large meta-analysis of deception studies in which average truth detection was measured at slightly above 50%).

[217] *See* Aldert Vrij & Jeannine Turgeon, *Evaluating Credibility of Witnesses – Are We Instructing Jurors on Invalid Factors*, 11 J. TORT L. 231, 233–37 (2018) (noting little evidence to support the "myth about the strong relationship between nonverbal behavior and deception"); Danielle Andrewartha, *Lie Detection in Litigation: Science or Prejudice?* 15 PSYCHIATRY, PSYCHOL. & L. 88, 92 (2008) (noting that nonverbal behavior like apparent nervousness is an especially questionable indicator of deception in the unnatural and confrontational setting of courtroom testimony); *see generally* Olin Guy Wellborn III, *Demeanor*, 76 CORNELL L. REV. 1075, 1075, 1091–93 (1991) (concluding that ordinary observers often cannot effectively use demeanor to assess truthfulness and that overreliance on demeanor can be misguided).

[218] Paul Ekman & Maureen O'Sullivan, *Who Can Catch a Liar?*, 46 AM. PSYCHOL. 913, 913 (1991)(summarizing twenty years of literature as providing little reason to trust human lie detection capabilities and reporting a study in which even professionals whose job involved lie detection typically fared little better than chance).

witness' words and demeanor that jurors are equipped to evaluate. Testimony does not exist as an isolated datapoint, sealed off from the justificatory web of evidentiary interdependencies that connects all truth. It is situated within a context, and threads from that context can provide additional evidence. Some threads will reveal (in)consistencies internal to the testimony. Others show whether the testimony is (in)consistent with truths external to the testimony itself. Courts just needed a process that would reliably generate such evidence and put it before the jurors. That process is the adversary trial.

Jurors do not assess testimony in isolation. They assess it against context supplied by other sources of information, like their common experience, cross examination, and evidence introduced by opposing counsel. Jeremy Bentham proclaimed, "Against erroneous or mendacious testimony, the grand security is cross-examination: cross-examination, by which, if the individual facts charged are false, true ones . . . may be brought out against them."[219] Wigmore opined that "no safeguard for testing the value of human statements is comparable to that furnished by cross-examination, [which] is beyond any doubt the greatest legal engine ever invented for the discovery of truth."[220] Of more recent vintage, the U.S. Supreme Court has described cross-examination as "the principal means by which the believability of a witness and the truth of his testimony are tested."[221]

Confidence in this power of cross-examination is justified, to some extent, by the robust rules of evidence that have developed since Gilbert's time to limit and focus witness testimony. The strictures of relevancy[222] and prohibition on character reasoning[223] discourage testimony from wandering away from material facts. Limits on the introduction of hearsay evidence force litigants to put witnesses before the jury.[224] A formidable machinery of impeachment and rehabilitation stands ready to test the credibility and plumb the character of every witness who takes the stand.[225] Layered atop of these testimonial screens, the threat and practice of skillful cross-

[219] JEREMY BENTHAM, RATIONALE OF JUDICIAL EVIDENCE, SPECIALLY APPLIED TO ENGLISH PRACTICE 212 n.* (Hunt and Clarke, 1995) (1827); *see also id. at* 230 ("Mendacious invention, then, having been either prevented, or encompassed with dangers, by the viva voce questions followed immediately by the viva voce answers . . . "); *id.* at 231 (arguing that rapid cross-examination, with consequently little time for careful fabrication, is the "only remedy" for mendacious invention).

[220] John Henry Wigmore, TREATISE ON THE SYSTEM OF EVIDENCE IN TRIALS AT COMMON LAW: INCLUDING THE STATUTES AND JUDICIAL DECISIONS OF ALL JURISDICTIONS OF THE UNITED STATES § 1367, at 1697 (1904).

[221] Davis v. Alaska, 415 U.S. 308, 316 (1974).

[222] *See* FED. R. EVID. 401–403 (defining relevance).

[223] *See* FED. R. EVID. 404 (prohibiting character evidence to prove conduct, with exceptions for specific purposes).

[224] Obviously, this forcing function is tempered by the many exemptions and exceptions to the hearsay rule. *See* FED. R. EVID. 801–807.

[225] *See* FED. R. EVID. 607–609, 613 (governing witness impeachment, including character for truthfulness and prior inconsistent statements).

examination surely does provide courts and juries with some impressive tools for spotting attempted deception.

The lesson is not that cross-examination is an unfailing engine of truth. It isn't. Deepfakes, like lies, are about deception. And people only try to deceive when they believe there is a chance they will succeed. But in evaluating the evidentiary challenge posed by deepfakes, we can't lose sight of the evidentiary challenges posed by *all* deceptive testimony. The possibility that every statement could be a lie is a challenge that has endured the centuries little diminished by anything the legal system has thought to throw at it. Rather than succumb to skepticism or minimize the lay juror's role, courts have leaned in by developing adversarial procedures to generate contextual information for enhancing jurors' truth-finding function.

B. The History of Photographs and Recordings

Deepfakes purport to be mechanical recordings of reality. But this deception is the just latest challenge in a long-running struggle to decide the evidentiary value of photographs and similar recordings. That struggle dates, unsurprisingly, to the proliferation of these technologies between the 1850s and 1950s. In the domain of recorded images, photography's reliance on a comparatively convenient and affordable paper medium soon won out over the daguerreotype's coated copper plates.[226] By the 1870s, it appears that nearly everyone, from all walks of life, had either sat for a photograph or at least seen photographs that had been taken of friends, family, and familiar places.[227] Audio recording devices developed over a similar timeline.[228] For technical and practices reasons, however, they found fewer applications in trial evidence before the proliferation of portable magnetic-tape recording devices in the 1950s.[229] Video recording matured and expanded in usage around the same time.[230] For our purposes, it is enough to trace the history of

[226] Library of Congress Collections, *The Daguerreotype Medium*, LIBRARY OF CONGRESS, https://www.loc.gov/collections/daguerreotypes/articles-and-essays/the-daguerreotype-medium.

