Nothing You See Is Evidence Anymore

The provenance crisis: when real and fake become indistinguishable, truth becomes a luxury good.

Kymata Labs Research · An independent research institution studying how AI systems actually behave in production.

~12 min read

Tags · Deepfakes · Trust & Provenance · AI Fraud · Digital Evidence

We have crossed a line and there is no walking back over it: any image, any voice, any video can now be faked well enough to fool the experts. Spotting the fakes was never going to be the hard part. The hard part is that real evidence has lost its power. Drawing on the 2024 Arup deepfake-CFO fraud, in which a finance worker in the firm's Hong Kong office wired roughly US$25.6 million after a video call where the "CFO" and every colleague were AI; FTC data showing imposter scams were the most-reported fraud category at $2.95 billion amid record $12.5 billion losses; the Chesney–Citron "liar's dividend"; the so-far-failing "it's a deepfake" courtroom defence; and research finding humans detect deepfakes at chance while only 0.1% of people spot every fake, this paper argues that the moment anything *can* be fake, the guilty get a universal alibi and the innocent lose the ability to prove anything happened. Truth stops being the default state of the world and becomes something you have to buy.

The argument, in moves

1. **The cost of a perfect fake has fallen to zero.** Generative tools collapse the three barriers that once made a convincing fake of a specific person rare, skill, time, and a source image, all at once. Anyone can now produce video and voice of anyone, saying anything, in minutes.
2. **The moment anything can be fake, two things happen at once.** The guilty get a universal alibi, the words "it's a deepfake." The innocent lose the ability to prove that anything happened at all.
3. **The deepfake fraud already happened, with names, dates, and dollar figures.** A multinational lost roughly US$25.6 million to a video call full of AI ghosts. Imposter fraud set records the same year.
4. **Detection cannot save us.** Every detector that gets good becomes the training target that teaches the next fake to slip past it, and even a perfect detector leaves the deeper wound untouched, because by then the doubt is already planted.
5. **The fix is provenance, not detection.** The shift is from proving something is fake after the fact to proving something is real at the moment of capture, which is harder, slower, and unevenly available, which is precisely the problem.

The one-liner: the fakes were never the point. The doubt was. Real footage of real events now arrives pre-doubted, and the answer is not to win the impossible race to spot every fake, but to make truth provable again.

The strongest evidence: this already happened

None of what follows is a forecast; it is a record. The most concrete result in this set is not a study but a wire transfer. In January 2024, a finance worker in the Hong Kong office of the UK engineering firm Arup wired roughly **US$25.6 million (HK$200 million)** on the instruction of his chief financial officer [1]. He had his doubts. The request carried the smell of a phishing email, and he suspected as much, so he joined a video conference to be sure. On the call were the CFO and several colleagues he recognised. He saw them. He heard them. He wired the money. Every person on that call except him was an AI deepfake. The fraud surfaced only when he later checked with head office, and Arup publicly confirmed the loss in May 2024 [2]. The lesson is the one that should keep every operator up at night: the live video call, the very thing we reach for to *defeat* our suspicions, was the weapon.

The price of impersonation, in public data

The Arup story is not an outlier; it is the macro trend repeated a million times over. The FTC's 2024 figures, released in March 2025, recorded **more than $12.5 billion** in consumer fraud losses, up 25% year over year [3]. **Imposter scams were the single most-reported category, at $2.95 billion.** The share of people who reported a scam and actually lost money jumped from **27% to 38%**. Voice cloning is one of the drivers the FTC names, and the agency launched a Voice Cloning Challenge and an impersonation rule in response. We do not put a specific dollar figure on synthetic voice, because the FTC does not, but the trend line under "imposter" is bending the way the technology predicts.

The liar's dividend: predicted in 2019, collected ever since

Two legal scholars, Robert Chesney and Danielle Citron, saw the shape of this in 2019. Their phrase, the **"liar's dividend,"** names the windfall the guilty collect simply because fakes now exist: once a public is primed to doubt, anyone caught on tape can dismiss a *genuine* recording as a fake [4]. The fakes do not even have to be good. They only have to be *possible*. The mere existence of deepfakes hands every wrongdoer a pre-loaded defence, and it grows more plausible the more the public learns to distrust what it sees. Nobody has to prove a video is fake. They only have to make the audience unsure, and uncertainty is enough.

