Generative AI and a Skeptical Challenge to Digital Testimony

Ian M. Church

Abstract: Generative AI threatens knowledge online not merely by producing falsehoods, but by changing the epistemic environment in which online beliefs are formed. I argue that the spread of convincing AI-generated counterfeits creates a local skeptical challenge to digital testimony. The challenge is not that ordinary subjects must rule out every skeptical possibility. Rather, for many beliefs formed solely from unauthenticated online materials, ordinary subjects are situated among nearby, method-indistinguishable bad cases. In such environments, even true beliefs formed from genuine digital items may fail to satisfy a familiar anti-luck condition on knowledge. This does not show that we never know from online sources. It shows instead that digital-only uptake often falls short of knowledge unless supported by provenance, corroboration, or other independent checks. I close by suggesting that this narrowing of online knowledge may reinforce existing patterns of epistemic injustice.

Key Words: digital testimony; skepticism; generative AI; epistemic injustice; live alternatives; testimonial knowledge

Call the online resources (texts, images, videos, "fact-checks," etc.) on the basis of which one comes to believe a given proposition *digital testimony*.[1] In *good cases,* digital testimony accurately reports how things are. In *bad cases*, let's say, the digital testimony is produced by generative AI and is *not* appropriately connected to the facts it purports to report (e.g. a deepfake video, a fabricated bot-written report, a counterfeit site, etc.).[2]

In this paper, I argue that the proliferation of generative AI content online—along with our growing inability to distinguish good cases of digital testimony from bad cases[3]—significantly undermines our ability to acquire knowledge via digital testimony.[4] First, I explore the core argument, and then I consider two plausible objections before concluding.

---

[1] I am using "digital testimony" in a broad sense here. The paradigm cases of testimony in the epistemology literature involve verbal or written assertion: a speaker's saying or telling that *p*. And much of the literature focuses on how we learn from a speaker's "say-so" (Leonard 2021). But the nature of testimony is contested, and some accounts understand testimony more broadly in terms of communicative acts that convey, or are reasonably taken to convey, information to an audience (see, for example, Lackey 2008, 28). My usage is meant to capture this broader phenomenon in digital environments: not only online assertions, but also images, videos, recordings, fact-checks, websites, and other digital items that are circulated, framed, shared, trusted, amplified, and used to tell people how things are.

[2] I am not here suggesting that all AI-generated content is epistemically defective or disconnected from reality. My concern is with AI-generated content that functions as counterfeit digital testimony: content that presents itself as reporting or representing how things are while lacking the appropriate connection to the facts it purports to report.

[3] See, for example, Nightingale and Farid 2022; Bray et al. 2023; Jakesch et al. 2023.

[4] Related worries about deepfakes appear in Rini's (2020) claim that synthetic media erodes the "epistemic backstop" provided by recordings and thus threatens testimonial practices, and in Fallis's (2021) analysis that deepfakes reduce the information that videos carry to viewers. My aim is different: I generalize from video to digital testimony across modalities

The Argument

> **P1.** For many propositions *p* believed solely on the basis of digital testimony, an ordinary subject *S* is not in a position to rationally discriminate the good case from a live incompatible alternative (the AI-generated, unconnected case).

When a photo-realistic picture of Pope Francis wearing a stylish Balenciaga jacket began to circulate in 2023, many people came to believe that "Pope Francis has a Balenciaga jacket." But that belief turned out to be false; the photo-realistic picture of Pope Francis in a white, puffy Balenciaga jacket was generated by AI.[5] In response, a flurry of online articles popped up with the goal of helping the average internet user distinguish AI-generated images from genuine photos.

But AI-generated content has become increasingly prominent. Sometimes the content is obviously fake, but other times it mirrors reality so well that it's increasingly difficult (if not impossible) to distinguish real images, video, and text from fake images, video, and text. And as AI tools continue to improve and AI-generated content continues to proliferate, the epistemic challenge posed by such content will only grow worse.

It's important to stress that this problem is most pressing in cases where there is no easy route to independent verification. If a social media post tells me that my local hospital is closed, I can drive to the hospital and see for myself. But if the claim is that a hospital in Ukraine was bombed this morning, then digital testimony may be all I have. It's important to see that the argument here is targeted at those cases where digital testimony is not just one source among many, but the *only* source available to us.