[227] *See* Udderzook v. Com., 76 Pa. 340, 353 (1874) ("The Daguerrean process was first given to the world in 1839. It was soon followed by photography, of which we have had nearly a generation's experience. It has become a customary and a common mode of taking and preserving views as well as the likenesses of persons, and has obtained universal assent to the correctness of its delineations.").

[228] NAT'L PARK SERVICE, THE ORIGINS OF SOUND RECORDING, https://www.nps.gov/edis/learn/historyculture/origins-of-sound-recording.htm. (Mar. 29, 2023)

[229] *See* ROBERT C. MAHER, PRINCIPLES OF FORENSIC AUDIO ANALYSIS 29 (2018) ("The first portable recorders using magnetic tape appeared in the 1950s, and soon these devices were used to obtain clandestine recordings of interviews and wiretaps, as well as to record interrogations and confessions.").

[230] Judith Keilbach, *Instant TV. The Forgotten History of Video Tape Recording (and the Coverage of the Eichmann Trial)*, 24 J. MEDIA HIST. 1, 1–12. (2024).

photographs as evidence, since the themes and principles generalize in obvious ways to other types of recorded evidence.[231]

The reason that photographs posed—and still pose—a challenge for courts and the law of evidence is that they fit imperfectly into a trial process fortified against the challenges of written documents and live witness testimony. Familiar infirmities in human expression—errors in perception, memory, mendacity, and narration—were all arguably addressed by the developing law of evidence at the time photographs arrived on the scene.[232] But this new technology deviated just enough from human expression that it presented uncomfortable problems. Like all human expression, recordings could be manipulated to deceive. In all but the rarest cases, photographs were also intractably tethered to witness testimony. At a minimum, a witness was needed to supply the context in which the photograph was meant to be understood and interpreted. But unlike human expression, photographs held claim to special capacities like mechanical objectivity, super-human perceptive accuracy, and near-perfect recall. These features made them obviously and alarmingly potent evidence.[233]

Themes of the latter sort—scientific objectivity, precision, and unbiased truth-telling—were common in much of the early discussion of photographs as evidence. Writing in 1869, one author effusively propounded the value of photographs specifically for their mechanical advantages over live witness testimony: "The photographic apparatus never intentionally falsifies nor do its products ever so fade as to distort the image they present, as do the figures of things committed to the treacherous memory of men." [234] Similar statements can be found in other

[231] *See, e.g.*, Maher, *supra*, at 229, 67 (summarizing the legal treatment of audio recordings). VIDEO EVIDENCE: A PRIMER FOR PROSECUTORS, BUREAU OF JUST. ASSISTANCE, U.S. DEPT. OF JUSTICE (Oct. 2016), https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/final-video-evidence-primer-for-prosecutors.pdf.

[232] *See, e.g.,* Edmund M. Morgan, *Hearsay Dangers and the Application of the Hearsay Concept*, 62 HARV. L.R. 177, 177–78 (1948) (describing these infirmities in the context of hearsay evidence); Laurence H. Tribe, Triangulating Hearsay, 87 HARV. L.R. 957, 958–61 (1974) (same).

[233] See Jennifer L. Mnookin, *The Image of Truth: Photographic Evidence and the Power of Analogy*, 10 YALE J. L. HUMANS. 1, 4 (1998) ("In the second half of the nineteenth century, two competing paradigms governed the understanding of the photograph. One emphasized its ability to transcribe nature directly, while the other highlighted the ways in which it was a human representation. From the first perspective, the photograph was viewed as an especially privileged kind of evidence; from the second perspective, the photograph was seen as a potentially misleading form of proof.").

[234] J.A.J, *The Legal Relations of Photographs*, 8 AM. L. REG. 1, 5 (1869) see also Mnookin, *supra* note 43, at 6–7 ("[I]f a difference exist, should we not give the greater credence to the photograph, whose testimony, we know, is perfectly truthful and generally commensurate with the fact, while that of the vouching witness, and also of the witness called to speak to the question of identity, may be mistaken or perjured?").

commentary of the time.[235] Jennifer Mnookin summarizes this then-prevalent perspective succinctly: "[T]he photograph was not merely evidence, but the best kind of evidence imaginable: mechanical, automatic, and not subject to those biases and foibles that may cloud human judgment."[236]

Sympathetic judges found no difficulty admitting photographs as evidence. In an 1882 case in which a photograph of a victim's cut throat was put before the jury, the Supreme Court of Georgia easily brushed aside objections to the introduction of the photograph into evidence:

> [T]he character of the wound was important to elucidate the issue; the man was killed and buried, and a description of the cut by witnesses must have been resorted to; *we cannot conceive of a more impartial and truthful witness than the sun*, as its light stamps and seals the similitude of the wound on the photograph put before the jury; it would be more accurate than the memory of witnesses, and as the object of all evidence is to show the truth, *why should not this dumb witness show it*? Usually the photograph is introduced to prove identity of person, but why not to show the character of the wound? In either case it is evidence; it throws light on the issue.[237]

The Georgia Supreme Court was undoubtedly correct that photographs offered mechanical advantages over human testimony, but its uncritical analogy of the camera as "dumb witnesses" without capacity to lie was less persuasive, even in the 1800s.

Indeed, the historic record indicates that, from the start, even casual observers were aware of the potential for manipulation and deception when photographs were used as evidence.[238] A short and unapologetic critique of

[235] See, e.g., Rodney G.S. Carter, *"Ocular Proof": Photographs as Legal Evidence,* 69 J. ASS'N CANADIAN ARCHIVISTS 23, 27 (2010) ("From the mid-nineteenth century, and continuing well into the latter part of the twentieth century, a dominant strain of the discourse surrounding photography centred [sic] on its ability to objectively reproduce what was before the lens. Given its technological origins in optics and chemistry, photography was viewed as being the product of a scientific, and therefore truthful, process, and the earliest texts announcing the invention of photography in France and Britain emphasize its mechanical nature."); Mnookin, *supra* note 233, at 17 (noting that "[I]n the inaugural volume of the Philadelphia Photographer, one author described how the camera 'sees everything and it represents just what it sees. It has an eye that cannot be deceived and a fidelity that cannot be corrupted.'").

[236] Mnookin, *supra* note 233, at 19.