The dividend is no longer theoretical; it is being claimed in court, and so far it is failing. In a 2023 wrongful-death suit over a fatal Tesla Autopilot crash, Tesla's lawyers argued that Elon Musk's recorded statements might be deepfakes; Judge Evette Pennypacker rejected the move as **"deeply troubling,"** warning it would let public figures escape their own words [5]. In January 6 Capitol-riot trials, defendants argued the footage of them could be AI-fabricated. They were convicted anyway. Read these as attempts rather than victories. The courts have held the line so far, but that line is now being tested in every venue, and not every venue has a federal judge and an expert like Hany Farid on hand.

The foundation: humans can't tell, and the machines barely can

Here is the foundation the whole crisis rests on. Peer-reviewed work from the University of Florida found AI detectors up to **97% accurate** on deepfake images while **humans performed at chance**, coin-flip odds, dragged down by a built-in "truth bias" that wants to believe what it sees [6]. The vendor iProov ran 2,000 people in the US and UK: only **0.1%** correctly identified every real-versus-fake item, even when warned to look for fakes, and people were **36% worse** at catching deepfake video than images [7]. The vendor Sumsub reports deepfakes rose **fourfold** from 2023 to 2024, now about **7% of all fraud attempts** [8]. Read the vendor

figures as directional. Read the Florida result as the bottom line: the human eye is no longer an instrument of proof.

What makes the thesis hard to dismiss is not any single figure; it is that the same shape recurs across the cases, which share almost nothing else.

Signal	Finding	Source
Live deepfake fraud	~US$25.6M wired to a video call where every colleague was AI	CNN, Feb 2024; Guardian, May 2024
Imposter losses	Most-reported fraud category at $2.95B; total $12.5B, up 25%	U.S. FTC, March 2025
Human detection	Detectors ~97% on images; humans at chance, with a "truth bias"	Univ. of Florida (PMC), peer-reviewed
Warned humans	Only 0.1% spotted every fake; 36% worse on video than images	iProov, 2025 (vendor)
Rising volume	Deepfakes rose fourfold 2023⟶2024, now ~7% of fraud attempts	Sumsub, 2024 (vendor)

For a century, the recording was the witness

The photograph, then the audio recording, then the video, built a quiet civilisation of proof. A photo could place you somewhere. A recording could convict you. Footage could exonerate you. Journalism, courts, insurance, history itself, all of it leaned on a shared assumption: a recording, absent obvious tampering, was a trace of something that really happened. Faking it convincingly was expensive enough to be rare, and rare enough to trust.

Generative AI dissolved every part of that assumption at once. The cost, the skill, the time, the need for a source, all of it gone. What used to take a studio now takes a sentence. And the decisive shift is not that good fakes exist; it is that *everyone knows they exist.* Universal awareness is what converts a technical capability into a social collapse. You no longer need to be fooled by a fake to be harmed by it. It is enough to know that one was possible. That is the engine of the liar's dividend, and it runs in both directions: the guilty point at the real recording and cry "fake," while the innocent hold up their proof and hear "prove it's not."

Truth becomes a luxury good

If you can no longer trust a recording on its face, then trust has to be manufactured and attached in advance: cryptographically signed at capture, carried as content credentials, attested by hardware. That machinery is real, and it works. But it is **expensive, unevenly available, and optional.** Newsrooms, courts, governments, and corporations will buy it. The rest of the world will not.

So a divide opens. On one side stand the institutions and individuals wealthy enough to provision provenance, able to prove their own recordings are real and to verify everyone else's. On the other stands everyone whose proof is just a phone video with no pedigree: the worker documenting abuse, the citizen filming the police, the family with a voicemail that meant something. Their evidence still exists. It just no longer *counts.* When truth costs money, the people who cannot pay are the first to lose the right to be believed.

What to do: assume the channel can be faked, verify the source

Doubt is not destiny, but it is not answered with better fake-spotting; it is answered with verifiable origin. What that requires depends on who you are.

Reader	What it requires
Individuals	Treat any urgent voice or video request for money or secrets as unproven until confirmed on a separate, trusted channel, the single step that would have saved Arup's clerk. Agree a private code word with family. Do not let "it could be a deepfake" quietly excuse the powerful.
Employers	A live video call is not proof of identity, and Arup's loss proves it. Require out-of-band confirmation for any high-value transfer, with no exception for the CFO. Budget for imposter fraud as an operational risk, and adopt content provenance before a forged recording relies on you.
Policymakers	Fund and standardise provenance, the signing, content credentials, and attestation, as public infrastructure, so truth is not rationed by who can pay. Give judges and juries the authentication standards to keep rejecting the "it's a deepfake" defence, and protect the powerless side of the dividend.