But it's also important to note that digital testimony is not a fringe phenomenon; it is now the default channel through which many people learn about what lies beyond the radius of their direct experience.[6] National and international news reaches us first, and often only, through screens. Legacy outlets themselves scrape social feeds, pick up footage from anonymous accounts, and run stories built on viral posts long before reporters arrive on scene—if they arrive at all.[7] In practice, then, a vast share of what we take ourselves to know about elections, conflicts, disasters, scientific breakthroughs, or distant cultural events is mediated exclusively by unauthenticated digital items that circulate faster than any analog check can keep up. That is the domain P1 targets, and it is, by any reasonable measure, significant.

But as that domain becomes increasingly epistemically "polluted" by AI-generated content, it becomes increasingly difficult to distinguish good digital testimony from bad. Put in terms of familiar epistemological thought experiments, the analogs of "fake barns" (cf. Goldman 1976), "cleverly disguised mules" (cf. Dretske 1970) and "sheep-shaped rocks" (cf. Chisholm 1966) are becoming distressingly common in online environments.[8]

---

and target knowledge at uptake via a discrimination/live-alternative premise rather than a recordings-backstop or information-carrying account (though the diagnoses are complementary).

[5] For contemporaneous coverage, see Reuters Fact Check 2023.

[6] See especially Fletcher et al. 2025. Also see Schmidt et al. 2017; Allcott and Gentzkow 2017.

[7] See, for example, Rauchfleisch et al. 2017; von Nordheim et al. 2018; Hermida 2010.

[8] Matthews (2023) develops an explicit fake-barn analogy for knowledge from videos, arguing that deepfakes create barn-like risk; my use of the analogy here broader, applying to digital testimony (images, clips, text, "fact-checks") in unauthenticated, digital-only uptake.

P2. If *S* cannot discriminate *p* from a live incompatible alternative *q*, then *S* does not know *p*.[9]

Knowledge is commonly understood as demanding more than lucking-into truth; it requires an ability to rule out relevant error possibilities. If I'm traveling through a county with an abundance of fake barns, then I can't know "that's a barn" when I see what looks like a barn. I'm not in a position to know "that's a barn" because I'm not able to discriminate between genuine barns and barn facades in this county. Or if I'm visiting a zoo where I know some of the exotic animals are cleverly disguised commonplace animals dressed up to look like exotic animals, then I can't know that "that's a zebra" since for all I know it could be a cleverly disguised mule. Or suppose I glance at the bird feeder and announce, "That's a canary." If canaries and goldfinches are both common in my yard, and I know I can't tell them apart, the live alternative—it's a goldfinch—is every bit as compatible with what I see (cf. Stroud 1996). In that situation I may still form a true belief, but I cannot know it's a canary, because my evidence leaves the canary-and-goldfinch possibilities undifferentiated. Cases like these within the epistemology literature are commonplace. The same discrimination requirement applies to digital testimony: when the good digital item and its AI-generated **counterfeit of the relevant kind** ~~doppelgänger~~ are evidentially on a par, the relevant alternative isn't a distant skeptical fantasy—it's a live option inseparable from the evidence on offer. Where such parity of live alternatives persists, knowledge is undermined.

But what does it mean for an incompatible alternative to be "live"? The core idea is something like this: an incompatible alternative *q* is live for *S* at the time of uptake *iff* (i) in *S*'s actual environment there are nearby, ordinary cases in which *q* holds, and (ii) given the very way *S* formed the belief then, those *q*-cases would be indistinguishable by that method.[10] Let's say that I come to believe that my colleague had toast for breakfast based on her testimony saying as much. Of course, she could have lied, and I wouldn't have had any way to tell; however, if there are no nearby, ordinary cases in which she *does* lie—she's not prone to lie about such things—such an incompatible alternative is not "live." If my colleague was heavily invested in a local bread company, such that she frequently lies about what she had for breakfast in hopes of drumming up more business for that bread company, then there are nearby, ordinary cases in which such an incompatible alternative holds that would be indistinguishable for me merely on the basis of her testimony. But more to the point of the current argument, bad cases of digital testimony are increasingly prevalent in our digital environment and, given the usual means of forming beliefs based on digital content, we're unable to distinguish good cases of digital testimony from bad. Such bad cases of digital testimony are, by these lights, "live" incompatible alternatives.