[237] Franklin v. State, 69 Ga. 36, 42–43 (1882) (emphasis added).

[238] Carter, *supra* at 235, 35–36 ("Staged and manipulated photographs – including photographs that had their negatives retouched, combined, or otherwise tampered with – were widely created and circulated from the very beginning of photographic history, and contemporaries readily understood the artifice employed in the creation of the images.").

photographic evidence appeared in a number of publications in 1886 under the title "The Photograph as a False Witness."[239] In this article, the anonymous author warns that unguarded acceptance of photographs as legal proof creates a danger of deception and perjury: "[T]he photograph may be made to speak for this or for that, according as the finger of mammon does point."[240] Careful selection of lighting, perspective, and equipment could be used to editorialize the content of a photograph in ways that an unsophisticated audience—or the Georgia Supreme Court—might not suspect.[241]

Post-exposure manipulation of photographs was also a concern well before "Photoshop" became a verb. In an 1861 article, Oliver Wendell Holmes (father of the later Supreme Court justice) quipped: "A simple photographic picture may be tampered with. A lady's portrait has been known to come out of the finishing-artist's room ten years younger than when it left the camera."[242] It seems this type of manipulation was widespread. In one sensational example from the 1860s, photographer William H. Mumler became the target of popular and legal controversy for his production of spirit photographs—portrait photos which, when developed, appeared to show spirits of the subjects' deceased relatives floating as ghostly apparitions above them.[243] Whatever technique Mumler used to doctor these photographs was clever enough to evade detection by experienced photographers who visited the studio to observe his process.[244]

The law of evidence eventually settled on handling photographs by analogy to paintings and other constructed representations of witness testimony.[245] The approach and its reasoning are well captured in an early and influential comment on the subject by the New York Court of Appeals:

[239] *The Photograph as False Witness*, PHOTOGRAPHIC NEWS, Jul. 23 1886 at 465; *The Photograph as False Witness*, 10 VA. L.J. 10 644 (1886); *The Photograph as a False Witness*, 34 ALB. L.J. 457 (1886). For additional references to reproductions in law journals, see Mnookin, *supra* note 233, at 26 n.94.

[240] *The Photograph as False Witness*, 10 VA. L.J. 10 644, 645–46 (1886).

[241] *The Photograph as False Witness*, 10 VA. L.J. 10 644, 645–46 (1886).
E.g., *id.* (providing an anecdote relating to ancient lights case); *see generally* CHARLES SCOTT, 1 PHOTOGRAPHIC EVIDENCE (1942) (illustrating how differences in composition could influence the resulting recordings).

[242] Oliver Wendell Holmes, *Sun-Painting and Sun-Sculpture*, ATLANTIC MONTHLY, Jul. 1861 at 13, 15.

[243] *See* Mnookin, *supra* note 233, at 27–43.

[244] *Id.* at 31.

[245] *Cf. id.* at 53–59 (considering ways in which the treatment of photographs diffused some of the discomfort that judges and jurors might otherwise have felt about fact-finding in a context bounded by photographs).

> A portrait or a miniature taken by a skilled artist, and proven to be an accurate likeness, would be received on a question of the identity or the appearance of a person not producible in court. Photographic pictures do not differ in kind of proof from the pictures of a painter. . . . It is the skill of the operator that takes care of [details like lighting, position, and equipment], as it is the skill of the artist that makes correct drawing of features, and nice mingling of tints, for the portrait. . . . So the signs of the portrait and the photograph, if authenticated by other testimony, may give truthful representations. When shown by such testimony to be correct resemblances of a person, we see not why they may not be shown to the triers of the facts, not as conclusive, but as aids in determining the matter in issue, still being open, like other proofs of identity or similar matter, to rebuttal or doubt.[246]

Put another way, the photograph, like the painting, could be authenticated by a testifying witness as an illustration of that witness's testimony.[247] Somewhere in the background, the photograph still retained its mechanical advantages. But, in the legal theory of the trial, these advantages were set aside as the photo's purpose was merely to help lend color and detail to a witness' spoken words. It was the testimony, not the photograph, that was the evidence before the court.[248] Any risk of deception was thus no different from the traditional risk of false testimony, addressed by existing rules and procedures that policed the accuracy of what witnesses said in the stand.

This limited and rather artificial understanding of photographic evidence survives today as what is sometimes called the "pictorial testimony" use of photographic evidence. In this approach, a photo, video, or similar recording is introduced at trial for the purpose of illustrating a witness's testimony,[249] usually after being authenticated by that witness as a fair and accurate representation of her testimony.[250] There is, in principle, no difference between a candid and a staged photograph in this approach; both are merely illustrations of what the witness is trying to explain. Indeed, photographs introduced only to illustrate a point are commonly

[246] Cowley v. People, 83 N.Y. 464, 477–78 (N.Y. 1881).

[247] *See id.* at 478 ("When shown by such testimony to be correct resemblances of a person . . . .").

[248] *See*, *e.g.*, Mnookin, *supra* note 233, at 44–45 (citing late 1800s authority for this understanding of photographs).

[249] *See* 22 Charles Alan Wright & Arthur R. Miller, Fed. Prac. & Proc. § 5172.4 (2d ed. 2024).

[250] *E.g.*, People v. Bowley, 59 Cal. 2d 855, 859 (1963) ("It is well settled that the testimony of a person who was present at the time a film was made that it accurately depicts what it purports to show is a legally sufficient foundation for its admission into evidence.").

said to be "not evidence" at all.[251] Such records have, in theory, no evidentiary weight and would typically not be made available to the jury during deliberations.[252]

At the opposite extreme, another modern use of photographic evidence often goes under the label of the "silent witness" theory.[253] In this approach, a photograph, video, or similar recording is authenticated by a witness with knowledge of its source to be the output of a system that produces reliable results.[254] It may then be introduced as substantive evidence of its content. As one common example, the maintainer of a bank's closed-circuit surveillance-camera system could take the stand to explain how the system works and why its recording of a robbery could be trusted as an accurate depiction of what took place.[255] So authenticated, the recording's probative value would arise directly from its unthinking, mechanical transcription of the world, not merely from its derivative value in illustrating the first-hand testimony of a human witness.[256] Subject to other relevant rules of evidence,[257] the silent-witness recording

[251] *E.g.*, FED. R. EVID. 107 ("An illustrative aid is not evidence…."); JOHN HENRY WIGMORE, WIGMORE'S CODE OF THE RULES OF EVIDENCE IN TRIALS AT LAW (3d ed. 1942).