Takeaways

1. The crisis is not that we will be fooled by a perfect forgery. It is that real footage of real events now arrives pre-doubted.
2. The deepfake fraud is already operational: roughly US$25.6 million wired to a meeting of ghosts, against a backdrop of record $12.5 billion in fraud losses led by imposter scams.
3. Detection is a losing arms race; every good detector trains the next fake to beat it, and even a perfect one leaves the doubt already planted.
4. The fix is provenance: proving what is real before the fact, not chasing what is fake after it.
5. The danger in that fix is access. If truth must be provisioned in advance and only the wealthy can buy it, the powerless lose the right to be believed.

Frequently asked questions

Isn't this just the old problem of photo retouching and Photoshop, but louder?

It differs in kind, not in degree. Photoshop demanded skill, time, and a source image to manipulate, which made a convincing fake of a specific person doing a specific thing expensive and rare. Generative tools collapse all three barriers at once. Anyone can now produce video and voice of anyone, saying anything, in minutes, for almost nothing. When the cost of a perfect fake falls to zero, the scarcity that made real evidence trustworthy disappears along with it.

If detectors are up to 97% accurate, hasn't technology already solved this?

Partly, and only in the lab. The University of Florida work found detectors near 97% on deepfake images while humans performed at chance, so machines can help where the human eye fails. But detection is an arms race, and every detector trains the next generation of fakes to beat it. Accuracy on a curated test set is also a long way from accuracy on a phone video forwarded six times through compression. The detector

is a tool rather than a cure. The deeper damage, the liar's dividend, survives even a perfect one, because by then the doubt is already planted.

What is the "liar's dividend," exactly?

It is the windfall the guilty collect simply because fakes now exist. Chesney and Citron named it in 2019: once the public knows any recording could be fabricated, a real recording can be waved away as fabricated. The liar profits from the very technology that threatens everyone else. Nobody has to prove a video is fake. They only have to make the audience unsure, and uncertainty is enough.

Has the "it's a deepfake" defence actually worked in court?

Not yet, and that is the encouraging part. In a wrongful-death suit over a fatal Tesla Autopilot crash, Tesla's lawyers floated that Elon Musk's recorded statements could be deepfakes; Judge Evette Pennypacker called the argument "deeply troubling" and rejected it. January 6 defendants who argued footage could be AI-fabricated were convicted anyway. So far the courts have refused the gambit. But these are attempts rather than anomalies, and the attempts will keep coming, in venues with far less expertise than a federal bench.

So what's the actual fix?

Provenance, not detection. Rather than asking whether something is fake after the fact, which is a losing game, we attach verifiable origin to real content at the moment of capture: cryptographic signing, content credentials, hardware-level attestation. The shift is from proving something is fake to proving something is real. It is harder, slower, and unevenly available, which is precisely the problem. Truth becomes something you have to provision in advance, and not everyone can.

References

1 CNN (2024). "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'." On the Arup Hong Kong deepfake fraud. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

2 The Guardian (May 17, 2024). "UK engineering firm Arup revealed as $25m deepfake scam victim." Arup's public confirmation of the Hong Kong office loss. https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video

3 U.S. Federal Trade Commission (March 10, 2025). "New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024." Imposter scams were the most-reported category, with $2.95B in reported losses. https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024

4 Chesney, R. & Citron, D. (2019). "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." 107 California Law Review 1753. Originators of the "liar's dividend." https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

5 NPR (May 8, 2023). "People are trying to claim real videos are deepfakes. The courts are not amused." On the Tesla Autopilot suit (Judge Evette Pennypacker) and Jan 6 trials; comments from Hany Farid and Rebecca Delfino. https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused

6 University of Florida et al. (peer-reviewed; PMC). Study finding AI detectors up to ~97% accurate on deepfake images while human accuracy sat at chance, with a measurable "truth bias." https://pmc.ncbi.nlm.nih.gov/articles/PMC12779810/

7 iProov (2025). "Study Reveals Deepfake Blindspot." A 2,000-person US/UK study: only 0.1% correctly identified every real-vs-deepfake item; people were 36% worse at spotting deepfake video than images. (Vendor research.) https://www.iproov.com/press/study-reveals-deepfake-blindspot-detect-ai-generated-content

8 Sumsub (2024). "Identity Fraud Report." Deepfakes rose fourfold from 2023 to 2024 and now make up roughly 7% of all fraud attempts. (Vendor research.) https://sumsub.com/blog/fraud-trends-sumsub-fraud-report/