Consider a zoo that contains many real animals, but also many highly realistic animatronic replicas—tigers, lions, monkeys, and giraffes—replicas that are indistinguishable from the real animals by ordinary viewing in that setting. A visitor sees a zebra in its enclosure and forms the belief there is a zebra. Even if the animal she is looking at is in fact a real zebra, there are nearby ordinary cases in that very environment in which visually similar animal-presentations, assessed by the same ordinary method of uptake, are animatronic replicas. In that setting, the relevant incompatible alternative is

[9] This can also be put in familiar externalist terms: S's belief is unsafe when, in nearby relevant cases, S would form the same belief by the same method though *p* is false. I use discrimination-language to emphasize the subject's position at uptake, but the anti-luck point can be stated either way.

[10] Nothing here requires that S possess any special discriminatory skill. "Live" is fixed by the environment and method of uptake (as defined above). In ordinary deference cases the incompatible option is not live; in unauthenticated digital-only uptake amid prevalent convincing fakes, it is. That is enough for P2.

not best understood as there is a fake zebra nearby. Rather, the visitor is forming animal-identification beliefs in an environment containing many indistinguishable replicas of the relevant general kind, and her ordinary method of uptake does not reliably discriminate real animals from such replicas. Given the prevalence of such replicas and the visitor's ordinary method of belief formation, the relevant alternative is live, and so her true belief will not amount to knowledge. Online environments increasingly resemble the fake-animals zoo: even when one encounters a genuine clip, the presence of widespread, indistinguishable AI-generated counterfeits of similar kinds makes the relevant counterfeit alternative live for ordinary subjects relying only on the usual cues.

> **C.** Therefore, for many propositions believed solely on the basis of digital testimony, *S* does not know that *p*.

If the argument above is sound, then something important follows: we are beginning to lose the ability to know things we badly need to know. Consider the following high-stakes case. Suppose, for example, the Department of Justice circulates a video online showing Ghislaine Maxwell testifying under oath that Donald Trump never visited Epstein's island or participated in any of Epstein's illicit activities. Even if it looks perfectly authentic, we no longer have any reliable way to tell that it's not a carefully engineered deepfake. In an environment saturated with indistinguishable AI-generated counterfeits of similar clips, the deepfake alternative is live even absent a token counterfeit of this particular hearing. That fact blocks knowledge for ordinary subjects who rely solely on the circulating video. In cases like this—where digital testimony is the only basis for the belief—the digital item alone, without independent provenance or offline corroboration from a trusted source, does not put us in a position to know.[11]

There's also a deeper worry here—one that presses beyond epistemology and into ethics and social epistemology. The internet was supposed to democratize access to knowledge. For a time, it seemed to. Ordinary people could find information, hear marginalized voices, learn about the world, and tell their stories. Digital testimony was hailed as a great epistemic equalizer, a tool for leveling unjust disparities in access and authority.[12] But if the argument above is right, that promise may be collapsing. The ability to know through digital testimony, a power once hoped would be extended to all, is being eroded.

And the cost may not be evenly distributed. When online testimony becomes broadly suspect, it is often the least institutionally credentialed voices that suffer first. If my story is published in a legacy outlet with named sources and editorial review, it may survive the scrutiny.[13] But if I post it myself, or share it through informal channels, it becomes just another unauthenticated digital item—indistinguishable from the ocean of AI-generated content. In this way, the collapse of digital testimony risks reproducing old hierarchies of credibility. It systematically undermines the very people who depended most on the openness and accessibility of online platforms to be heard and believed. That is a kind of epistemic injustice—not because their stories are disbelieved on the basis of identity or bias, but because hearers are no longer in a position to *know*.

---

[11] This is distinct from the further question of whether subjects who recognize this predicament ought to suspend belief. My central claim here is only that the digital item alone does not suffice for knowledge.

[12] For a widely cited, seminal text exploring these ideas, see Benkler 2006.

[13] Though even here it's worth noting how unauthenticated digital testimony is actively eroding even trust in legacy news sources. See Fletcher et al. 2025.

None of this requires bad faith. We may want to believe the whistleblower's account, or the video from the conflict zone, or the activist's testimony. But if the evidence is digital-only, and the fakes are too good and prevalent, and the provenance too obscure, then we may not be in a position to rationally tell what's real. The result is a quiet, corrosive silencing—not through censorship, but through epistemic breakdown. The tragedy is not just that people are lied to. It's that even when they're told the truth, they may no longer be able to know it.