[252] *See, e.g.*, FED. R. EVID. 107 ("An illustrative aid is not evidence and must not be provided to the jury during deliberations unless: (1) all parties consent; or (2) the court, for good cause, orders otherwise."); WRIGHT & MILLER, 22 FED. PRAC. & PROC. EVID. § 5174 (2d ed. 2024) ("[M]ost courts seem to follow the suggestion by the commentators that illustrative objects should not be sent to the jury room during deliberations").

[253] People v. Bowley, 59 Cal. 2d 855, 860 (1963) ("[P]hotographs are useful for different purposes. When admitted merely to aid a witness in explaining his testimony they are, as Wigmore states, nothing more than the illustrated testimony of that witness. But they may also be used as probative evidence of what they depict. Used in this manner they take on the status of independent 'silent' witnesses.").

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

[255] *E.g.*, United States v. Clayton, 643 F.2d 1071, 1073 (5th Cir. 1981) ("[P]hotographs made from bank camera films were sufficiently authenticated by Government witnesses who were not present at the robbery when the testimony adduced stated the manner in which the films were used in the camera, how the camera was activated, that the film was removed immediately after the robbery, and the chain of possession of the film and the development of the prints.").

[256] *E.g.*, United States v. Taylor, 530 F.2d 639, 641–42 (5th Cir. 1976) ("In the case before us it was, of course, impossible for any of the tellers to testify that the film accurately depicted the events as witnessed by them, since the camera was activated only after the bank personnel were locked in the vault. The only testimony offered as foundation for the introduction of the photographs was by government witnesses who were not present during the actual robbery. These witnesses, however, testified as to 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 made from it. Under the circumstances of this case, we find that such testimony furnished sufficient authentication for the admission of the contact prints into evidence.").

[257] *E.g.*, FED. R. EVID. 1001–1004 (requiring the production of originals or mechanical duplicates of a recording in most such circumstances).

could be introduced as substantive evidence itself, essentially as the testimony of the type of "dumb witness" that the Georgia Supreme Court imagined.

The space between photographs as "pictorial testimony" and photographs as "silent witnesses," remains uncomfortably wide. Photographs introduced as pictorial testimony obviously convey gratuitous details beyond the words being uttered by the authenticating witness. Testimony that such a photo is a "fair and accurate representation" of the scene typically does nothing to establish the value of its fine details, and it is fantasy to believe that jurors interpret such photographs as mere illustrations to be doubted in every respect.[258] At the other extreme, photographs introduced under silent witness theories may fail to disclose their exposure to human manipulation. Even setting aside more complicated issues, like selection bias when interested parties identify and produce photographic evidence, the simple capacity images and video recordings to be manipulated often seems to go underexplored, a concern that has spawned decades of frustrated legal commentary.[259]

As background context for the future treatment of deepfakes, the history of photographic evidence is again a curious mix of causes for concern and comfort. Deepfakes are photographic manipulation carried to its logical extreme. But opportunities for manipulation, deception, and simple overweighing of photographic evidence have existed since the dawn of this technology. Whether and how deepfakes are really all that different is the subject we next consider.

V. Deepfakes and Proposed Reforms

Deepfakes are novel. They are shocking. And they are generating a buzz of worried analysis and calls for reform in academic and legislative-policy circles (Part

[258] *Cf.* Mnookin, *supra* note 233, at 26 ("If the photograph was properly understood as equivalent to any other form of human testimony, then the widespread belief in inherent photographic certainty might make the legal use of this new technology highly misleading.").

[259] See, e.g., WRIGHT & MILLER, 22 FED. PRAC. & PROC. EVID. § 5172.4 Demonstrative Evidence—Photographs (2d ed. June 2024 Update) ("[I]t is rare to find a federal court excluding photographic evidence. So far as we can detect, the availability of computer programs that can fake photographs has not made courts any more cautious about admitting photos."); 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, 271–72 (2002) ("Digital images are highly susceptible to manipulation. Manipulation, as distinct from enhancement, consists of changing the elements of a photograph or image by changing the colors, moving items from place to place on the image, or otherwise altering the original image. . . . The electronic nature of the image file makes undetectable manipulation of a digital image easy, in part because no traditional "original image" is made. Unlike traditional cameras, which produce one negative, digital cameras create an electronic file from which the image can be generated.").

II.C). But deepfakes are also just the newest version of the common lie. We humans have been guarding ourselves against lies and other acts of trickery for a very, very long time (Parts III and IV). How do deepfakes stand when viewed through the lenses of epistemology and the law of evidence? Are new laws and social interventions as urgent and as necessary as they appear to be?

For the most part, we think not. In the following pages, we evaluate common justifications for alarm and corresponding proposals for policy reform. We argue that the case for panic is overstated; the justifications for reform are insubstantial. We do not deny that deepfakes present new and worrying opportunities for deception in the courtroom. And deception should never be treated lightly—least not in as important a social context as trials. But to accord deepfakes appropriate gravity is not necessarily to treat them differently than other forms of lies and deception. The novel expression of an ancient problem does not necessarily require novel solutions.

Our analysis draws on justificatory frameworks from both evidentialism and reliabilism. We assume that in the courtroom, factfinders use evidentialist methods to form the beliefs that determine case outcomes. In so doing, we are only taking courts at their word when, for example, they instruct jurors to "Your first duty is to decide the facts from the evidence in the case."[260] When we evaluate existing or proposed rules of evidence, we employ a reliabilist point of view. In other words, we assess rules of evidence by how reliably they enable evidentialist jurors to exercise human judgment in arriving at the truth.

Our conclusion is that knee-jerk proposals in the literature tend to focus on deepfakes as isolated pieces of evidence. They mistakenly assume that deepfakes will always bear some mark of their false provenance. Or they forget that, like any piece of evidence, digital media need not bear their own mark of authenticity to be deemed trustworthy. All evidence is situated within a web of co-dependencies, and the law has long relied on human judgment about context to help disentangle fact from fiction.

Before turning to the arguments, one clarification may be helpful. Our focus, here, is on the use of deepfakes to deceive. That is, generated media being presented to the judge and jury as if it were simple, mechanical recording of reality. This limited scope of analysis is important because different issues are raised by something like the clearly disclosed use of computer-generated content to illustrate a witness's testimony—what we call "deep fabrications." "Deepfabs" are interesting in their own right but they are not our focus in this Article. Different issues are likewise raised by the autonomous editing decisions of smart devices. Smartphones use filters, exposure settings, and post-processing to convert raw recordings of nighttime scenes into clear

[260] Southern District of Illinois, Court's Jury Instructions, https://www.ilsd.uscourts.gov/sites/ilsd/files/CourtsJuryInstructions.pdf.

and attractive photographs.[261] This type of transparent background editing by "silent *smart* witnesses" presents many interesting evidentiary and epistemological challenges. But these are, again, not our focus in this Article.

A. Conduct-Oriented Prohibitions and Penalties

Turning to deepfakes as deception, and corresponding proposals for reform, we can start with what might seem like the most targeted responses to the challenge: proposals that would prohibit deepfakes from being produced and distributed in the first place. Examples includes call for compelled origin-disclosure statements on all generated media[262] and calls to ban and penalize specific abuses of deepfakes (like the production of a video portraying a targeted person engaging in a sex act).[263] At the extreme, this strategy could be implemented as a flat ban on the production and distribution of *any* deepfake content.[264] Reframed in epistemological terms, these proposals evince confidence in the reliability of existing court rules, but worry about the future justificatory power of digital media as deepfakes dampen the truth signal digital media provides to jurors. Banning deepfakes boosts the signal of digital media that remain, making it easier for jurors to form justified beliefs based on it—or so the reasoning apparently goes.

Proponents of bans on deepfakes may appropriately aspire to address more than our specific focus on the deceptive use of deepfakes as trial evidence.[265] But if their proposals are to address this challenge, then they should at least provide some identifiable advantages over existing rules of evidence and related restrictions. For this to happen, two conditions would have to be satisfied. First, the existing legal safeguards would have to be inadequate to deter the introduction of deepfakes into evidence. Second, the proposed bans would have to offer credible improvements in deterrence over existing law.

The first of these conditions is almost surely satisfied. True, there are many deterrents to presenting false evidence in the courtroom. A lawyer cannot ethically

[261] *See, e.g.*, iPhone User Guide, *Take Great Photos and Videos*, https://support.apple.com/guide/iphone/take-great-photos-and-videos-iph9bbc8619e/ios ("Night mode automatically takes bright, detailed photos in low-light settings.").

[262] *See* Delfino, *supra* note 14, at 303 (describing an act that would have "mandated that most classes of deepfakes" would need to conspicuously disclose their fabrication, with penalties available to enforce this requirement).

[263] *See, e.g.*, Brown, *supra* note 20, at 45-47 (describing state law proposals for banning the use of deepfakes in attempting to influence elections and in generating sexually explicit content without consent).

[264] *Cf.* Chesney & Citron, *supra* note 22, at 1788-89 (arguing that "a flat ban is not desirable because digital manipulation is not inherently problematic").

[265] *See, e.g., id.* at 1771-86 (describing social, political, and other problems that could be caused by the proliferation of deepfakes).

mislead a court or facilitate the presentation of evidence that the lawyer knows or reasonably believes to be untrue.[266] Every witness who testifies must first "give an oath or affirmation to testify truthfully."[267] Since no purported recording can be introduced as evidence without being authenticated as accurate by the testimony of a witness with appropriate knowledge of its accuracy, known deepfakes cannot be introduced without someone lying to the tribunal and thus subjecting themselves to the penalty of perjury.[268] But the actual enforcement of penalties for perjury is infrequent at best,[269] and almost the entire history of the law of evidence betrays the commonsense understanding that promising to tell the truth is little obstacle to lying.[270]

The second condition is where the proposed reforms fall flat. If the oath and all related penalties for lying in court are not already adequate to prevent deepfakes from being presented as legitimate evidence, what contribution does one more rule against deception stand to make? Unless proposed legislation offers greater or more certain penalties for deepfake deception than for other examples of lying under oath, the promises of additional penalties are hard to spot.

For deepfake bans to have any additional deterrent effect, deepfakes must also be at least reasonably detectable. How else would production and promulgation be punishable except if the result was identifiably fake upon inspection? As we have already discussed, deepfake detection is an active area of research,[271] but the arms-race between deepfake detectors and generators looks unpromising for detectors.[272] For early deepfakes and crude manipulations, conduct-oriented bans may perhaps do some work. But in a world of undetectable deepest fakes, these interventions are entirely toothless.

B. Prophylactic Exclusionary Rules

If conduct-oriented prohibitions cannot stem the predicted tide of deepest-fakes, then the law of evidence is the next logical place to look for solutions. There

[266] *See, e.g.*, MODEL RULES OF PRO. CONDUCT R. 3.3 (Am. Bar Ass'n 2023) ("A lawyer shall not knowingly ... (3) offer evidence that the lawyer knows to be false. If a lawyer, the lawyer's client, or a witness called by the lawyer, has offered material evidence and the lawyer comes to know of its falsity, the lawyer shall take reasonable remedial measures, including, if necessary, disclosure to the tribunal. A lawyer may refuse to offer evidence, other than the testimony of a defendant in a criminal matter, that the lawyer reasonably believes is false.").

[267] *See* FED. R. EVID. 603.

[268] *See, e.g.*, Congressional Research Service, False Statements and Perjury: An Overview of Federal Criminal Law (Updated October 8, 2024), https://crsreports.congress.gov/product/pdf/RL/98-808.