Objection 1: Over-breadth

But surely we still know plenty of things online. Reputable outlets apply editorial checks, many videos now carry cryptographic provenance tags, and major events are eventually corroborated by on-the-ground reporting. Perhaps the skeptical damage done by the above argument isn't substantial.[14]

Reply

The argument only targets those beliefs formed solely on the basis of unauthenticated digital testimony. That domain is large—breaking news, viral images, fast-moving political clips—but it is not universal. Where an item comes with verifiable provenance (hardware-backed capture, C2PA manifests actually checked, on-the-record sourcing that can be followed up), the relevant AI-bad alternative may no longer be live, and the conclusion may not apply. What the argument shows is that, in the very circumstances where most users first form beliefs about distant events, the discriminatory gap remains: provenance is usually absent, dropped in reposts, or ignored by design. Later editorial confirmation may restore knowledge, but that does nothing for the initial level-one uptake on which public opinion, sharing, and action often depend. The result is not that we never know from screens, but that a wide and epistemically significant share of our first-pass online beliefs—precisely the ones that shape rapid discourse—fall short of knowledge until independent verification arrives (if it ever does).

Objection 2: No New Threat

Digital fakery is old news. People have been doctoring photographs, forging documents, and editing audio for decades. Generative AI is just a faster version of this technology; it doesn't introduce a novel epistemic problem.

Reply

Traditional manipulation was costly, slow, and often detectable. High-skill editors could alter a single photograph or splice a brief audio clip, but large-scale fabrication—coherent images, video, and matching text—required resources few people possessed, and tell-tale artifacts (jagged edges, mismatched lighting, compression seams) often betrayed the forgery. Generative AI systems lower all of these barriers simultaneously. Anyone with a web prompt can now produce photorealistic images and fluid text; no specialist training is needed. A single model can generate thousands of variants in minutes, flooding timelines before any manual check can keep pace. The same engine can output an

[14] For caution against "epistemic apocalypse" narratives about deepfakes, see Habgood-Coote (2023). My thesis is deliberately local: it concerns unauthenticated, digital-only uptake and is compatible with later independent verification restoring knowledge.

image, a matching "behind-the-scenes" video, synthetic eyewitness tweets, and a long-form article—each supporting the others and erasing the old tell-tale seams. And current models eliminate many artifacts that traditional detection methods relied on; the usual pixel-level heuristics—odd hands, warped shadows—no longer reliably appear.

The epistemic upshot is a shift from occasional, detectable fakes to ubiquitous, method-matched alternatives that are, for ordinary users at belief-formation time, indistinguishable from genuine testimony. That ubiquity and parity are what drive P1; the difference in degree and kind are substantial, and they are what underwrite the conclusion that knowledge fails for a wide class of digital-only beliefs.

Conclusion

The argument here is not that the digital sphere has collapsed, but that it faces new epistemic threats. In an environment where AI-generated lookalikes can be polished, pervasive, and engineered to pass the checks ordinary users employ, the bad cases of digital testimony are live options for a significant set of the propositions we come to believe solely on the basis of online resources. When such incompatible alternatives are live, knowledge is undermined. Of course, this does not lead us to global skepticism—it does not deny that truth, and even knowledge, can be had—but it does narrow the range of cases where digital-only testimony can give it.[15] That narrowing comes at a cost: much of the internet's value has rested on its ability to give us knowledge of the wider world from afar, and when that is lost, so too is a distinctive kind of epistemic reach. And when access to knowledge depends on who can supply verifiable provenance, the losses will fall unevenly, exacerbating familiar patterns of epistemic injustice.

Bibliography

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36. https://doi.org/10.1257/jep.31.2.211.

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Bray, Sergi D., Shane D. Johnson, and Bennett Kleinberg. 2023. "Testing Human Ability to Detect 'Deepfake' Images of Human Faces." *Journal of Cybersecurity* 9 (1): tyad011. https://doi.org/10.1093/cybsec/tyad011.

Chisholm, Roderick M. 1966. *Theory of Knowledge*. Prentice-Hall.

Dretske, Fred I. 1970. "Epistemic Operators." *The Journal of Philosophy* 67 (24): 1007–23.