[269] *See Perjury: The Forgotten Offense*, 65 J. CRIM. L. & CRIMINOLOGY 361 (1974). *But cf.* Chris William Sanchirico, *Evidence Tampering*, 53 Duke L.J. 1215 (2004) (presenting restrictions on lies and evidence tampering in a more optimistic light).

[270] *See supra* Part IV.B.

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

[272] *See supra* Part I.D.

is an immediate positive note, here, as increasingly searching scrutiny of things presented as standard mechanical recordings seems likely to be an organic byproduct of the adversarial system's growing awareness of deepfake technology.

The reason for this is simply that the authentication standard is a fact question embedded in the changing social context. In order to introduce photos, videos, and other recordings into evidence, the proponent must be able to defend the authenticity of the evidence as being what the proponent claims it is.[273] The proponent must also persuade the factfinder to give that evidence whatever weight it deserves. It takes little imagination to see why opposing counsel, in a world where deepfakes are plentiful, would be more apt to challenge the authenticity of apparently recorded evidence than they are today.[274]

In litigating these challenges, proponents of deepfake-able evidence are also likely to be chasing increasingly demanding targets. To see why, consider the lowly authentication standard, usually articulated as requiring "evidence sufficient to support a finding [by a preponderance of the evidence] that the item is what the proponent claims it is."[275] Now consider this standard in relation to an audio file that presented as a recording of the defendant's verbal confession. In a world without deepfake voice generation, this audio file could be convincingly authenticated by simple means. The jury could compare the recorded voice to that of the defendant in deciding that the recording probably was the defendant's spoken words.[276] But in a world of deepest fakes, that simple demonstration may fail to persuade. Even if the evidence is admitted, the jury may assign it little weight out of fear that it could have been artificially generated by the prosecution. Authenticity and persuasion are both context-dependent requirements, and as the ease of producing deepfakes increases, it is only natural to suppose that factfinders will grow increasingly skeptical of the purportedly recorded evidence put before them.

Some commentators fear the realization of this prediction—a reaction we take up in the next section. Others demand more than what organic change promises to produce. These commentators propose changes to the law of evidence to further reduce the opportunities for deepfake deception.[277] One version of this proposal

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

[274] *See* Pfefferkorn, *supra* note 30, at 268 (predicting more frequent litigation of authenticity to mean that "successfully getting a video admitted into evidence may require additional motion practice, witness testimony, and forensic tools").

[275] FED. R. EVID. 901.

[276] *See* FED. R. EVID. 901(b)(5).

[277] *E.g.*, Delfino, *supra* note 14, at 297 ("The current Rules [of Evidence] will need to be adapted to solve the problem of how to show when a video is fake and when it is not."); *id.* at 332 ("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 . . . .").

would withdraw the "silent witness" theory of authentication altogether.[278] Broader yet, evidence law could be changed to prophylactically exclude all deepfake-able evidence, perhaps on the reasoning that it is impossible to demonstrate that such evidence is conditionally relevant.[279]

In epistemological terms, these proposals sound more in the vein of digital media skepticism. If we cannot prevent deepest fakes from proliferating, the thinking goes, digital media will eventually carry a very low truth signal. As a consequence, jurors will not be able to form justified beliefs on the basis digital media evidence. To maintain courts' reliability as a truth finding process, digital media must be excluded from trial or its use severely limited.

Here, again, we see these proposals as poorly calibrated to the challenge they purport to address. The problem is not that they are necessarily misguided, but that they overstate the severity of the deepfake threat—and even the deepest fake threat. They do this by failing to account for context in how the evidence will be assessed as authentic or fake.

To illustrate, imagine the trial of a civil action arising from a car collision at an intersection. The plaintiff wishes to introduce a video recording that purportedly shows the light was green as the plaintiff's car entered the intersection. This video was shot on the smartphone of a disinterested third-party witness. This third party was trying to record a video of her dog doing a trick but accidentally caught footage of the collision in the background. The witness takes the stand and testifies that she did not observe the collision when it happened (her attention was on the dog), but she is sure that the video was made using default settings on her phone. She observes the video and testifies that it looks today exactly as it did when she filmed it. She also produces her phone for inspection; the recording, still present in her photos reel, is identical in every way to the video file that the plaintiff seeks to introduce as evidence.

Our question for digital media skeptics is this: Is the mere technical feasibly of deepfake-video generation sufficient ground for excluding the video evidence in this hypothetical? Just at the conceptual possibility of Descartes' demon is insufficient to justify external world skepticism, we think the answer is an emphatic "No." True, the scene could have been generated from nothing more than an AI prompt. But why would a disinterested third-party generate a false video and then lie about the doctored origin of that video under oath, all in relation to a legal dispute of no interest to her? Broadening the reasoning beyond this example, why should the technical

[278] *See id.* at 341 ("[T]he 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"); *see also* Danielle C. Breen, Silent No More: How Deepfakes Will Force Courts to Reconsider Video Admission Standards, 21 J. HIGH TECH. L. 122, 160 (2021) ("Absent significant deepfake legislation, courts should adopt the pictorial evidence theory to combat heightened public skepticism of photographic and video evidence.").

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

feasibility of deepfake content generation pose a problem for introducing *any* evidence that is credibly authenticated by disinterested witnesses?

A useful way of reframing these types of proposals to prophylactically exclude deepfake-able evidence is to note that this is simply a modern refashioning of the old competency rules of witness testimony. As we have previously discussed, trial practice once struck witnesses from the stand in contexts where they were not expected to be truthful despite their promise not to lie.[280] Prophylactic exclusion rules operate the same way and for essentially the same reason. This framing also helps to illustrate the problem with the recent proposals. Not even the harshest competency rules excluded *all* witnesses from giving testimony. Parties and other interested witnesses were, for a time, excluded because of their special motivation to lie.

We would not go so far as to endorse revival of old competency rules for the problem of deepfake evidence. The deficiencies of the competency system were not limited to its dubious utility in helping juries resolve hard cases.[281] But if prophylactic exclusion is ever deemed a worthy intervention to pursue, then we would at most suggest that the new exclusionary rules be limited to evidence produced by witnesses with a plausible motivation to lie under oath.