Fallis, Don. 2021. "The Epistemic Threat of Deepfakes." *Philosophy & Technology* 34 (4): 623–43. https://doi.org/10.1007/s13347-020-00419-2.

[15] And since the skeptical upshot of the above argument isn't global, it's worth noting that some of the standard responses to global skeptical arguments (Moore 1939; Wittgenstein 1975) seemingly won't be effective. Also see Pritchard 2018.

Fletcher, Richard, Simge Andı, Sumitra Badrinathan, et al. 2025. "The Link between Changing News Use and Trust: Longitudinal Analysis of 46 Countries." *Journal of Communication* 75 (1): 1–15. https://doi.org/10.1093/joc/jqae044.

Goldman, Alvin I. 1976. "Discrimination and Perceptual Knowledge." *The Journal of Philosophy* 73 (20): 771–91.

Habgood-Coote, Joshua. 2023. "Deepfakes and the Epistemic Apocalypse." *Synthese* 201 (3): 103. https://doi.org/10.1007/s11229-023-04097-3.

Hermida, Alfred. 2010. "Twittering the News: The Emergence of Ambient Journalism." *Journalism Practice* 4 (3): 297–308. https://doi.org/10.1080/17512781003640703.

Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. 2023. "Human Heuristics for AI-Generated Language Are Flawed." *Proceedings of the National Academy of Sciences of the United States of America* 120 (11): e2208839120. https://doi.org/10.1073/pnas.2208839120.

Lackey, Jennifer. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.

Leonard, Nick. 2021. "Epistemological Problems of Testimony." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Stanford University. https://plato.stanford.edu/archives/spr2023/entries/testimony-episprob/.

Matthews, Taylor. 2023. "Deepfakes, Fake Barns, and Knowledge from Videos." *Synthese* 201 (2): 41. https://doi.org/10.1007/s11229-022-04033-x.

Moore, G. E. 1939. "Proof of an External World." *Proceedings of the British Academy* 25 (5): 273–300.

Nightingale, Sophie J., and Hany Farid. 2022. "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy." *Proceedings of the National Academy of Sciences of the United States of America* 119 (8): e2120481119. https://doi.org/10.1073/pnas.2120481119.

Nordheim, Gerret von, Karin Boczek, and Lars Koppers. 2018. "Sourcing the Sources: An Analysis of the Use of Twitter and Facebook as a Journalistic Source over 10 Years in The New York Times, The Guardian, and Süddeutsche Zeitung." *Digital Journalism* 6 (7): 807–28. https://doi.org/10.1080/21670811.2018.1490658.

Pritchard, Duncan. 2018. "Epistemic Angst." *Philosophy and Phenomenological Research* 96 (1): 70–90.

Rauchfleisch, Adrian, Xenia Artho, Julia Metag, Senja Post, and Mike S. Schäfer. 2017. "How Journalists Verify User-Generated Content during Terrorist Crises. Analyzing Twitter Communication during the Brussels Attacks." *Social Media + Society* 3 (3): 2056305117717888. https://doi.org/10.1177/2056305117717888.

Reuters Fact Check. 2023. "Fact Check: Image of Pope Francis Wearing Oversized White Puffer Coat Is AI-Generated." Fact Check. *Reuters*, March 29. https://www.reuters.com/article/fact-check/image-of-pope-francis-wearing-oversized-white-puffer-coat-is-ai-generated-idUSL1N36120G/.

Rini, Regina. 2020. "Deepfakes and the Epistemic Backstop." *Philosopher's Imprint* 20 (24). http://hdl.handle.net/2027/spo.3521354.0020.024.

Schmidt, Ana Lucía, Fabiana Zollo, Michela Del Vicario, et al. 2017. "Anatomy of News Consumption on Facebook." *Proceedings of the National Academy of Sciences of the United States of America* 114 (12): 3035–39. https://doi.org/10.1073/pnas.1617052114.

Stroud, Barry. 1996. "Epistemological Reflection on Knowledge of the External World." *Philosophy and Phenomenological Research* 56 (2): 345–58. https://doi.org/10.2307/2108525.

Wittgenstein, Ludwig. 1975. *On Certainty*. Edited by G. E. M. Anscombe and G. H. von Wright. Translated by Denis Paul and G. E. M. Anscombe. Blackwell Publishing.