C. The "Deepfake Defense" and Juror Skepticism

Finally, recent discussions about deepfakes and their role in trials have raised alarm about the possibility of a blanket deepfake defense that could be lobbed against even genuine and accurate media evidence.[282] A related line of reasoning worries that digital media skepticism will overtake jurors in a world where deepfakes are everywhere. For the most part these concerns are presented as systemic worries about how deepfakes will change the way that factfinders process evidence. Of a more functional mindset, one recent proposal attempts to tackle the deepfake defense and juror skepticism by amending the rules of evidence to concentrate authentication decisions in the hands of judges.[283] Under the proposed rule, the judge who finds a

[280] *See supra* notes 202–210 and accompanying text.

[281] *See* Fisher, *supra* note 197, at 662-97; *see also* Langbein, *supra* note 198, at 1185 (describing even "disqualification for interest" as "a grievous shortcoming in common law civil procedure").

[282] *E.g.*, Delfino, *supra* note 14, at 310 ("This "deepfake defense" will debut in court in the foreseeable future, if it has not already."); Pfefferkorn, *supra* note 30, at 255 (commenting that "The opponent of an authentic video may allege that it is a deepfake in order to try to exclude it from evidence or at least sow doubt in the jury's minds."); Chesney & Citron, *supra* note 17, at 1785 ("As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authentic video and audio as deep fakes. Put simply: a skeptical public will be primed to doubt the authenticity of real audio and video evidence. . . . Hence what we call the liar's dividend: this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes.").

[283] Delfino, *supra* note 14, at 341-42.

proffered item of evidence authentic would instruct the jury not to doubt its authenticity.[284] Epistemologically, such proposals, like those just considered, worry that digital media skepticism is inevitable outside of court. However, they believe that more robust gatekeeping in the courtroom could preserve the truth signal digital media send at trial.

We understand the bases for predicting increased invocation of the deepfake defense and rising juror skepticism—but we struggle to understand why either of these is a problem. In considering the deepfake defense, it cannot be forgotten that *every* item of evidence is vulnerable to attack for its lack of authenticity, accuracy, and reliability. True, one party can now claim that the other party's evidence is a deepfake. But what is the significance of this claim if credible evidence is not available to back it up? Return, again, to the third-party auto-collusion footage. Would a claim that the video is fake, based on nothing but theoretical possibility of deepfake generation, be worthy of deep analysis? Courts that have responded to this question to date offer an answer that largely mirrors our own: the mere conceptual possibility of fabrication is insufficient to support an authenticity challenge.[285]

But what about runaway juror skepticism? Suppose the overwhelming spread of deepfakes in everyday life pushes jurors to the point of the accuracy of all media evidence. We concede that this is a dystopian vision of the world. But we, again, fail to see why is a problem. Trials nearly always present jurors with conflicting evidence of variable quality, some of questionable reliability. Jurors are expected to bring their common sense, their life experience, and their own healthy skepticism to the task of evaluating the evidence they are presented with at trial. If the community at large comes to distrust media evidence, then that distrust can and should make its way into jurors' trial deliberations. Far from a problem to be corrected, this is the system working as intended.

VI. Conclusion: It's Not that Deep

[284] *Id.* at 342 ("The court would . . . admonish the jury to weigh that evidence, but not question its authenticity.").

[285] E.g., People v. Foreman, 2020 IL App (2d) 180178-U, ¶ 145 ("Defendant . . . points out that improperly-authenticated recordings are inherently suspect in this age of deep-fake videos and easily-manipulated audio records. We reject defendant's arguments. . . . We also reject defendant's argument that recent technological advancements render all recordings suspect, because they can be easily manipulated. In the absence of any evidence of tampering or other such manipulation in this case, there are no foundational issues with the recordings."); Pittman v. Commonwealth, No. 0681-22-1, 2023 WL 3061782, at *6 (Va. Ct. App. Apr. 25, 2023 ("[T]here is no evidence of or contention that would call into question the veracity of the video or the possibility of a 'deep fake.' And we reiterate that where there is 'mere speculation that contamination or tampering could have occurred, it is not an abuse of discretion to admit the evidence and let what doubt there may be go to the weight to be given the evidence.'" (quoting Reedy v. Commonwealth, 9 Va. App. 386, 391 (1990))).

What lessons does our tour through epistemology and evidence law hold for how society will respond to deepfakes? Alarmists warn of an epistemic apocalypse that will erode not only our ability to know true things, but perhaps our very concept of truth. Law scholars in this camp predict that our adjudicatory practices will fare no better. Because "[t]he courtroom is a microcosm of society in general,"[286] people bring their everyday frameworks and assumptions with them when they become judges and jurors. If a post-truth dystopia reigns on the outside, it must reign in the courtroom too.

In this concluding Part, we reverse the courtroom-society analogy. If courtrooms mirror society, we can also look to adjudication for insights that could unfold in ordinary life. Since the beginning of modern evidentiary practice, courts have grappled with the fact that most evidence is created by humans, and human creations can be tools for deception. The duplicitous twins of testimony, documents, and photographs are perjury, forgeries, and the fauxtographs. As discussed above, the predecessors of deepfake alarmists were equally concerned that each new type of manmade evidence would irreversibly corrupt courts' truth finding function. These alarmists defended varieties of philosophical skepticism about whether the artifacts they considered could ever facilitate truth-finding in court.

History has closed the book on generations of alarmists. Today, testimony, documents, and photographs are alive and well as fixtures of evidentiary practice. This is not because alarmists were wrong about the nature of these artifacts. Since the 1860s, professional photographers could generate false images that no expert could detect. The capacity to lie is even older—and requires no technical expertise.[287] The moral that courts repeatedly affirm and that alarmists seem periodically to forget is that human epistemic judgment is nuanced, flexible, and crucially multivariate. We look beyond the four corners of a statement, document, or photograph to evaluate its truth. We can catch out a lie not because we are astute lie detectors (we are not) but because we recognize that the statement is embedded within a web of other evidence, contexts, and commonsense intuitions that bear on its veracity. The epistemic weight we attach to different nodes in this web constantly evolves in light of individual and collective experience.

Courts' historical solution to falsifiable evidence has not been less evidence, but more. This, we have predicted, will be their response to deepfakes too. The existence of deepest fakes does not mean that jurors will be left with coin tosses and guesses when presented with digital media evidence. Rather, jurors will simply require more before believing and will become more astute judges of the fuller array of considerations that bear on truth.

[286] Pfefferkorn, *supra* note 30, at 257.

[287] *See* Mark Twain, private communication to Margery H. Clinton on August 18, 1908, available at https://twainsgeography.com/node/11416 ("'Let a sleeping dog lie.' It is a poor old maxim, & nothing in it: anybody can do it, you don't have to employ a dog.").

Before envisioning how deepfakes will play out beyond the courtroom, consider a related example from recent memory. In the not-too-distant past, people could generally trust the claims of disinterested third parties. Where's the best burger in town? What's a home remedy to whiten my teeth? Where was Kamala Harris born? There was never a guarantee that the response of a random passerby would be true, but you could generally trust her to provide her best answer, if she answered at all. As discussed above, this trust is grounded in the generally accepted norms of sincerity and competency that govern the ethics of interpersonal testimony. The internet changed this, or at least limited its scope. The information superhighway is not the country road. Trolls were born in the faceless corners of the web. Unchecked by any reputational consequences for violating the norms of sincerity and competence, they asserted falsehoods simply to bait clicks or stoke response. Unwary netizens could be misguided by outdated testimonial expectations to believe what the trolls wrote. Many an uncle has pounded the Thanksgiving table, swearing by some preposterous claim he read in some unremembered online forum.

But the troll did not kill testimony. Rather, society simply updated its epistemic norms to weaken the prima facie evidentiary value accorded to online statements—the brute fact that some text appears on the internet is little reason to believe what it says. Now everyone knows, you can't believe everything you read on the internet. Before believing it, people must exercise judgment that contextualizes the statement. It is relevant, for example, who made the claim, what their interests are, where the statement appears, how it connects to other facts, whether it conforms to common sense, etc. The trolls are still out there, and their knowing falsehoods are still indistinguishable on their face from everyone else's attempted truths. Many netizens still fall prey to their tricks, particularly when motivated reasoning clouds sober judgment. But the trolls' epistemic power has diminished as we have become more astute consumers of online text.

The same natural cycle of epistemic updating will play out as society emerges from the present alarm over deepfakes. Just five years ago, people could generally trust what they saw in videos. Of course, with big budgets, lots of time, and sufficient expertise, highly motivated actors could use computer-generated imagery to create persuasive fakes. Since most videos didn't and couldn't implicate these concerns, the signal videos sent to people searching for truth were pretty high. Deepfakes change this, and as their prevalence increases, the average signal that videos send will weakening. The bumpy period is now, as the number of deepfakes grows faster than the public's awareness of them. The signal is weaker, but not everyone knows it. This is, of course, rapidly changing. English-language corpora searches show that references to deepfakes are doubling every two years,[288] and references in news outlets

[288] Google Books Ngram Viewer, https://books.google.com/ngrams/graph?content=deepfake&year_start=2015&year_end=2022&corpus=en&smoothing=3&case_insensitive=false (last visited Jan. 21, 2025).

are currently doubling every year.[289] Growing awareness of deepfakes will "alter our trust in audio and video for good."[290] We anticipate the imminent rise of the dinner table meme, "You can't trust everything you see in a video."

That is where the alarmists end the story, but it is not the story's end. You can see their error most clearly in statements like: "If viewers cannot distinguish authentic videos from fabricated ones on their own, they will be disinclined to trust *any* video."[291] Not trusting everything is a far cry from trusting nothing. As jurors have done for testimony, documents, and photographs, and as society has recently done for text online, we will learn to use contextual factors to discern high signal from low signal content. The alarmists are right that society will become "a skeptical public . . . primed to doubt the authenticity of real audio and video evidence."[292] But they are wrong that this means people will become skeptics. People will simply become more astute consumers of digital media.

More concretely, this means the human element will become more important for truth finding, not less. Rather than moving passively from seeing to believing (the old model) or from seeing to disbelieving (the alarmist prediction), consumers will intervene in their own belief-forming processes with increasingly refined judgment. They will pause to look beyond the four corners of their video players to the sort of factors that digital media literacy advocates[293] and scam advisories[294] have long promoted. Is the content too good, or bad, or bizarre to be true? Are the stakes high? Am I presently in a calm state of mind? Where am I accessing the video?[295] Is the source trustworthy? Is the video asking me to do anything? Do other sources confirm the content? None of these questions, in isolation or combination, can provide certainty. Certainty was never the goal. But as we come to ask them, we will become as discriminating consumers of video as we already are of testimony and text. In a

[289] NOW Corpus (News on the Web), https://www.english-corpora.org/now/ (last visited Jan. 21, 2025).

[290] Donie O'Sullivan, *When Seeing Is No Longer Believing, Inside the Pentagon's Race Against Deepfake Videos*, CNN Bus. (Jan. 2019), https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

[291] Nina I. Brown, *supra* note 20, at 1.

[292] Chesney & Citron, *supra* note 22, at 1785. *See* Delfino, *supra* note 23, at 1082 ("As public knowledge of deepfakes continues to grow and people become increasingly skeptical about the credibility of audiovisual images").

[293] MediaSmarts, *Digital Media Literacy Fundamentals*, https://mediasmarts.ca/digital-media-literacy/general-information/digital-media-literacy-fundamentals (last visited Jan. 21, 2025).

[294] Minnesota Attorney General Keith Ellison, *How to Spot a Scam*, https://www.ag.state.mn.us/consumer/publications/howtospotascam.asp (last visited Jan. 21, 2025).

[295] "You don't have to become a detective. You don't have to become a forensic analyst. Just get off of social media. You will thank me." Bill Chappell, *LA's Wildfires Prompted a Rash of Fake Images. Here's Why*, NPR (Jan. 16, 2025) (quoting Berkeley Information Professor Hany Farid), https://www.npr.org/2025/01/16/nx-s1-5259629/la-wildfires-fake-images.

sense, deepfakes will be an epistemic boon rather than the epistemic harm philosophers fear. They will force us to become better believers.